Nós de Modelagem do IBM SPSS Modeler 17

Observação

Antes de usar estas informações e o produto por elas suportadas, leia as informações em “Avisos” na página 341.

Informações do produto

Esta edição se aplica à versão 17, release 0, modificação 0 do IBM(r) SPSS(r) Modeler e a todas as liberações e modificações subsequentes, até que seja indicado de outra forma em novas edições.

Índice

Prefácio

IBM® SPSS Modeler é o ambiente de trabalho de mineração de dados de força corporativa do IBM Corp.. O SPSS Modeler ajuda as organizações a melhorarem as relações com o cliente e com o cidadão por meio de um entendimento profundo dos dados. As organizações utilizam o insight adquirido do SPSS Modeler para reter clientes rentáveis, identificar oportunidades de venda cruzada, atrair novos clientes, detectar fraude, reduzir o risco e melhorar a entrega de serviço de governo.

A interface visual do SPSS Modeler convida os usuários a aplicarem seus conhecimentos de negócios específicos, levando a modelos preditivos mais poderosos e reduzindo o tempo para a solução. O SPSS Modeler oferece muitas técnicas de modelagem, como predição, classificação, segmentação e algoritmos de detecção de associação. Quando os modelos são criados, o IBM SPSS Modeler Solution Publisher permite entregá-los aos tomadores de decisão na empresa ou a um banco de dados.

Sobre o IBM Business Analytics

O software IBM Business Analytics fornece informações completas, consistentes e exatas nas quais os tomadores de decisão confiam para melhorar o desempenho de negócios. Um portfólio abrangente de inteligência de negócios, analise preditiva, gerenciamento de desempenho e estratégia financeira e aplicativos analíticos fornece insights claros, imediatos e práticos sobre o desempenho atual e a capacidade de prever resultados futuros. Combinado com soluções para segmentos do mercado, práticas comprovadas e serviços profissionais completos, organizações de qualquer tamanho poderão conduzir maior produtividade, automatizar as decisões de modo confiável e entregar melhores resultados.

Como parte deste portfólio, o software IBM SPSS Predictive Analytics ajuda as organizações a preverem eventos futuros e agirem proativamente nesse insight para conduzir os melhores resultados de negócios. Clientes comerciais, governamentais e acadêmicos do mundo todo confiam na tecnologia IBM SPSS como uma vantagem competitiva para atrair, reter e aumentar clientes, enquanto reduz a fraude e minimiza riscos. Ao incorporar o software IBM SPSS em suas operações diárias, as organizações se tornam empresas preditivas, ou seja, capazes de direcionar e de automatizar as decisões para atender às metas de negócios e obter vantagem competitiva mensuráveis. Para obter mais informações ou entrar em contato com um representante, visite http://www.ibm.com/spss.

Suporte técnico

O suporte técnico está disponível para clientes de manutenção. Os clientes podem entrar em contato com o Suporte Técnico para obterem assistência com o uso de produtos IBM Corp. ou para obterem ajuda com a instalação de um dos ambientes de hardware suportados. Para entrar em contato com o Suporte Técnico, consulte o website do IBM Corp. em http://www.ibm.com/support. Esteja preparado para se identificar, identificar sua organização e sua concordância de suporte ao solicitar assistência.

Capítulo 1. Sobre o IBM SPSS Modeler

O IBM SPSS Modeler é um conjunto de ferramentas de mineração de dados que permite desenvolver rapidamente modelos preditivos usando o conhecimento de negócios, e implementá-los em operações de negócios para melhorar a tomada de decisão. Projetado em torno do modelo CRISP-DM padrão de mercado, o IBM SPSS Modeler suporta todo o processo de mineração de dados, a partir dos dados para melhores resultados de negócios.

O IBM SPSS Modeler oferece uma variedade de métodos de modelagem a partir do aprendizado de máquina, inteligência artificial e estatísticas. Os métodos disponíveis na paleta Modelagem permitem derivar informações novas a partir dos dados e desenvolver modelos preditivos. Cada método possui determinadas intensidades e é mais bem adequado para tipos de problemas específicos.

O SPSS Modeler pode ser comprado como um produto independente, ou usado como um cliente na combinação com o SPSS Modeler Server. Várias opções adicionais também estão disponíveis, conforme sumarizadas nas seções a seguir. Para obter mais informações, consulte http://www.ibm.com/software/analytics/spss/products/modeler/.

Produtos do IBM SPSS Modeler

A família de produtos e software associado do IBM SPSS Modeler consistem no seguinte.

- IBM SPSS Modeler
- IBM SPSS Modeler Server
- IBM SPSS Modeler Administration Console
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- Adaptadores do IBM SPSS Modeler Server para IBM SPSS Collaboration and Deployment Services

IBM SPSS Modeler

SPSS Modeler é uma versão funcionalmente completa do produto que você instala e executa em seu computador pessoal. É possível executar o SPSS Modeler no modo local como um produto independente ou usá-lo no modo distribuído com IBM SPSS Modeler Server para melhorar o desempenho em conjuntos de dados grandes.

Com o SPSS Modeler, é possível construir modelos preditivos exatos de maneira rápida e intuitiva, sem programação. Usando a interface visual exclusiva, é possível visualizar facilmente o processo de mineração de dados. Com o suporte da análise avançada integrada ao produto, é possível descobrir tendências e padrões ocultos anteriormente em seus dados. É possível modelar resultados e entender os fatores que os influenciam, permitindo aproveitar as vantagens das oportunidades de negócios e diminuir os riscos.

SPSS Modeler está disponível em duas edições: SPSS Modeler Professional e SPSS Modeler Premium. Consulte o tópico “Edições do IBM SPSS Modeler” na página 2 para obter mais informações.

IBM SPSS Modeler Server

O SPSS Modeler utiliza uma arquitetura de cliente/servidor para distribuir as solicitações de operações intensivas em recursos para um software do servidor potente, resultando em desempenho mais rápido em conjuntos de dados maiores.

O SPSS Modeler Server é um produto licenciado separadamente que executa continuamente em modo de análise distribuída em um host do servidor em conjunto com uma ou mais instalações do IBM SPSS Modeler. Dessa forma, o SPSS Modeler Server fornece desempenho superior em conjuntos de dados grandes porque as operações intensivas em memória podem ser feitas no servidor sem fazer download dos dados para o computador cliente. O IBM SPSS Modeler Server também fornece suporte para otimização SQL e para recursos de modelagem dentro da base de dados, proporcionando benefícios adicionais em desempenho e automação.

IBM SPSS Modeler Administration Console

O Modeler Administration Console é um aplicativo gráfico para gerenciar muitas das opções de configuração do SPSS Modeler Server, que também são configuráveis por meio de um arquivo de opções. O aplicativo fornece uma interface com o usuário do console para monitorar e configurar suas instalações do SPSS Modeler Server e está disponível gratuitamente para os clientes atuais do SPSS Modeler Server. O aplicativo pode ser instalado apenas em computadores Windows, no entanto, ele pode administrar um servidor instalado em qualquer plataforma suportada.

IBM SPSS Modeler Batch

Embora geralmente a mineração de dados seja um processo interativo, também é possível executar o SPSS Modeler a partir de uma linha de comandos, sem a necessidade de uma interface gráfica com o usuário. Por exemplo, você pode ter tarefas repetidas ou de longa execução que deseja executar sem intervenção do usuário. SPSS Modeler Batch é uma versão especial do produto que fornece suporte para capacidades de análise completa do SPSS Modeler sem acessar a interface com o usuário regular. SPSS Modeler Server é necessário para usar o SPSS Modeler Batch.

IBM SPSS Modeler Solution Publisher

O SPSS Modeler Solution Publisher é uma ferramenta que permite criar uma versão compactada de um fluxo do SPSS Modeler que pode ser executada por um mecanismo de tempo de execução externo ou integrada em um aplicativo externo. Desta maneira, é possível publicar e implementar fluxos completos do SPSS Modeler para uso em ambientes que não tiverem o SPSS Modeler instalado. O SPSS Modeler Solution Publisher é distribuído como parte do serviço do IBM SPSS Collaboration and Deployment Services - Scoring, para o qual uma licença separada é necessária. Com esta licença, você receberá o SPSS Modeler Solution Publisher Runtime que permite executar os fluxos publicados.

Para obter mais informações sobre o SPSS Modeler Solution Publisher, consulte a documentação do IBM SPSS Collaboration and Deployment Services. O Centro de Conhecimento do IBM SPSS Collaboration and Deployment Services contém seções chamadas "IBM SPSS Modeler Solution Publisher" e "IBM SPSS Analytics Toolkit".

Adaptadores do IBM SPSS Modeler Server para IBM SPSS Collaboration and Deployment Services

Diversos adaptadores para o IBM SPSS Collaboration and Deployment Services estão disponíveis que permitem que o SPSS Modeler e o SPSS Modeler Server interajam com um repositório do IBM SPSS Collaboration and Deployment Services. Dessa forma, um fluxo do SPSS Modeler implementado no repositório pode ser compartilhado por diversos usuários ou acessado a partir do aplicativo thin client do IBM SPSS Modeler Advantage. Instale o adaptador no sistema que hospeda o repositório.

Edições do IBM SPSS Modeler

O SPSS Modeler está disponível nas edições a seguir.

SPSS Modeler Professional

O SPSS Modeler Professional fornece todas as ferramentas necessárias para trabalhar com a maioria dos tipos de dados estruturados, como comportamentos e interações rastreados em sistemas CRM,

informações demográficas, comportamento de compras e dados de vendas.

SPSS Modeler Premium

O SPSS Modeler Premium é um produto licenciado separadamente que estende o SPSS Modeler Professional para trabalhar com dados especializados, como dados utilizados para análise de entidade ou de rede social, e com dados de texto não estruturado. O SPSS Modeler Premium inclui os componentes a seguir.

O **IBM SPSS Modeler Entity Analytics** inclui uma dimensão extra à análise preditiva do IBM SPSS Modeler. Considerando que a análise preditiva tenta prever comportamento futuro de dados passados, a análise de entidade foca melhorar a coerência e a consistência dos dados atuais resolvendo conflitos de identidade nos registros em si. Uma identidade pode ser de um indivíduo, de uma organização, de um objeto ou de qualquer outra entidade para a qual possa existir ambiguidade. A resolução de identidade pode ser vital em diversos campos, incluindo gerenciamento de relacionamento com o cliente, detecção de fraude, medidas contra lavagem de dinheiro e segurança nacional e internacional.

IBM SPSS Modeler Social Network Analysis transforma informações sobre relacionamentos em campos que caracterizam o comportamento social dos indivíduos e grupos. Usando dados que descrevem os relacionamentos subjacentes das redes sociais, o IBM SPSS Modeler Social Network Analysis identifica os líderes sociais que influenciam o comportamento dos outros na rede. Além disso, é possível determinar quais pessoas são mais afetadas por outros participantes da rede. Combinando esses resultados com outras medidas, é possível criar perfis abrangentes de indivíduos nos quais basear modelos preditivos. Os modelos que incluem essas informações sociais serão executados melhor do que os modelos que não incluem.

IBM SPSS Modeler Text Analytics usa tecnologias de linguística avançada e processamento de linguagem natural (NLP) para processar rapidamente uma grande variedade de dados de texto não estruturados, extrair e organizar conceitos chave e agrupar esses conceitos em categorias. Categorias e conceitos extraídos podem ser combinados com dados estruturados existentes, como demográficos, e aplicados à modelagem usando o conjunto completo de ferramentas de mineração de dados do IBM SPSS Modeler para gerar decisões melhores e mais focadas.

Documentação do IBM SPSS Modeler

A documentação em formato de ajuda online está disponível no menu Ajuda do SPSS Modeler. Isso inclui documentação para SPSS Modeler, SPSS Modeler Server, bem como o Guia de Aplicativos (também chamado de Tutorial) e outros materiais de apoio.

A documentação completa para cada produto (incluindo instruções de instalação) está disponível em formato PDF sob a pasta *Documentation* em cada produto DVD. Os documentos de instalação também podem ser transferidos por download na web em http://www.ibm.com/support/docview.wss?uid=swg27043831.

A documentação em ambos os formatos também está disponível no SPSS Modeler Knowledge Center em http://www-01.ibm.com/support/knowledgecenter/SS3RA7_17.0.0.0.

Documentação do SPSS Modeler Professional

O Conjunto de Documentações do SPSS Modeler Professional (exceto instruções de instalação) é o seguinte.

- **Guia do Usuário do IBM SPSS Modeler.** Introdução geral ao uso do SPSS Modeler, incluindo como construir fluxos de dados, manipular valores omissos, construir expressões do CLEM, trabalhar com projetos e relatórios e empacotar fluxos para implementação no IBM SPSS Collaboration and Deployment Services, em Aplicativos Preditivos ou no IBM SPSS Modeler Advantage.

- **Nós de Origem, de Processo e de Saída do IBM SPSS Modeler.** Descrições de todos os nós utilizados para ler, processar e gerar dados em formatos diferentes. Efetivamente, isso significa todos os nós diferentes dos nós de modelagem.
- **Nós de Modelagem do IBM SPSS Modeler.** Descrições de todos os nós utilizados para criar modelos de mineração de dados. O IBM SPSS Modeler oferece uma variedade de métodos de modelagem a partir do aprendizado de máquina, inteligência artificial e estatísticas.
- **Guia de Algoritmos do IBM SPSS Modeler.** Descrições das bases matemáticas dos métodos de modelagem utilizados no IBM SPSS Modeler. Este guia está disponível apenas em formato PDF.
- **Guia de Aplicativos do IBM SPSS Modeler.** Os exemplos neste guia fornecem introduções breves e destinadas aos métodos e técnicas de modelagem específicos. Uma versão online deste guia também está disponível a partir do menu Ajuda. Consulte o tópico "Exemplos de Aplicativos" na página 5 para obter mais informações.
- **Script e Automação Python do IBM SPSS Modeler.** Informações sobre como automatizar o sistema por meio de scripts Python, incluindo as propriedades que podem ser utilizadas para manipular nós e fluxos.
- **Guia de Implementação do IBM SPSS Modeler.** Informações sobre como executar os fluxos e cenários do IBM SPSS Modeler como passos no processamento de tarefas no IBM SPSS Collaboration and Deployment Services Deployment Manager.
- **Guia do Desenvolvedor do IBM SPSS Modeler CLEF.** O CLEF fornece a capacidade de integrar programas de terceiros, como rotinas de processamento de dados ou algoritmos de modelagem como nós no IBM SPSS Modeler.
- **Guia de Mineração Dentro do Banco de Dados do IBM SPSS Modeler.** Informações sobre como utilizar o poder de seu banco de dados para melhorar o desempenho e estender a variedade de recursos analíticos por meio de algoritmos de terceiros.
- **Guia de Desempenho e de Administração do IBM SPSS Modeler Server.** Informações sobre como configurar e administrar o IBM SPSS Modeler Server.
- **Guia do Usuário do Console de Administração do IBM SPSS Modeler.** Informações sobre como instalar e utilizar a interface com o usuário do console para monitoramento e configuração do IBM SPSS Modeler Server. O console é implementado como um plug-in para o aplicativo Deployment Manager.
- **Guia do IBM SPSS Modeler CRISP-DM.** Guia passo a passo para usar a metodologia CRISP-DM para mineração de dados com o SPSS Modeler.
- **Guia do Usuário do IBM SPSS Modeler Batch.** Guia completo para usar o IBM SPSS Modeler no modo em lote, incluindo detalhes sobre a execução no modo em lote e argumentos da linha de comandos. Este guia está disponível apenas em formato PDF.

Documentação do SPSS Modeler Premium

O Conjunto de Documentações do SPSS Modeler Premium (exceto instruções de instalação) é o seguinte.

- **Guia do Usuário do IBM SPSS Modeler Entity Analytics.** Informações sobre como utilizar analítica de entidade com o SPSS Modeler, abrangendo instalação e configuração de repositório, nós de analítica de entidade e tarefas administrativas.
- **Guia do Usuário do IBM SPSS Modeler Social Network Analysis.** Um guia para executar análise de rede social com o SPSS Modeler, incluindo análise do grupo e análise de difusão.
- **Guia do Usuário do SPSS Modeler Text Analytics .** Informações sobre como utilizar analítica de texto com o SPSS Modeler, que abrange os nós de mineração de texto, ambiente de trabalho interativo, modelos e outros recursos.

Exemplos de Aplicativos

Enquanto as ferramentas de mineração de dados no SPSS Modeler podem ajudar a resolver uma ampla variedade de negócios e problemas organizacionais, os exemplos de aplicativos fornecem introduções breves e destinadas aos métodos e técnicas de modelagem específicos. Os conjuntos de dados usados aqui são muito menores do que os enormes armazenamentos de dados gerenciados por alguns mineradores de dados, mas os conceitos e métodos envolvidos devem ser escaláveis para aplicativos reais.

É possível acessar os exemplos clicando em **Exemplos de Aplicativos** no menu de Ajuda no SPSS Modeler. Os arquivos de dados e os fluxos de amostra são instalados na pasta *Demos* no diretório de instalação do produto. Consulte o tópico "Pasta Demos" para obter mais informações.

Exemplos de modelagem da base de dados. Consulte os exemplos no *Guia de Mineração dentro do Banco de Dados do IBM SPSS Modeler*.

Exemplos de script. Consulte os exemplos no *Guia de Script e Automação do IBM SPSS Modeler*.

Pasta Demos

Os arquivos de dados e os fluxos de amostra utilizados com os exemplos de aplicativos são instalados na pasta *Demos* no diretório de instalação do produto. Esta pasta também pode ser acessada a partir do grupo do programa IBM SPSS Modeler no menu Iniciar do Windows ou clicando em *Demos* na lista de diretórios recentes na caixa de diálogo Abrir Arquivo.

Capítulo 2. Introdução à Modelagem

Um modelo é um conjunto de regras, fórmulas ou equações que podem ser utilizadas para prever um resultado com base em um conjunto de campos de entrada ou variáveis. Por exemplo, uma instituição financeira pode utilizar um modelo para prever se os solicitantes de empréstimo poderão representar um bom ou mau risco, com base nas informações que já se conhece sobre os solicitantes passados.

A capacidade de prever um resultado é o objetivo central de análise preditiva e entender o processo de modelagem é a chave para utilizar o IBM SPSS Modeler.

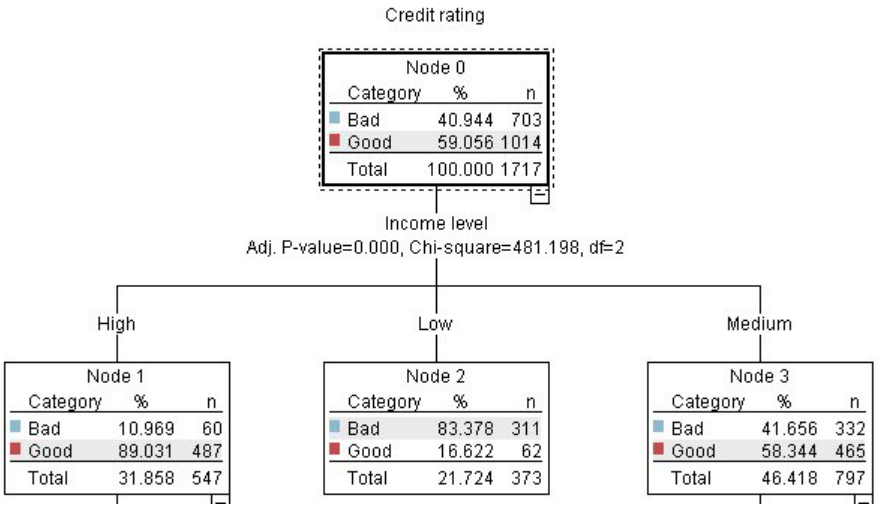


Figura 1. Um modelo de árvore de decisão simples

Esse exemplo utiliza um modelo de **árvore de decisão**, que classifica registros (e prediz uma resposta) usando uma série de regras de decisão, por exemplo:

```
IF income = Medium
AND cards <5
THEN -> 'Good'
```

Embora este exemplo utilize um modelo CHAID (Chi-squared Automatic Interaction Detection), ele é destinado como uma introdução geral e a maioria dos conceitos se aplica amplamente a outros tipos de modelagem no IBM SPSS Modeler.

Para entender qualquer modelo, primeiro deve-se entender os dados que entram nele. Os dados neste exemplo contêm informações sobre os clientes de um banco. Os campos a seguir são utilizados:

Nome do campo	Descrição
Credit_rating	Classificação de crédito: 0=Bad, 1=Good, 9=missing values
Idade	Idade em anos
Renda	Nível de renda: 1=Low, 2=Medium, 3=High
Credit_cards	Número de cartões de crédito que possui: 1=Less than five, 2=Five or more
Educação	Nível de educação: 1=High school, 2=College
Car_loans	Número de empréstimos para compra de carro contraídos: 1=None or one, 2=More than two

O banco mantém um banco de dados de informações históricas sobre os clientes que contraíram empréstimos do banco, incluindo se eles pagaram os empréstimos (Classificação de crédito = Bom) ou se ficaram inadimplentes (Classificação de crédito = Ruim). Utilizando esses dados existentes, o banco constrói um modelo que permitirá prever quão provavelmente futuros solicitantes de empréstimo se tornarão inadimplentes.

Utilizando um modelo de árvore de decisão, é possível analisar as características dos dois grupos de clientes e prever a probabilidade de inadimplência no empréstimo.

Esse exemplo usa o fluxo denominado *modelingintro.str*, disponível na pasta *Demos* sob a subpasta *streams*. O arquivo de dados é *tree_credit.sav*. Consulte o tópico "Pasta Demos" na página 5 para obter mais informações.

Vamos dar uma olhada no fluxo.

1. Escolha o seguinte no menu principal:
   Arquivo > **Abrir Fluxo**
2. Clique no ícone de pepita de ouro na barra de ferramentas da caixa de diálogo Abrir e escolha a pasta Demos.
3. Clique duas vezes na pasta *streams*.
4. Clique duas vezes no arquivo denominado *modelingintro.str*.

Construindo o Fluxo

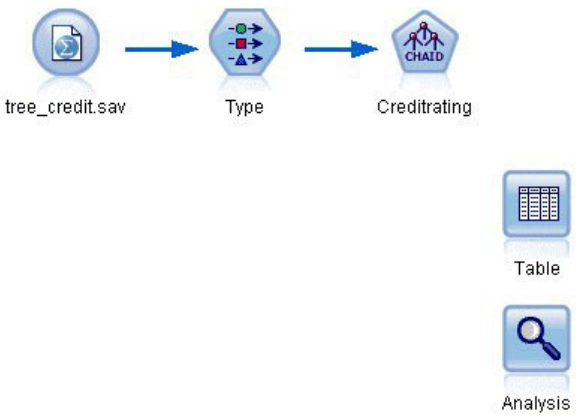


Figura 2. Fluxo de Modelagem

Para construir um fluxo que criará um modelo, pelo menos três elementos são necessários:

- Um nó de origem que lê dados a partir de alguma origem externa, nesse caso, um arquivo de dados do IBM SPSS Statistics.
- Uma origem ou nó Tipo que especifica as propriedades do campo, como nível de medição (o tipo de dados que o campo contém) e o papel de cada campo como um destino ou entrada na modelagem.
- Um nó de modelagem que gera um nugget do modelo quando o fluxo é executado.

Neste exemplo, estamos utilizando um nó de modelagem CHAID. O CHAID, ou Chi-squared Automatic Interaction Detection, é um método de classificação que constrói as árvores de decisão usando um tipo específico de estatísticas conhecido como estatísticas qui-quadrado para descobrir os melhores locais para fazer as divisões na árvore de decisão.

Se os níveis de medição forem especificados no nó de origem, o nó Tipo separado poderá ser eliminado. Funcionalmente, o resultado é o mesmo.

Este fluxo também tem os nós Tabela e Análise que serão usados para visualizar os resultados da escoragem após o nugget do modelo ter sido criado e incluído no fluxo.

O nó de origem Arquivo de Estatísticas lê dados no formato IBM SPSS Statistics a partir do arquivo de dados *tree_credit.sav*, que é instalado na pasta *Demos*. (Uma variável especial denominada *$CLEO_DEMOS* é usada para referenciar essa pasta na instalação atual do IBM SPSS Modeler. Isso assegura que o caminho seja válido, independentemente da pasta de instalação ou da versão atual).

Figura 3. Lendo dados com um nó de origem Arquivo de Estatísticas

O nó Tipo especifica o **nível de medição** para cada campo. O nível de medição é uma categoria que indica o tipo de dados no campo. Nosso arquivo de dados de origem utiliza três níveis diferentes de medição.

Um campo **Contínuo** (como o campo *Idade*) contém valores numéricos contínuos, ao passo que um campo **Nominal** (como o campo *Classificação de crédito*) possui dois ou mais valores distintos, por exemplo, *Ruim*, *Bom* ou *Nenhum histórico de crédito*. Um campo **Ordinal** (como o campo *Nível de renda*) descreve os dados com diversos valores distintos que possuem uma ordem inerente – nesse caso *Baixo*, *Médio* e *Alto*.

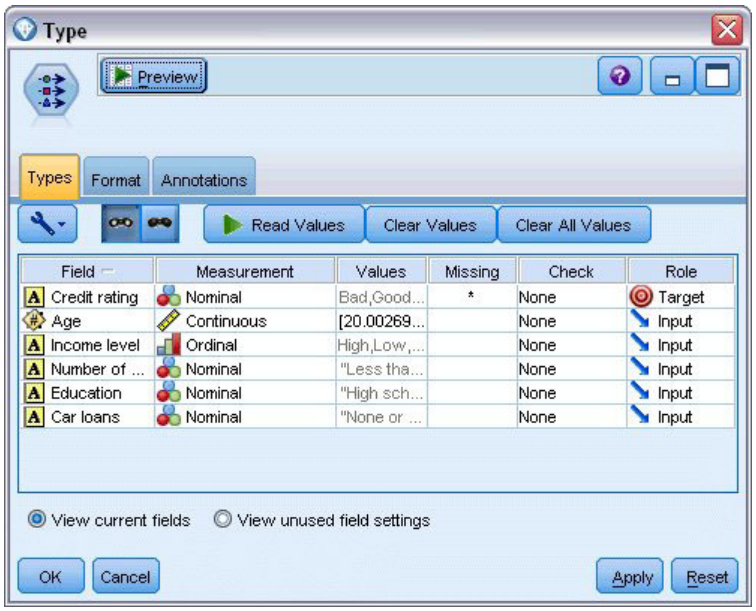


Figura 4. Configurando os campos de destino e de entrada com o nó Tipo

Para cada campo, o nó Tipo também especifica um **papel**, para indicar a função que cada campo atua na modelagem. O papel é configurado para *Resposta* para o campo *Classificação de crédito*, que é o campo que indica se um determinado cliente está inadimplente em um empréstimo ou não. Esta é a **resposta**, ou o campo para o qual queremos prever o valor.

O papel é configurado como *Entrada* para os outros campos. Os campos de entrada às vezes são conhecidos como **preditores** ou campos cujos valores são utilizados pelo algoritmo de modelagem para prever o valor do campo de destino.

O nó de modelagem CHAID gera o modelo.

Na guia Campos no nó de modelagem, a opção **Usar papéis predefinidos** é selecionada, o que significa que o destino e as entradas serão utilizados conforme especificado no nó Tipo. Poderíamos alterar os papéis do campo neste momento, mas para este exemplo, eles serão usados no estado em que se encontram.

1. Clique na guia Opções de Criação.

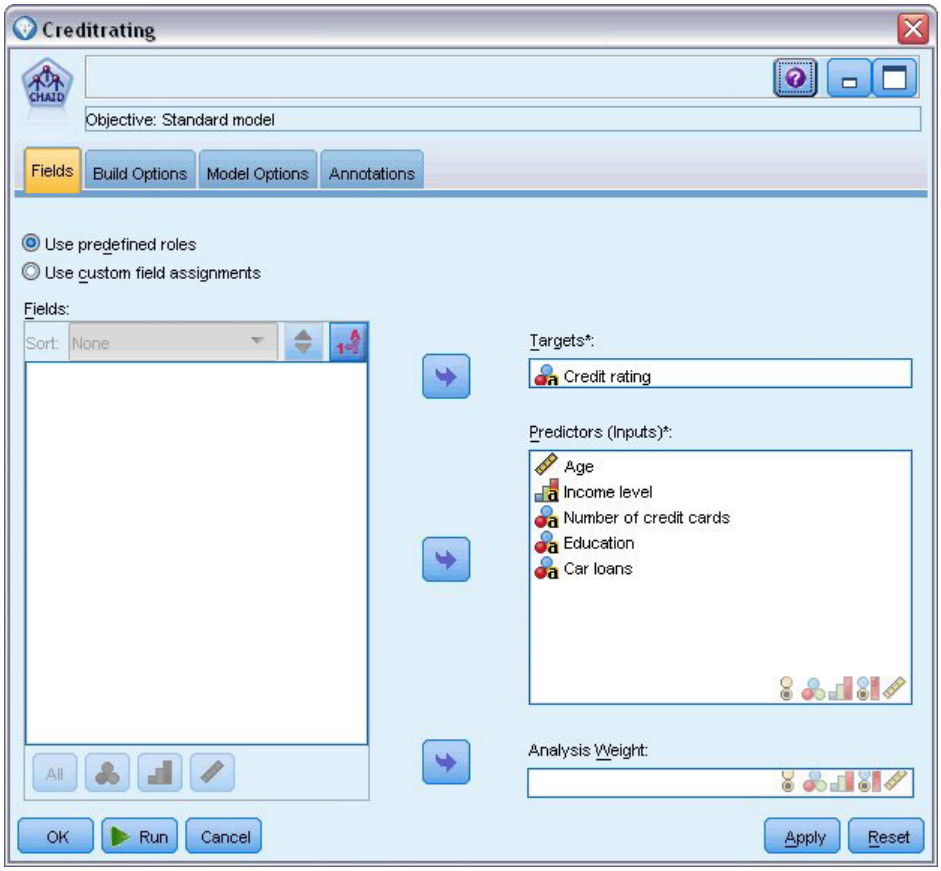


Figura 5. Nó de modelagem CHAID, guia Campos

Aqui há várias opções que permite especificar o tipo de modelo que queremos construir.

Como queremos um modelo novo, vamos usar a opção **Construir novo modelo** padrão.

Também queremos um único modelo de árvore de decisão padrão sem quaisquer aprimoramentos, portanto, manteremos também a opção objetiva padrão **Construir uma árvore única**.

Embora seja possível, opcionalmente, ativar uma sessão de modelagem interativa que permite fazer um ajuste preciso do modelo, este exemplo simplesmente gera um modelo utilizando a configuração de modo padrão **Gerar modelo**.

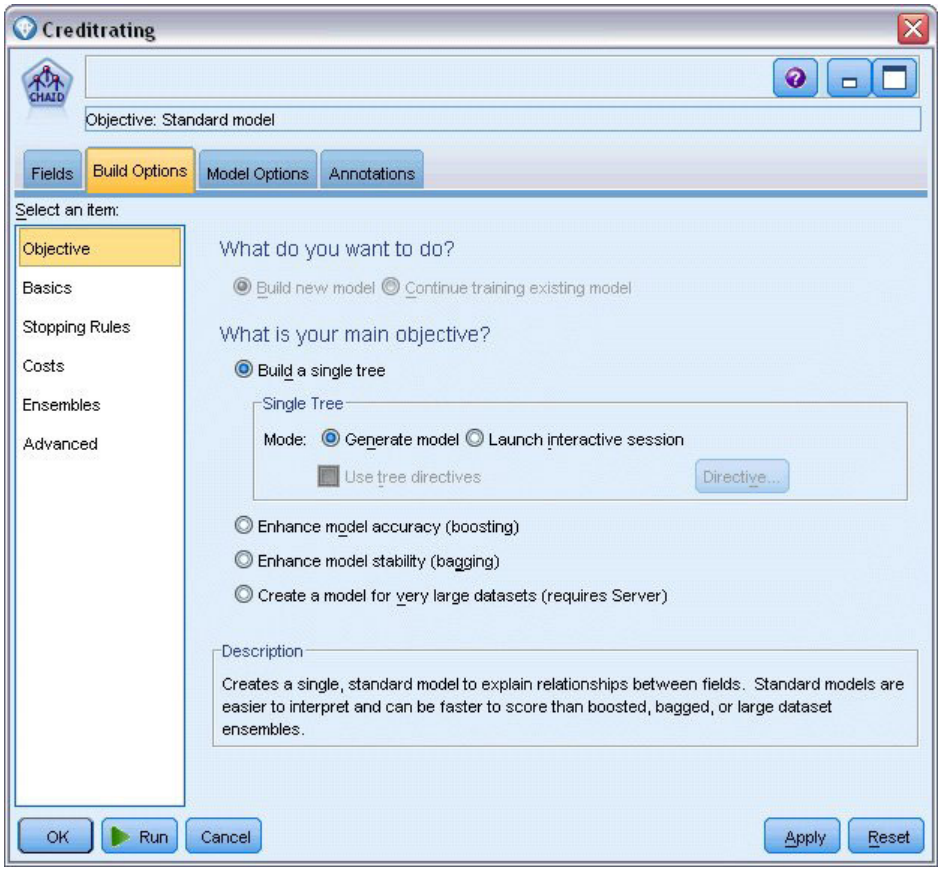


Figura 6. Nó de modelagem CHAID, guia Opções de Criação

Para esse exemplo, como queremos manter a árvore muito simples, vamos limitar o crescimento dela ao aumentar o número mínimo de casos para nós pais e filhos.

2. Na guia Opções de Criação, selecione **Regras de Parada** na área de janela do navegador à esquerda.
3. Selecione a opção **Usar valor absoluto**.
4. Configure o **Mínimo de registros na ramificação pai** para 400.
5. Configure o **Mínimo de registros na ramificação filha** para 200.

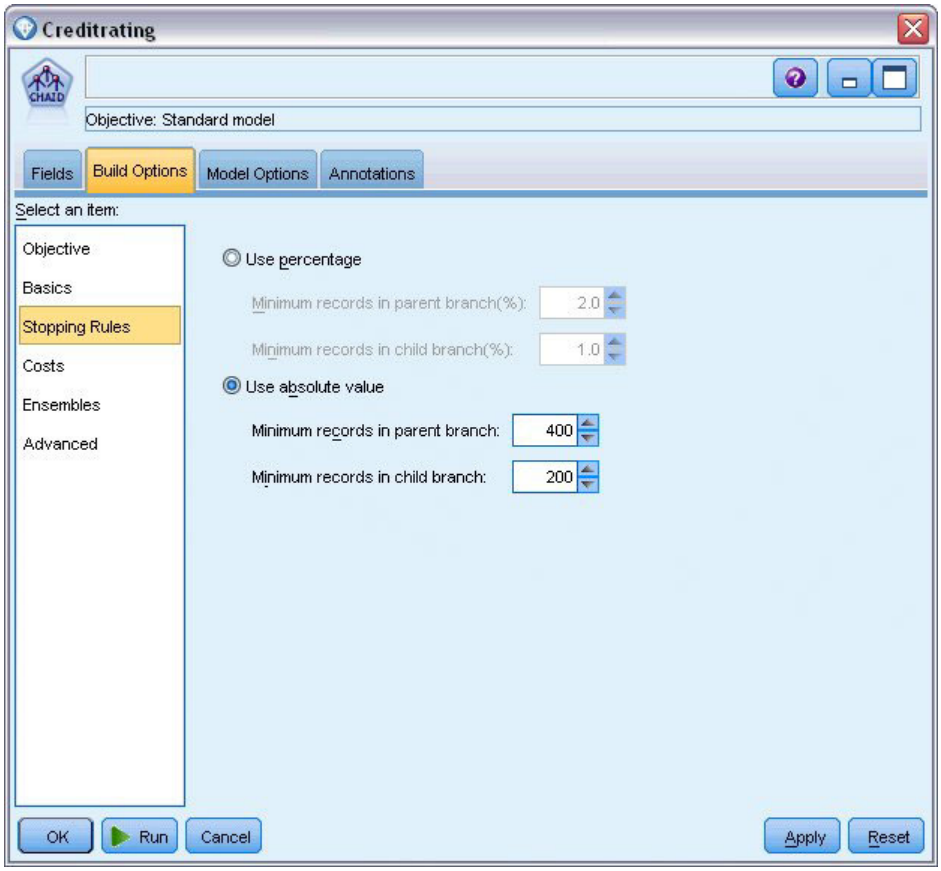


Figura 7. Configurando os critérios de parada para construção de árvore de decisão

Como é possível utilizar todas as outras opções padrão para este exemplo, clique em **Executar** para criar o modelo. (Como alternativa, clique com o botão direito no nó e escolha **Executar** no menu de contexto ou selecione o nó e escolha **Executar** no menu Ferramentas).

Procurando o Modelo

Quando a execução é concluída, o nugget do modelo é incluído na paleta Modelos no canto superior direito da janela do aplicativo, e também é colocado na tela de fluxo com uma ligação com o nó de modelagem a partir da qual ele foi criado. Para visualizar os detalhes do modelo, clique com o botão direito no nugget do modelo e escolha **Procurar** (na paleta de modelos) ou **Editar** (na tela).

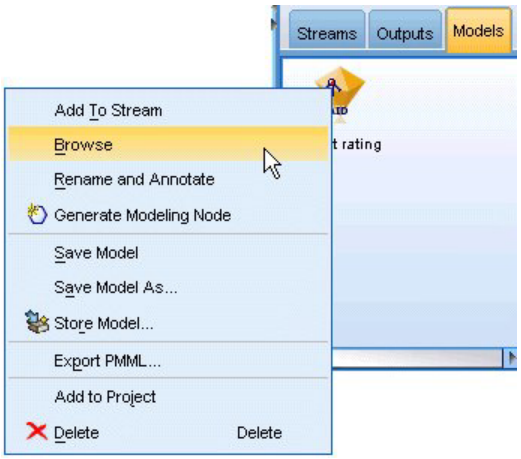


Figura 8. Paleta de Modelos

No caso do nugget CHAID, a guia Modelo exibe os detalhes na forma de um conjunto de regras -- essencialmente, uma série de regras que podem ser utilizadas para designar registros individuais para os nós filhos com base nos valores de diferentes campos de entrada.

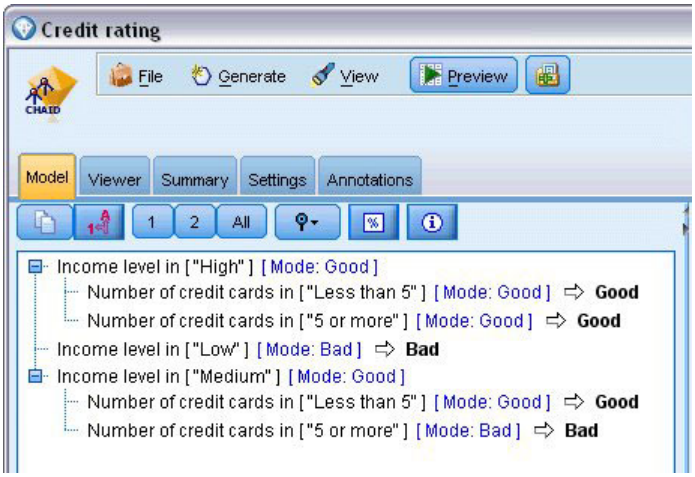


Figura 9. Nugget do modelo CHAID, conjunto de regras

Para cada nó terminal de árvore de decisão -- significando os nós de árvore que não são divididos ainda mais -- uma predição de *Bom* ou *Ruim* é retornada. Em cada caso, a predição é determinada pelo **Modo**, ou resposta mais comum, para registros que caírem nesse nó.

À direita do conjunto de regras, a guia Modelo exibe o gráfico Importância do Preditor, que mostra a importância relativa de cada preditor na estimativa do modelo. A partir disso, podemos ver que o *Nível de renda* é facilmente o mais significativo neste caso, e que o único outro fator significativo é *Número de cartões de crédito*.

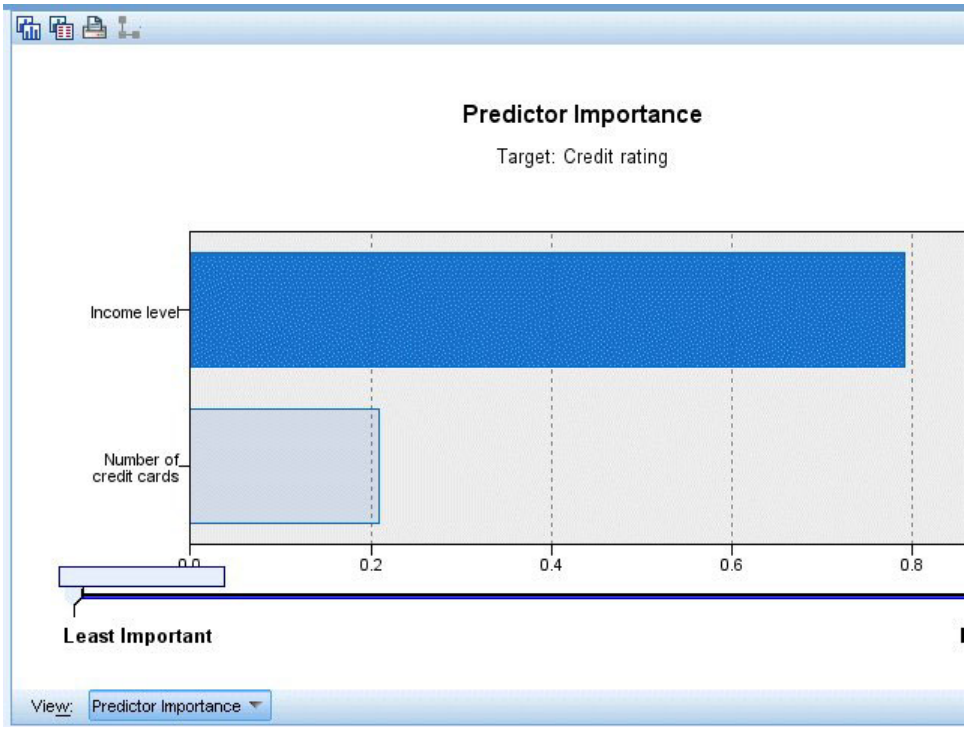


Figura 10. Gráfico de Importância do Preditor

A guia Visualizador no nugget do modelo exibe o mesmo modelo na forma de uma árvore, com um nó em cada ponto de decisão. Utilize os controles de Zoom na barra de ferramentas para aumentar o zoom em um nó específico ou diminuir o zoom para ver mais da árvore.

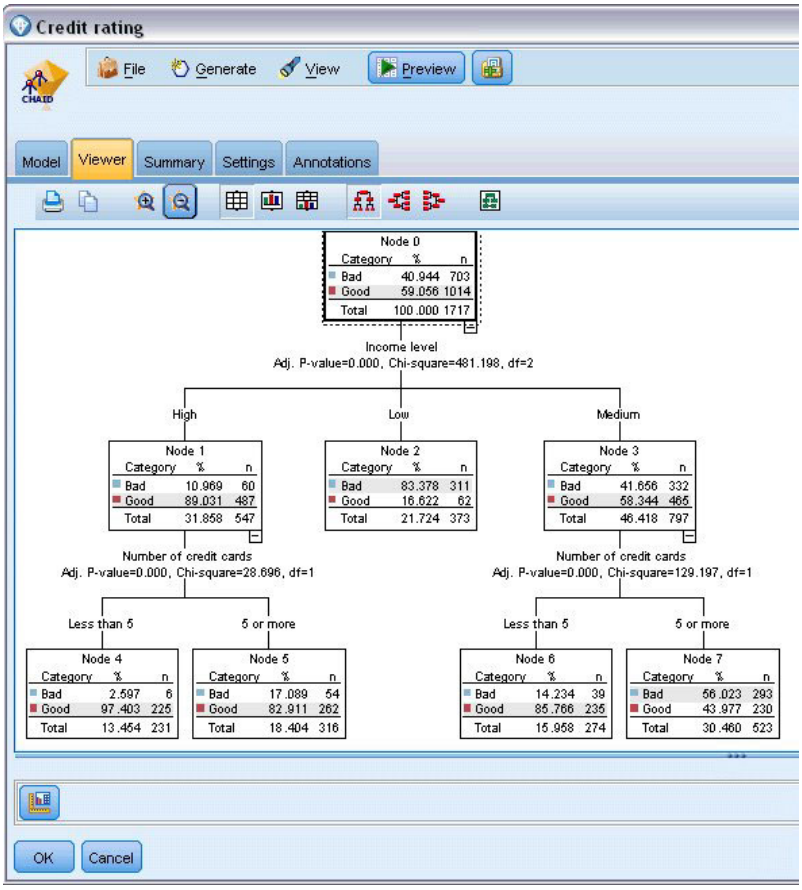


Figura 11. Guia Visualizador no nugget do modelo, com diminuir zoom selecionado

Observando a parte superior da árvore, o primeiro nó (Nó 0) fornece uma sumarização de todos os registros no conjunto de dados. Mais de 40% dos casos no conjunto de dados são classificados como um mau risco. Como esta é uma proporção muito alta, vamos verificar se a árvore pode dar dicas sobre quais fatores podem ser responsáveis.

Podemos ver que a primeira divisão é por *Nível de Renda*. Os registros em que o nível de renda estiver na categoria *Baixa* são designados ao Nó 2, e não é de surpreender que esta categoria contém a porcentagem mais alta de inadimplentes com empréstimos. É evidente que conceder empréstimos para clientes nesta categoria é altamente arriscado.

No entanto, como 16% dos clientes nesta categoria *não* estão realmente inadimplentes, a predição nem sempre estará correta. Nenhum modelo poderá prever de modo factível cada resposta, no entanto, um bom modelo deverá permitir prever a resposta *mais provável* para cada registro com base nos dados disponíveis.

Da mesma forma, se olharmos os clientes de alta renda (Nó 1), vemos que a grande maioria (89%) representa um bom risco. No entanto, mais de 1 a cada 10 desses clientes também estiveram inadimplentes. Nós podemos refinar os critérios de empréstimo para minimizar o risco aqui?

Observe como o modelo dividiu esses clientes em duas subcategorias (Nós 4 e 5), com base no número de cartões de crédito que eles possuem. Para clientes de alta renda, se um empréstimo for concedido somente para clientes com menos de 5 cartões de crédito, poderemos aumentar a taxa de sucesso de 89% para 97% -- que é um resultado ainda mais satisfatório.

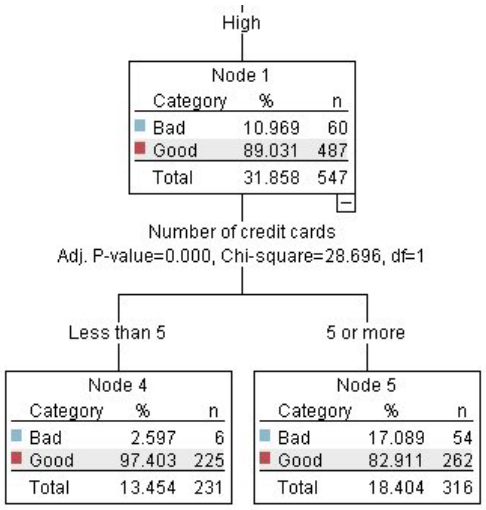


Figura 12. Visualização em árvore de clientes de alta renda

E quanto aos clientes na categoria Renda média (Nó 3)? Eles são muito mais igualmente divididos entre as classificações Bom e Ruim.

Mais uma vez, as subcategorias (Nós 6 e 7 nesse caso) podem nos ajudar. Desta vez, conceder empréstimo apenas para clientes com renda média com menos de 5 cartões de crédito aumenta a porcentagem de classificações Bom de 58% para 85%, que é uma melhoria significativa.

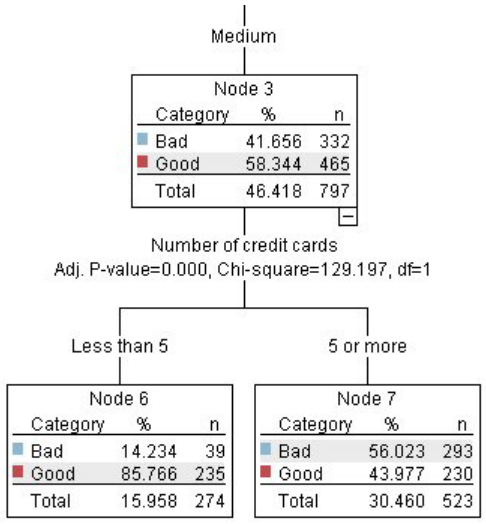


Figura 13. Visualização em árvore de clientes de renda média

Portanto, vimos que cada registro que é inserido nesse modelo é designado a um nó específico, e uma predição de *Bom* ou *Ruim* é designada com base na resposta mais comum para esse nó.

Este processo de designar predições para registros individuais é conhecido como **escoragem**. Ao escorar os mesmos registros usados para estimar o modelo, podemos avaliar precisamente como será o desempenho desse modelo nos dados de treinamento – os dados para os quais sabemos o resultado. Vamos ver como fazer isso.

Avaliando o Modelo

Nós temos procurado o modelo para entender como a escoragem funciona. Mas para avaliar *exatamente* como ela funciona, é necessário escorar alguns registros e comparar as respostas preditas pelo modelo com os resultados reais. Iremos escorar os mesmos registros que foram utilizados para estimar o modelo, permitindo comparar as respostas observadas e preditas.

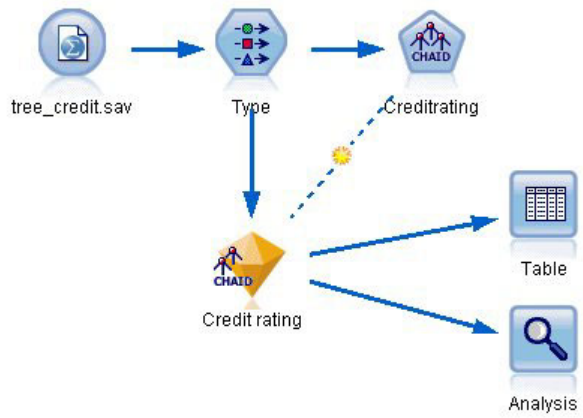
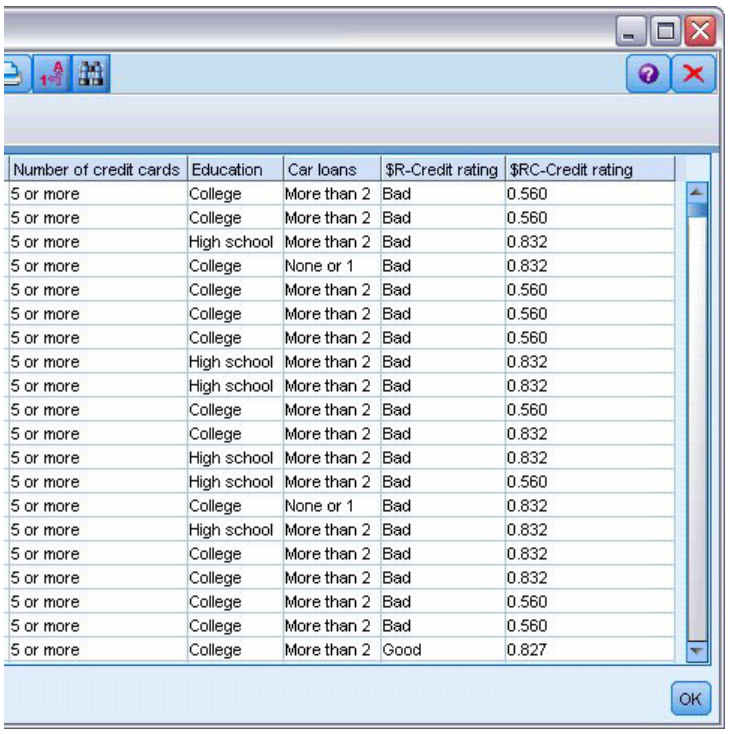


Figura 14. Anexado o nugget do modelo aos nós de saída para avaliação de modelo

1. Para ver as escoragens ou predições, anexe o nó Tabela ao nugget do modelo, clique duas vezes no nó Tabela e clique em **Executar**.

   A tabela exibe as escoragens preditas em um campo denominado *Classificação de $R-Credit*, que foi criado pelo modelo. É possível comparar esses valores com o campo *Classificação de crédito* original que contém as respostas reais.

   Por convenção, os nomes dos campos gerados durante a escoragem baseiam-se no campo de destino, mas com um prefixo padrão. Os prefixos *$G* e *$GE* são gerados pelo Modelo Linear Generalizado, *$R* é o prefixo utilizado para a predição gerada pelo modelo CHAID neste caso, *$RC* é para valores de confiança, *$X* é normalmente gerado utilizando uma combinação e *$XR*, *$XS* e *$XF* são utilizados como prefixos em casos em que o campo de destino é um campo Contínuo, Categórico, Conjunto ou de Sinalização, respectivamente. Tipos de modelo diferentes utilizam conjuntos diferentes de prefixos. Um **valor de confiança** é a estimação do próprio modelo do grau de precisão de cada valor predito, em uma escala de 0,0 a 1,0.



Number of credit cards	Education	Car loans	$R-Credit rating	$RC-Credit rating
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.560
5 or more	High school	More than 2	Bad	0.832
5 or more	College	None or 1	Bad	0.832
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.560
5 or more	High school	More than 2	Bad	0.832
5 or more	High school	More than 2	Bad	0.832
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.832
5 or more	High school	More than 2	Bad	0.832
5 or more	High school	More than 2	Bad	0.560
5 or more	College	None or 1	Bad	0.832
5 or more	High school	More than 2	Bad	0.832
5 or more	College	More than 2	Bad	0.832
5 or more	College	More than 2	Bad	0.832
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Good	0.827

OK

Figura 15. Tabela mostrando escoragens geradas e valores de confiança

Conforme esperado, o valor predito corresponde às respostas reais para muitos registros, mas não para todos. O motivo para isso é que cada nó terminal CHAID possui uma combinação de respostas. A predição corresponde à predição *mais comum*, mas estará errada para todas as outras nesse nó. (Lembre-se da minoria de 16% de clientes de baixa renda que não estiveram inadimplentes).

Para evitar isso, podemos continuar dividindo a árvore em ramificações cada vez menores, até que cada nó esteja 100% puro, ou seja, todos sendo *Bom* ou *Ruim* sem respostas combinadas. No entanto, esse modelo seria extremamente complicado e provavelmente não generalizaria tão bem para os demais conjuntos de dados.

Para descobrir exatamente quantas predições estão corretas, é possível ler a tabela e somar o número de registros nos quais o valor do campo predito *Classificação de $R-Credit* corresponde ao valor de *Classificação de crédito*. Felizmente, há outra maneira muito mais fácil que é utilizar o nó Análise que faz todo esse processo automaticamente.

2. Conecte o nugget do modelo ao nó Análise.
3. Clique duas vezes no nó Análise e clique em **Executar**.

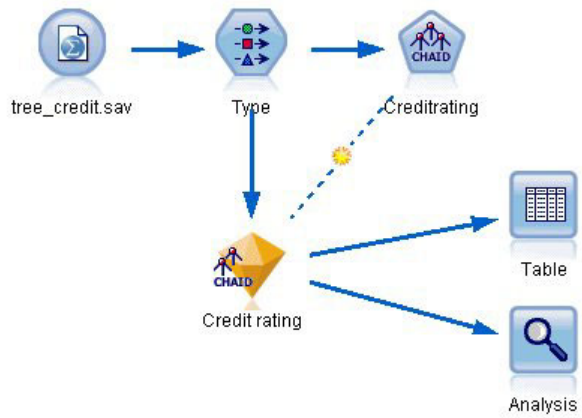


Figura 16. Anexando um nó Análise

A análise mostra que, para 1899 de 2464 registros -- mais de 77% -- o valor predito pelo modelo correspondeu à resposta real

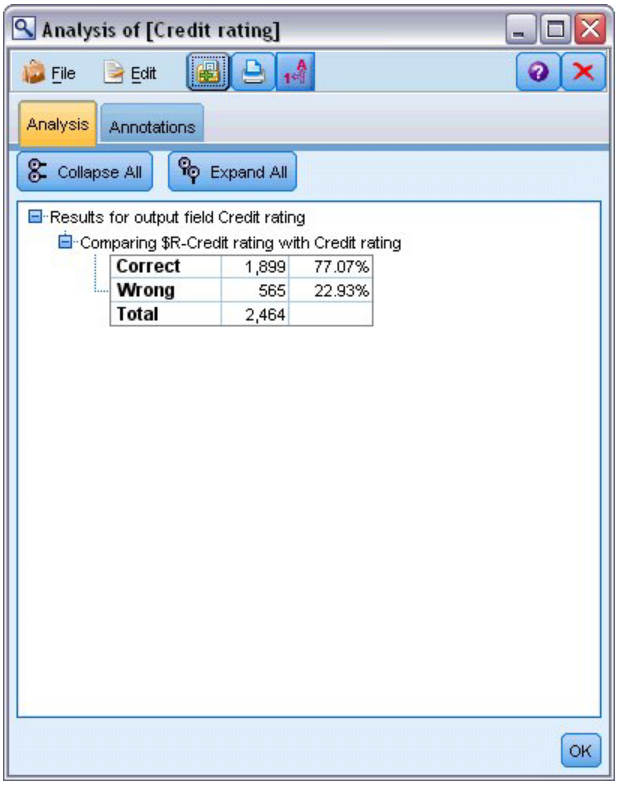


Figura 17. Resultados da análise comparando respostas observadas e preditas

Esse resultado é limitado pelo fato de que os registros que estão sendo escorados são os mesmos utilizados para estimar o modelo. Em uma situação real, é possível utilizar um nó Partição para dividir os dados em amostras separadas para treinamento e avaliação.

Ao utilizar uma partição de amostra para gerar o modelo e outra amostra para testá-lo, é possível obter uma indicação muito melhor do quão bem ele será generalizado para outros conjuntos de dados.

O nó Análise permite testar o modelo com relação aos registros para os quais nós já sabemos o resultado real. O próximo passo ilustra como é possível utilizar o modelo para escorar registros para os quais não

sabemos o resultado. Por exemplo, isso pode incluir pessoas que não forem atualmente clientes de um banco, mas que são possíveis alvos de receberem um email promocional.

Escorando Registros

Anteriormente, nós escoramos os mesmos registros utilizados para estimar o modelo para avaliar o nível de precisão do modelo. Agora vamos ver como escorar um conjunto diferente de registros a partir daqueles utilizados para criar o modelo. Este é o objetivo de modelagem com um campo de destino: Registros de estudo para os quais você sabe o resultado, para identificar padrões que permitirão prever resultados que você ainda não sabe.

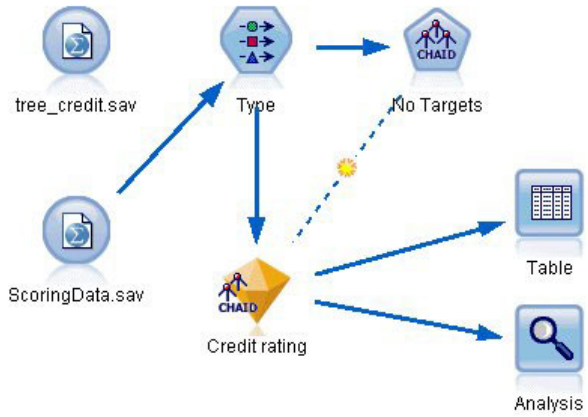


Figura 18. Anexando novos dados para escoragem

É possível atualizar o nó de origem Arquivo de Estatísticas para apontar para um arquivo de dados diferente, ou incluir um novo nó de origem que lê nos dados que você deseja escorar. De qualquer maneira, o novo conjunto de dados deverá conter os mesmos campos de entrada utilizados pelo modelo (*Idade*, *Nível de renda*, *Educação*, e assim por diante), mas não o campo de destino *Classificação de Crédito*.

Como alternativa, é possível incluir o nugget do modelo em qualquer fluxo que inclua os campos de entrada esperados. Independentemente de ler a partir de um arquivo ou de um banco de dados, o tipo de origem não importa desde que os nomes e tipos de campos correspondam aos utilizados pelo modelo.

Também é possível salvar o nugget do modelo como um arquivo separado, exportar o modelo no formato PMML para uso com outros aplicativos que suportam este formato ou armazenar o modelo em um repositório do IBM SPSS Collaboration and Deployment Services, que oferece implementação, escoragem e gerenciamento de modelos corporativos.

Independentemente da infraestrutura utilizada, o próprio modelo funciona da mesma maneira.

Sumarização

Este exemplo demonstra os passos básicos para criar, avaliar e escorar um modelo.

- O nó de modelagem estima o modelo ao estudar os registros para os quais o resultado é conhecido, e cria um nugget do modelo. Às vezes isso é referido como treinamento do modelo.
- O nugget do modelo pode ser incluído em qualquer fluxo com os campos esperados para escorar registros. Ao escorar os registros para os quais você já sabe o resultado (como clientes existentes), é possível avaliar o seu grau de desempenho.
- Quando estiver satisfeito com o desempenho do modelo, será possível escorar novos dados (como clientes esperados) para prever como eles responderão.

- Os dados usados para treinar ou estimar o modelo podem ser referidos como dados de analítica ou históricos, e os dados de escoragem também podem ser referidos como os dados operacionais.

Capítulo 3. Visão Geral de Modelagem

Visão Geral de Nós de Modelagem

O IBM SPSS Modeler oferece uma variedade de métodos de modelagem a partir do aprendizado de máquina, inteligência artificial e estatísticas. Os métodos disponíveis na paleta Modelagem permitem derivar informações novas a partir dos dados e desenvolver modelos preditivos. Cada método possui determinadas intensidades e é mais bem adequado para tipos de problemas específicos.

O *Guia de Aplicativos do IBM SPSS Modeler* fornece exemplos para muitos desses métodos, além de uma introdução geral ao processo de modelagem. Este guia está disponível como um tutorial online e também em formato PDF. Consulte o tópico "Exemplos de Aplicativos" na página 5 para obter informações adicionais.

Os métodos de modelagem são divididos em três categorias:

- Classificação
- Associação
- Segmentação

Modelos de Classificação

Modelos de classificação usam os valores de um ou mais campos de **entrada** para prever o valor de um ou mais campos de saída ou **resposta**. Alguns exemplos dessas técnicas são: árvores de decisão (algoritmo Árvore C e R, algoritmo de árvore estatística eficiente, rápido e imparcial, algoritmo Detector de Interação Automático Qui-Quadrado e C5.0), regressão (algoritmos linear, de logística, linear generalizado e algoritmo de regressão de Cox), redes neurais, Support Vector Machines e redes bayesianas.

Os modelos de classificação ajudam as organizações a preverem um resultado conhecido, como se um cliente irá comprar ou ir embora ou se uma transação se enquadra em um padrão conhecido de fraude. As técnicas de modelagem incluem aprendizado de máquina, indução de regra, identificação de subgrupo, métodos estatísticos e geração de diversos modelos.

Nós de classificação

O nó Classificador Automático cria e compara um número de modelos diferentes de resultados binários (sim ou não, rotatividade ou não rotatividade, e assim por diante), permitindo escolher a melhor abordagem para uma análise específica. Diversos algoritmos de modelagem são suportados, possibilitando selecionar os métodos que deseja utilizar, as opções específicas para cada um deles e os critérios para comparar os resultados. O nó gera um conjunto de modelos com base nas opções especificadas e classifica os melhores candidatos de acordo com os critérios que você especificar.

O nó Previsor Contínuo Automático estima e compara modelos de resultados de intervalo numérico contínuos utilizando um número de métodos diferentes. O nó funciona da mesma maneira que o nó Previsor Categórico Automático, permitindo escolher os algoritmos a serem utilizados e experimentá-los com as diversas combinações de opções em uma única passagem de modelagem. Os algoritmos suportados incluem redes neurais, Árvore C&R, CHAID, regressão linear, regressão linear generalizada e Support Vector Machines (SVMs). Os modelos podem ser comparados com base na correlação, no erro relativo ou no número de variáveis utilizadas.

O nó Árvore de Classificação e Regressão (C&R) gera uma árvore de decisão que permite prever ou classificar observações futuras. O método utiliza particionamento recursivo para dividir os registros de treinamento em segmentos ao minimizar a impureza em cada passo, em que um nó na árvore será considerado "puro" se 100% dos casos no nó caírem em uma categoria específica do campo de destino. Os campos de destino e de entrada podem ser intervalos numéricos ou categóricos (nominal, ordinal ou sinalizadores) e todas as divisões são binárias (somente dois subgrupos).

O nó QUEST fornece um método de classificação binária para construir árvores de decisão, projetado para reduzir o tempo de processamento necessário para grandes análises de Árvore C&R enquanto também reduz a tendência localizada nos métodos de árvore de classificação para favorecer entradas que permitem mais divisões. Os campos de entrada podem ser intervalos numéricos (conínuo), ao passo que o campo de destino deve ser categórico. Todas as divisões são binárias.

O nó CHAID gera árvores de decisão usando estatísticas qui-quadrado para identificar divisões ideais. Diferentemente dos nós Árvore C&R e QUEST, o CHAID pode gerar árvores não binárias, o que significa que algumas divisões possuem mais de duas ramificações. Os campos de destino e de entrada podem ser um intervalo numérico (contínuo) ou categóricos. Um Exhaustive CHAID é uma modificação de CHAID que executa uma tarefa mais completa de examinar todas as possíveis divisões, mas demora mais tempo para calcular.

O nó C5.0 constrói uma árvore de decisão ou um conjunto de regras. O modelo funciona dividindo a amostra com base no campo que fornece o ganho máximo de informações em cada nível. O campo de destino deve ser categórico. Diversas divisões em mais de dois subgrupos são permitidas.

O nó Lista de Decisão identifica os subgrupos ou segmentos, que mostram uma probabilidade maior ou menor de um resultado binário fornecido com relação à população geral. Por exemplo, é possível procurar por clientes que forem menos propensos a migrarem para o concorrente ou que responderão favoravelmente a uma campanha. É possível incorporar o conhecimento dos negócios no modelo ao incluir seus próprios segmentos customizados e visualizar modelos alternativos lado a lado para comparar os resultados. Os modelos de Lista de Decisão consistem em uma lista de regras em que cada regra possui uma condição e um resultado. As regras são aplicadas na ordem, e a primeira regra que corresponder determina o resultado.

Os modelos de regressão lineares preveem uma variável resposta contínua com base em relacionamentos lineares entre o destino e um ou mais preditores.

O nó PCA/Fator fornece técnicas poderosas de redução de dados para reduzir a complexidade de seus dados. A análise de componentes principais (PCA) localiza combinações lineares dos campos de entrada que executam a melhor tarefa de capturar a variância no conjunto inteiro de campos, em que os componentes são ortogonais (perpendiculares) entre si. A análise fatorial tenta identificar fatores subjacentes que explicam o padrão de correlações dentro de um conjunto de campos observados. Para ambas as abordagens, o objetivo é encontrar um número pequeno de campos derivados que efetivamente sumarizem as informações no conjunto de campos original.

O nó Seleção de Variável verifica campos de entrada para remoção com base em um conjunto de critérios (como a porcentagem de valores omissos) e, em seguida, classifica a importância das entradas restantes com relação a um destino especificado. Por exemplo, dado um conjunto de dados com centenas de possíveis entradas, quais delas mais poderão ser úteis para modelar os resultados do paciente?

A análise discriminante faz suposições mais rígidas do que a regressão logística, mas pode ser uma alternativa ou um complemento poderoso para uma análise de regressão logística quando essas suposições forem atendidas.

A regressão logística é uma técnica estatística para classificar registros com base em valores de campos de entrada. Ela é semelhante a uma regressão linear, mas usa um campo de destino categórico ao invés de um intervalo numérico.

O modelo Linear Generalizado expande o modelo linear geral para que a variável dependente seja linearmente relacionada aos fatores e covariáveis por meio de uma função de ligação especificada. Além disso, o modelo permite à variável dependente ter uma distribuição não normal. Ele cobre a funcionalidade de um grande número de modelos estatísticos, incluindo regressão linear, regressão logística, modelos de log-linear para dados de contagem e modelos de sobrevivência censurados por intervalo.

Um modelo linear generalizado misto (GLMM) estende o modelo linear para que o destino possa ter uma distribuição não normal, esteja linearmente relacionado aos fatores e covariáveis por meio de uma função de ligação especificada e para que as observações possam ser correlacionadas. Os modelos lineares generalizados mistos abrangem uma ampla variedade de modelos, desde regressão linear simples até modelos multiníveis complexos para dados de longitude não normais.

O nó Regressão de Cox permite construir um modelo de sobrevivência para dados de sobrevivência na presença de registros censurados. O modelo produz uma função de sobrevivência que prediz a probabilidade de que o evento de interesse tenha ocorrido em um determinado momento (*t*) de acordo com os valores fornecidos para as variáveis de entrada.

O nó Support Vector Machine (SVM) permite classificar dados em um dos dois grupos sem super ajuste. O SVM funciona bem com conjuntos de dados grandes, como aqueles com um número muito grande de campos de entrada.

O nó Rede Bayesiana permite construir um modelo de probabilidade ao combinar evidências observada e registrada com conhecimento do mundo real para estabelecer a probabilidade das ocorrências. O nó foca nas redes Tree Augmented Naïve Bayes (TAN) e Markov Blanket que são utilizadas principalmente para classificação.

O nó Self-Learning Response Model (SLRM) permite construir um modelo no qual um novo caso único, ou um pequeno número de casos novos, pode ser usado para estimar novamente o modelo sem precisar treinar novamente esse modelo usando todos os dados.

O nó Séries Temporais estima modelos Média Móvel Integrada AutoRegressiva (ARIMA) univariados de suavização exponencial e modelos ARIMA multivariados (ou de função de transferência) para os dados de séries temporais e produz previsões de desempenho futuro. Um nó Séries Temporais deve sempre ser precedido por um nó Intervalos de Tempo.

O nó *k*-Nearest Neighbor (KNN) associa um novo caso à categoria ou valor dos *k* objetos mais próximos a ele no espaço do preditor, em que *k* é um número inteiro. Os casos semelhantes estão próximos uns dos outros e os casos dissimilares estão distantes.

O nó Spatio-Temporal Prediction (STP) utiliza dados que contêm dados do local, campos de entrada para predição (preditores), um campo de tempo e um campo de destino. Cada local possui várias linhas nos dados que representam os valores de cada preditor em cada momento da medição. Após os dados serem analisados, eles podem ser usados para prever valores de destino em qualquer local dentro dos dados de formato que são usados na análise.

Modelos de Associação

Os *modelos de associação* localizam padrões nos seus dados nos quais uma ou mais entidades (como eventos, compras ou atributos) estão associadas a uma ou mais outras entidades. Os modelos constroem conjuntos de regras que definem esses relacionamentos. Aqui, os campos nos dados podem agir como entradas e destinos. É possível localizar essas associações manualmente, mas os algoritmos de regra de associação fazem isso muito mais rápido e podem explorar padrões mais complexos. Os modelos a priori e Carma são exemplos do uso de tais algoritmos. Outro tipo de modelo de associação é um modelo de detecção de sequência, que localiza os padrões sequenciais em dados estruturados por tempo.

Os modelos de associação são mais úteis quando prever diversos resultados, por exemplo, clientes que compraram um produto X também compraram Y e Z. Os modelos de associação associam uma conclusão específica (como a decisão de comprar algo) a um conjunto de condições. A vantagem dos algoritmos de regra de associação sobre os algoritmos de árvore de decisão mais padrão (C5.0 e C&RT) é que as associações podem existir entre qualquer um dos atributos. Um algoritmo de árvore de decisão construirá regras com apenas uma única conclusão, ao passo que os algoritmos de associação tentam localizar muitas regras, cada qual podendo ter uma conclusão diferente.

Nós de associação

O nó a priori extrai um conjunto de regras a partir dos dados, retirando as regras com o conteúdo mais alto de informações. O a priori oferece cinco métodos diferentes de selecionar regras e utiliza um esquema de indexação sofisticado para processar grandes conjuntos de dados de maneira eficiente. Para grandes problemas, o a priori geralmente é mais rápido para treinar e, além disso, ele não possui um limite arbitrário quanto ao número de regras que podem ser retidas e pode manipular regras com até 32 condições prévias. O a priori requer que todos os campos de entrada e de saída sejam categóricos e oferece melhor desempenho porque é otimizado para este tipo de dados.

O modelo do CARMA extrai um conjunto de regras de dados sem a necessidade de especificar campos de entrada ou de destino. Ao contrário do a priori o nó CARMA oferece configurações de construção para suporte de regra (suporte para ambos antecedente e subsequente) ao invés de apenas o suporte de antecedente. Isso significa que as regras geradas podem ser utilizadas para uma variedade maior de aplicativos, por exemplo, para localizar uma lista de produtos ou serviços (antecedentes) cujo subsequente é o item que você deseja promover nesta temporada de férias.

O nó Sequência descobre as regras de associação nos dados sequenciais ou orientados a tempo. Uma sequência é uma lista de conjuntos de itens que tendem a ocorrer em uma ordem previsível. Por exemplo, um cliente que compra uma lâmina de barbear e uma loção pós-barba poderá comprar um creme de barbear na próxima compra. O nó Sequência baseia-se no algoritmo de regras de associação do CARMA que utiliza um método de duas passagens eficiente para localizar sequências.

O nó Regras de Associação é semelhante ao nó a priori, no entanto, ao contrário do a priori, o nó Regras de Associação pode processar dados da lista. Além disso, o nó Regras de Associação poderá ser utilizado com o IBM SPSS Analytic Server para processar Big Data e aproveitar processamento paralelo mais rápido.

Modelos de Segmentação

Modelos de segmentação dividem os dados em segmentos, ou clusters, de registros com padrões semelhantes de campos de entrada. Como eles estão interessados somente nos campos de entrada, os modelos de segmentação não têm um conceito de campos de saída ou resposta. Os exemplos de modelos de segmentação são redes Kohonen, armazenamento em cluster de k-médias, clusterização em dois passos e detecção de anomalias.

Os modelos de segmentação (também conhecidos como "modelos de armazenamento em cluster") são úteis nos casos em que o resultado específico é desconhecido (por exemplo, ao identificar novos padrões de fraude ou ao identificar grupos de interesse em sua base de clientes). Os modelos de armazenamento em cluster focam na identificação de grupos de registros semelhantes e na rotulagem dos registros de acordo com o grupo ao qual eles pertencem. Isso é feito sem o benefício de conhecer previamente os grupos e suas características e também o que diferencia os modelos de armazenamento em cluster de outras técnicas de modelagem por não haver um campo de destino ou de saída predefinido para o modelo prever. Não há respostas certas ou erradas para esses modelos. Seus valores são determinados pela capacidade de capturar agrupamentos de interesse nos dados e de fornecer descrições úteis desses agrupamentos. Os modelos de armazenamento geralmente são usados para criar clusters ou segmentos, que são, então, usados como entradas em análises subsequentes (por exemplo, segmentação de possíveis clientes em subgrupos homogêneos).

Nós de segmentação

O nó Cluster Automático estima e compara modelos de armazenamento em cluster que identificam grupos de registros que possuem características semelhantes. O nó funciona da mesma maneira que outros nós de modelagem automatizados, permitindo experimentá-los com as diversas combinações de opções em uma única passagem de modelagem. Os modelos podem ser comparados utilizando medidas básicas com as quais é possível tentar filtrar e classificar a utilidade dos modelos de cluster e fornecer uma medida com base na importância de campos específicos.

O nó K-Médias armazena em cluster o conjunto de dados em grupos distintos (ou clusters). O método define um número fixo de clusters, designa iterativamente registros para clusters e ajusta os centros do cluster até que o refinamento adicional não consiga mais melhorar o modelo. Ao invés de tentar prever um resultado, o *k-médias* utiliza um processo conhecido como aprendizado não supervisionado para descobrir padrões no conjunto de campos de entrada.

O nó Kohonen gera um tipo de rede neural que pode ser utilizado para armazenar em cluster o conjunto de dados em grupos distintos. Quando a rede for totalmente treinada, os registros similares deverão estar próximos no mapa de saída, ao passo que registros que forem diferentes estarão distantes. É possível examinar o número de observações capturadas por cada unidade no nugget do modelo para identificar as unidades fortes. Isto poderá dar uma ideia do número apropriado de clusters.

O nó TwoStep utiliza um método de clusterização em dois passos. O primeiro passo faz uma única passagem através dos dados para compactar os dados de entrada brutos em um conjunto gerenciável de subclusters. O segundo passo usa um método de armazenamento em cluster hierárquico para mesclar progressivamente os subclusters em clusters cada vez maiores. O TwoStep tem a vantagem de estimar automaticamente o número ideal de clusters para os dados de treinamento. Ele pode manipular tipos de campo combinados e grandes conjuntos de dados de maneira eficiente.

O nó Detecção de Anomalia identifica casos incomuns, ou valores discrepantes, que não estiverem em conformidade com os padrões de dados "normais". Com este nó, é possível identificar valores discrepantes, mesmo se eles não se ajustarem a nenhum dos padrões anteriormente conhecidos e mesmo se você não souber exatamente o que procura.

Modelos de mineração dentro da base de dados

O IBM SPSS Modeler suporta integração com ferramentas de mineração e modelagem de dados que estiverem disponíveis a partir de fornecedores de banco de dados, incluindo o Oracle Data Miner, o IBM DB2 InfoSphere Warehouse e o Microsoft Analysis Services. É possível construir, escorar e armazenar modelos dentro do banco de dados - tudo isso no aplicativo IBM SPSS Modeler. Para obter detalhes completos, consulte o *Guia de Mineração Dentro da Base de Dados do IBM SPSS Modeler* , disponível no produto DVD.

IBM SPSS StatisticsModelos

Se você tiver uma cópia do IBM SPSS Statistics instalada e licenciada em seu computador, será possível acessar e executar determinadas rotinas do IBM SPSS Statistics a partir do IBM SPSS Modeler para construir e escorar os modelos.

Informações Adicionais

Uma documentação detalhada sobre os algoritmos de modelagem também está disponível. Para obter mais informações, consulte o *Guia de Algoritmos do IBM SPSS Modeler*, disponível no produto DVD.

Construindo Modelos de Divisão

A modelagem de divisão permite utilizar um único fluxo para construir modelos separados para cada valor possível de um campo de entrada flag, nominal ou contínuo, com todos os modelos resultantes podendo ser acessados a partir de um nugget do modelo único. Os valores possíveis para os campos de entrada podem ter efeitos muito diferentes no modelo. Com a modelagem de divisão, é possível construir facilmente o modelo de melhor ajuste para cada valor de campo possível em uma única execução do fluxo.

Observe que as sessões de modelagem interativa não podem utilizar divisão. Como a modelagem interativa permite especificar cada modelo individualmente, não é vantagem utilizar a divisão que constrói diversos modelos automaticamente.

A modelagem de divisão funciona ao designar um campo de entrada específico como um campo de divisão. Isso poderá ser feito ao configurar o papel do campo para **Divisão** na especificação de Tipo.

É possível designar apenas os campos com um nível de medição de **Flag**, **Nominal**, **Ordinal** ou **Contínuo** como campos de divisão.

É possível designar mais de um campo de entrada como um campo de divisão. Neste caso, porém, o número de modelos criados poderá aumentar enormemente. Um modelo é construído para cada combinação possível dos valores dos campos de divisão selecionados. Por exemplo, se três campos de entrada, cada um tendo três valores possíveis, forem designados como campos de divisão, isto resultará na criação de 27 modelos diferentes.

Mesmo após designar um ou mais campos como campos de divisão, ainda é possível escolher se deseja criar modelos de divisão ou um modelo único, por meio de uma configuração de caixa de seleção no diálogo de nó de modelagem.

Se os campos de divisão estiverem definidos, mas a caixa de seleção não for marcada, apenas um único modelo será gerado. Da mesma forma, se a caixa de seleção estiver marcada, mas nenhum campo de divisão for definido, a divisão será ignorada e um modelo único será gerado.

Ao executar o fluxo, os modelos separados são construídos atrás dos cenários para cada valor possível do campo ou campos de divisão, mas apenas um único nugget do modelo será colocado na paleta de modelos e na tela de fluxo. Um nugget de modelo de divisão é indicado pelo símbolo de divisão, que são dois retângulos cinzas sobrepostos na imagem do nugget.

Ao procurar pelo nugget do modelo de divisão, você verá uma lista de todos os modelos separados que foram construídos.

É possível investigar um modelo individual a partir de uma lista clicando duas vezes no ícone do nugget no visualizador. Fazer isso abre uma janela do navegador padrão para o modelo individual. Quando o nugget estiver na tela, clicar duas vezes em uma miniatura de gráfico abre o gráfico em tamanho integral. Consulte o tópico “Visualizador de Modelo de Divisão” na página 48 para obter mais informações.

Depois que um modelo tiver sido criado como um modelo de divisão, não será possível remover o processamento de divisão, nem desfazer a divisão posteriormente a partir de um nó ou nugget de modelagem de divisão.

Exemplo. Um varejista nacional deseja estimar as vendas por categoria do produto em cada uma de suas lojas no país. Usando a modelagem de divisão, ele designa o campo Loja de seus dados de entrada como um campo de divisão, permitindo construir modelos separados para cada categoria em cada loja em uma única operação. Em seguida, ele pode usar as informações resultantes para controlar os níveis de estoque de modo muito mais preciso do que faria com apenas um único modelo.

Dividindo e Particionando

A divisão possui alguns recursos em comum com o particionamento, mas são usados de maneiras muito diferentes.

O **Particionamento** divide o conjunto de dados aleatoriamente em duas ou três partes: treinamento, teste e (opcionalmente) validação, e é utilizado para testar o desempenho de um modelo único.

A **Divisão** divide o conjunto de dados em tantas partes quanto houver valores possíveis para um campo de divisão, e é usada para construir diversos modelos.

O particionamento e a divisão operam de modo totalmente independente um do outro. É possível escolher qualquer um deles, ambos ou nenhum em um nó de modelagem.

Nós de Modelagem que Suportam Modelos de Divisão

Um número de nós de modelagem pode criar modelos de divisão. As exceções são Cluster Automático, Séries Temporais, PCA/Fator, Seleção de Variáveis, SLRM, os modelos de associação (a priori, Carma e Sequência), os modelos de armazenamento em cluster (K-Médias, Kohonen, TwoStep e Anomalia), Modelo do Statistics e os nós utilizados para modelagem dentro da base de dados.

Os nós de modelagem que suportam modelagem de divisão são:

Árvore C&R

Rede Bayesiana

QUEST

GenLin

CHAID

KNN

C5.0

Cox

Rede neural

Classificador Automático

Lista de Decisão

Numeração Automática

Regressão

Logística

Discriminante

SVM

Variáveis Afetadas pela Divisão

O uso de modelos de divisão afeta diversas variáveis do IBM SPSS Modeler de várias formas. Esta seção fornece orientação sobre como utilizar os modelos de divisão em conjunto com outros nós em um fluxo.

Nós Operações de Registro

Ao utilizar modelos de divisão em um fluxo que contém um nó **Amostra**, estratifique os registros pelo campo de divisão para atingir uma amostragem igual de registros. Essa opção está disponível quando escolher Complexo como o método de amostra.

Se o fluxo contiver um nó **Balanceamento**, observe que o balanceamento se aplicará ao conjunto geral de registros de entrada, e não ao subconjunto de registros dentro de uma divisão.

Ao agregar registros por meio de um nó **Agregar**, configure os campos de divisão para serem campos-chave, se desejar calcular agregados para cada divisão.

Nós Operações de Campo

O nó **Tipo** é onde você especifica qual campo ou campos serão utilizados como campos de divisão.

Observe que, embora o nó **Combinação** seja utilizado para combinar dois ou mais nuggets do modelo, ele não pode ser utilizado para reverter a ação de divisão, já que os modelos de divisão estão contidos em um nugget do modelo único.

Nós de modelagem

Os modelos de divisão não suportam o cálculo da importância do preditor (a importância relativa dos campos de entrada do preditor na estimativa do modelo). As configurações da importância do preditor são ignoradas durante a construção de modelos de divisão.

O nó **KNN** (vizinho mais próximo) suportará os modelos de divisão somente se ele estiver configurado para prever um campo de destino. A configuração alternativa (apenas identificar vizinhos mais próximos) não cria um modelo. Se a opção "Selecionar automaticamente k" for escolhida, cada um dos modelos de divisão poderá ter um número diferente de vizinhos mais próximos. Assim, o número de colunas geradas em um modelo global será igual ao maior número de vizinhos mais próximos localizados em todos os modelos de divisão. Para os modelos de divisão em que o número de vizinhos mais próximos é menor que esse máximo, haverá um número correspondente de colunas preenchidas com valores `$null`. Consulte o tópico “Nó KNN” na página 331 para obter mais informações.

Nós de modelagem da base de dados

Os nós de modelagem dentro da base de dados não suportam modelos de divisão.

Nuggets do modelo

Exportar para o PMML a partir de um nugget do modelo de divisão não é possível porque o nugget contém diversos modelos e o PMML não suporta esse empacotamento. No entanto, é possível exportar em HTML ou em texto.

Opções de Campos do Nó de Modelagem

Todos os nós de modelagem possuem uma guia Campos, na qual é possível especificar os campos a serem utilizados na construção do modelo.

Antes de poder construir um modelo, é necessário especificar quais campos você deseja utilizar como destinos e como entradas. Com algumas exceções, todos os nós de modelagem utilizarão as informações de campo a partir de um nó Tipo de envio de dados. Se você estiver utilizando um nó Tipo para selecionar campos de entrada e de destino, não será necessário alterar nada nesta guia. (As exceções incluem o nó Sequência e o nó Extração de Texto, que requerem que as configurações de campo sejam especificadas no nó de modelagem).

Usar configurações do nó de tipo. Essa opção instrui o nó a utilizar as informações de campo a partir de um nó Tipo de envio de dados. Esse é o padrão.

Usar configurações customizadas. Essa opção instrui o nó a utilizar as informações de campo especificadas aqui ao invés das informações fornecidas em qualquer nó ou nós Tipo de envio de dados. Após selecionar esta opção, especifique os campos abaixo conforme necessário.

Nota: nem todos os campos são exibidos para todos os nós.

- **Usar formato transacional (apenas nós a priori, CARMA, Regras de Associação da MS e Oracle a priori).** Marque esta caixa de seleção se os dados de origem estiverem em **transacional formato**. Os registros nesse formato possuem dois campos, um para ID e outro para conteúdo. Cada registro representa uma única transação ou item, e os itens associados estão vinculados por terem o mesmo ID. Desmarque essa caixa se os dados estiverem em **formato tabular**, em que os itens são representados por flags separados, cada campo de flag representa a presença ou ausência de um item específico e cada registro representa um conjunto completo de itens associados. Consulte o tópico "Dados Tabulares versus Transacionais" na página 246 para obter mais informações.
  - **ID.** Para dados transacionais, selecione um campo de ID na lista. Campos numéricos ou simbólicos podem ser utilizados como o campo de ID. Cada valor exclusivo deste campo deve indicar uma unidade específica de análise. Por exemplo, em um aplicativo de cesta de mercado, cada ID pode representar um cliente único. Para um aplicativo de análise de log da web, cada ID pode representar um computador (pelo endereço IP) ou um usuário (pelos dados de login).
  - **IDs são contíguos.** (apenas nós a priori e CARMA) Se seus dados estiverem pré-ordenados de forma que todos os registros com o mesmo ID sejam agrupados no fluxo de dados, selecione esta opção para acelerar o processamento. Se seus dados não estiverem pré-ordenados (ou se você não tiver certeza), deixe essa opção desmarcada e o nó ordenará os dados automaticamente.

    Nota: se seus dados não estiverem classificados e você selecionar essa opção, resultados inválidos poderão ser obtidos no seu modelo.
  - **Conteúdo.** Especifique um ou mais campos de conteúdo para o modelo. Esses campos contêm os itens de interesse na modelagem de associação. É possível especificar diversos campos de sinalização (se os dados estiverem em formato tabular) ou um campo nominal único (se os dados estiverem em formato transacional).
- **Resposta.** Para modelos que requerem um ou mais campos de destino, selecione os campos de destino. Isso é semelhante a configurar o papel do campo para *Destino* em um nó Tipo.
- **Avaliação.** (apenas para modelos de Cluster Automático). Nenhum destino é especificado para modelos de cluster, no entanto, é possível selecionar um campo de avaliação para identificar seu nível de importância. Além disso, é possível avaliar o quão bem os clusters diferenciam os valores deste campo que, por sua vez, indica se os clusters podem ser utilizados para prever este campo. *Nota* o campo de avaliação deve ser uma sequência com mais de um valor.
  - **Entradas.** Selecione um ou mais campos de entrada. Isso é semelhante a configurar o papel do campo para *Entrada* em um nó Tipo.
  - **Partição.** Este campo permite especificar um campo utilizado para particionar os dados em amostras separadas para os estágios de treinamento, de teste e de validação de construção de modelo. Ao utilizar uma amostra para gerar o modelo e uma amostra diferente para testá-lo, é possível obter uma boa indicação do quão bem o modelo será generalizado para conjuntos de dados maiores que forem semelhantes aos dados atuais. Se diversos campos de partição tiverem sido definidos usando os nós Tipo ou Partição, um campo de partição único deverá ser selecionado na guia Campos em cada nó de modelagem que utiliza particionamento. (Se apenas uma partição estiver presente, ela será utilizada automaticamente sempre que o particionamento estiver ativado). Além disso, observe que para aplicar a partição selecionada à sua análise, o particionamento também deverá ser ativado na guia Opções de Modelo para o nó. (Desmarcar esta opção permite desativar o particionamento sem alterar as configurações do campo).
- **Divisões.** Para os modelos de divisão, selecione o campo ou campos de divisão. Isso é semelhante a configurar o papel do campo para *Divisão* em um nó Tipo. É possível designar apenas os campos com um nível de medição de **Flag**, **Nominal**, **Ordinal** ou **Contínuo** como campos de divisão. Os campos escolhidos como campos de divisão não podem ser utilizados como campos de destino, de entrada, de partição, de frequência ou de ponderação. Consulte o tópico "Construindo Modelos de Divisão" na página 28 para obter mais informações.
- **Usar campo de frequência.** Esta opção permite selecionar um campo como uma ponderação de frequência. Use esta opção se os registros em seus dados de treinamento representarem mais de uma unidade cada, por exemplo, se estiver usando dados agregados. Os valores do campo devem ser o

número de unidades representadas por cada registro. Consulte o tópico "Usando Campos de Frequência e de Ponderação" para obter mais informações.

Nota: se a mensagem de erro **Metadados (em campos de entrada/saída) inválidos** for exibida, assegure-se de ter especificado todos os campos que forem necessários, como o campo de frequência.

- **Usar campo de ponderação.** Esta opção permite selecionar um campo como uma ponderação de caso. As ponderações de caso são utilizadas para considerar as diferenças na variância nos níveis do campo de saída. Consulte o tópico "Usando Campos de Frequência e de Ponderação" para obter mais informações.
- **Subsequentes.** Para nós de indução de regra (a priori), selecione os campos a serem usados como subsequentes no conjunto de regras resultante. (Isso corresponde aos campos com papel *Destino* ou *Ambos* em um nó Tipo).
- **Antecedentes.** Para nós de indução de regra (a priori), selecione os campos a serem usados como antecedentes no conjunto de regras resultante. (Isso corresponde aos campos com papel *Entrada* ou *Ambos* em um nó Tipo).

Alguns modelos possuem uma guia Campos que difere daqueles descritos nesta seção.

- Consulte o tópico "Opções de Campos do Nó Sequência" na página 262 para obter mais informações.
- Consulte o tópico "Opções de Campos do Nó CARMA" na página 250 para obter mais informações.

Usando Campos de Frequência e de Ponderação

Os campos de frequência e de ponderação são utilizados para dar importância extra para alguns registros sobre outras, por exemplo, porque você sabe que uma parte da população está sub-representada nos dados de treinamento (ponderação) ou porque um registro representa um número de casos idênticos (frequência).

- Os valores para um campo de frequência devem ser números inteiros positivos. Os registros com uma ponderação de frequência negativa ou zero são excluídos da análise. As ponderações de frequência de números não inteiros são arredondadas para o número inteiro mais próximo.
- Os valores de ponderação de caso devem ser positivos, mas não precisam ser valores de número inteiro. Os registros com uma ponderação de caso negativa ou zero são excluídos da análise.

Escorando Campos de Frequência e de Ponderação

Os campos de frequência e de ponderação são utilizados nos modelos de treinamento, mas não são utilizados na escoragem porque o escore para cada registro baseia-se em suas características, independentemente de quantos casos ele representar. Por exemplo, suponha que você tenha os dados na tabela a seguir.

Tabela 1. Exemplo de dados

Casado	**Respondeu**
Sim	Sim
Sim	Sim
Sim	Sim
Sim	Não
Não	Sim
Não	Não
Não	Não

Com base nisso, você conclui que três das quatro pessoas casadas responderam à promoção, e duas das três pessoas não casadas não responderam. Portanto, você escorará quaisquer novos registros de acordo, conforme mostrado na tabela a seguir.

Tabela 2. Exemplo de registros escorados

Casado	**$-Respondeu**	**$RP-Respondido**
Sim	Sim	0,75 (três/quatro)
Não	Não	0,67 (dois/três)

Como alternativa, é possível armazenar seus dados de treinamento de modo mais compacto utilizando um campo de frequência, conforme mostrado na tabela a seguir.

Tabela 3. Exemplo alternativo de registros escorados

Casado	**Respondeu**	**Frequência**
Sim	Sim	3
Sim	Não	1
Não	Sim	1
Não	Não	2

Como isso representa exatamente o mesmo conjunto de dados, você construirá o mesmo modelo e preverá as respostas com base exclusivamente no estado civil. Se você tiver dez pessoas casadas em seus dados de escoragem, você preverá *Sim* para cada uma delas, independentemente se forem apresentadas como dez registros separados ou como um registro com um valor de frequência de 10. A ponderação, embora geralmente não seja um número inteiro, pode ser considerada como indicando da mesma forma a importância de um registro. É por isso que os campos de frequência e de ponderação não são utilizados quando escorar os registros.

Avaliando e Comparando Modelos

Alguns tipos de modelo suportam campos de frequência, outros suportam campos de ponderação e outros suportam ambos. Mas em todos os casos em que eles se aplicam, eles são utilizados apenas para construção de modelo e não são considerados ao avaliar os modelos utilizando um nó Avaliação ou um nó Análise, ou quando classificar modelos utilizando a maioria dos métodos suportados pelos nós Classificador Automático e Numeração Automática.

- Ao comparar modelos (usando gráficos de avaliação, por exemplo), os valores de frequência e de ponderação serão ignorados. Isso permite uma comparação de nível entre modelos que utilizam esses campos e modelos que não utilizam, mas significa que, para uma avaliação precisa, um conjunto de dados que represente com precisão a população sem depender de um campo de frequência ou de ponderação deve ser utilizado. Na prática, isso pode ser feito ao assegurar que os modelos sejam avaliados utilizando uma amostra de teste na qual o valor do campo de frequência ou de ponderação é sempre nulo ou 1. (Esta restrição se aplica apenas quando avaliar modelos; se os valores de frequência ou de ponderação forem sempre 1 para ambas as amostras de treinamento e de teste, não haverá motivo para utilizar esses campos em primeira instância).
- Se utilizar o Classificador Automático, a frequência poderá ser levada em conta se os modelos forem classificados com base no Lucro, de modo que esse método é recomendado nesse caso.
- Se necessário, é possível dividir os dados em amostras de treinamento e de teste utilizando um nó Partição.

Opções de Análise do Nó de Modelagem

Muitos nós de modelagem incluem uma guia Análise que permite obter informações de importância do preditor com escores de propensão bruta e ajustada.

Avaliação de modelo

Calcular a importância do preditor. Para modelos que produzem uma medida apropriada de importância, é possível exibir um gráfico que indica a importância relativa de cada preditor na estimativa do modelo. Geralmente, você deseja concentrar seus esforços de modelagem nos preditores que forem de maior importância e considerar eliminar ou ignorar aqueles que forem de menor importância. Observe que a importância do preditor poderá levar mais tempo para calcular para alguns modelos, principalmente quando estiver trabalhando com grandes conjuntos de dados, e é desativada por padrão para alguns modelos como resultado. A importância do preditor não está disponível para modelos de lista de decisão. Consulte o tópico “Importância do preditor” na página 44 para obter mais informações.

Escores de Propensão

Os escores de propensão podem ser ativados no nó de modelagem e na guia Configurações no nugget do modelo. Esta funcionalidade estará disponível apenas quando o destino selecionado for um campo de flag. Consulte o tópico “Escores de Propensão” na página 36 para obter mais informações.

Calcular escores de propensão bruta. Os escores de propensão bruta são derivados do modelo com base apenas nos dados de treinamento. Se o modelo prever o valor *true* (responderá), então a propensão será a mesma que P, em que P é a probabilidade da predição. Se o modelo prever o valor false, então e propensão é calculada como (1-P).

- Se você escolher essa opção ao construir o modelo, os escores de propensão serão ativados no nugget do modelo por padrão. No entanto, sempre é possível optar por ativar os escores de propensão bruta no nugget do modelo independentemente se você selecioná-los no nó de modelagem ou não.
- Ao escorar o modelo, os escores de propensão bruta serão incluídos em um campo com as letras *RP* anexadas ao prefixo padrão. Por exemplo, se as predições estiverem em um campo denominado *$R-churn*, o nome do campo de escore de propensão será *$RRP-churn*.

Calcular escores de propensão ajustada. As propensões brutas baseiam-se puramente nas estimativas fornecidas pelo modelo, que podem ser super ajustadas e gerar estimativas de propensão super otimistas. As propensões ajustadas tentam compensar isso ao examinar como o modelo é executado nas partições de teste ou de validação e ajustar as propensões para fornecer uma estimativa melhor de acordo.

- Essa configuração requer que um campo de partição válido esteja presente no fluxo.
- Diferentemente dos escores de confiança bruta, os escores de propensão ajustada devem ser calculados ao construir o modelo; caso contrário, eles não estarão disponíveis quando escorar o nugget do modelo.
- Ao escorar o modelo, os escores de propensão ajustada serão incluídos em um campo com as letras *AP* anexadas ao prefixo padrão. Por exemplo, se as predições estiverem em um campo denominado *$R-churn*, o nome do campo de escore de propensão será *$RAP-churn*. Os escores de propensão ajustada não estão disponíveis para modelos de regressão logística.
- Ao calcular os escores de propensão ajustada, a partição de teste ou de validação utilizada para o cálculo não deverá ter sido balanceada. Para evitar isso, assegure-se de que a opção **Balancear somente dados de treinamento** esteja selecionada em qualquer nó Balanceamento de envio de dados. Além disso, se uma amostra complexa tiver sido obtida anteriormente, isto invalidará os escores de propensão ajustada.
- Os escores de propensão ajustada não estão disponíveis para modelos de árvore e de conjunto de regras "impulsionados". Consulte o tópico “Modelos do C5.0 Impulsionados” na página 118 para obter mais informações.

Baseado em. Para que os escores de propensão ajustada sejam calculados, um campo de partição deverá estar presente no fluxo. É possível especificar se deseja utilizar a partição de teste ou de validação para este cálculo. Para obter melhores resultados, a partição de teste ou de validação deverá incluir pelo menos tantos registros quanto a partição usou para treinar o modelo original.

Escores de Propensão

Para modelos que retornam uma predição *sim* ou *não*, é possível solicitar escores de propensão além da predição padrão e de valores de confiança. Os escores de propensão indicam a probabilidade de um determinado resultado ou resposta. A tabela a seguir contém um exemplo.

Tabela 4. Escores de propensão

Cliente	Propensão para responder
Joe Smith	35%
Jane Smith	15%

Os escores de propensão estão disponíveis apenas para modelos com respostas flag e indicam a probabilidade do valor *True* definido para o campo, conforme especificado em uma origem ou nó Tipo.

Escores de Propensão Versus Escores de Confiança

Os escores de propensão diferem dos escores de confiança que se aplicam à predição atual, independentemente de *sim* ou *não*. Nos casos em que a predição é *não*, por exemplo, uma alta confiança na verdade significa uma alta probabilidade de *não* como resposta. Os escores de propensão contornam essa limitação para permitir a comparação em todos os registros. Por exemplo, uma predição *não* com uma confiança de *0,85* é convertida em uma propensão bruta de *0,15* (ou *1 menos 0,85*).

Tabela 5. Escores de confiança

Cliente	Predição	Confiança
Joe Smith	Irá responder	.35
Jane Smith	Não irá responder	.85

Obtendo Escores de Propensão

- Os escores de propensão podem ser ativados na guia Análise no nó de modelagem ou na guia Configurações no nugget do modelo. Esta funcionalidade estará disponível apenas quando o destino selecionado for um campo de flag. Consulte o tópico "Opções de Análise do Nó de Modelagem" na página 35 para obter mais informações.
- Os escores de propensão também podem ser calculados pelo nó Combinação, dependendo do método de combinação utilizado.

Calculando Escores de Propensão Ajustada

Os escores de propensão ajustada são calculados como parte do processo de construção do modelo, e não estarão disponíveis de outra forma. Depois que o modelo é construído, ele é escorado utilizando os dados da partição de teste ou de validação e um novo modelo para entregar os escores de propensão ajustada é construído ao analisar o desempenho do modelo original nessa partição. Dependendo do tipo de modelo, um dos dois métodos pode ser utilizado para calcular os escores de propensão ajustada.

- Para os modelos de conjunto de regras e de árvore, os escores de propensão ajustada são gerados ao recalcular a frequência de cada categoria em cada nó da árvore (para modelos de árvore) ou o suporte e a confiança de cada regra (para modelos de conjunto de regras). Isso resulta em um novo conjunto de regras ou modelo de árvore que é armazenado com o modelo original, a ser utilizado sempre que os

escores de propensão ajustada forem solicitados. Cada vez que o modelo original é aplicado aos novos dados, o novo modelo poderá subsequentemente ser aplicado ao escores de propensão bruta para gerar os escores ajustados.

- Para outros modelos, os registros produzidos ao escorar o modelo original na partição de teste ou de validação são então categorizados pelos seus escores de propensão bruta. Em seguida, um modelo de rede neural é treinado e define uma função não linear que é mapeada da propensão bruta média em cada categoria para a propensão média observada na mesma categoria. Conforme observado anteriormente para os modelos de árvore, o modelo de rede neural resultante é armazenado com o modelo original e pode ser aplicado aos escores de propensão bruta sempre que os escores de propensão ajustada forem solicitados.

Cuidado ao considerar valores omissos na partição de teste. O tratamento de valores de entrada omissos na partição de teste/validação varia por modelo (consulte algoritmos de escoragem de modelo individuais para obter detalhes). O modelo C5 não pode calcular propensões ajustadas quando houver entradas omissas.

Custos de classificação errada

Em alguns contextos, determinados tipos de erros são mais caros que outros. Por exemplo, pode ser mais caro classificar um solicitante de crédito de alto risco como baixo risco (um tipo de erro) do que classificar um solicitante de baixo risco como alto risco (um tipo diferente de erro). Os custos de classificação errada permitem especificar a importância relativa de diferentes tipos de erros de predição.

Os custos de classificação errada são basicamente ponderações aplicadas a resultados específicos. Essas ponderações são fatoradas no modelo e podem, na realidade, alterar a predição (como uma forma de proteger contra erros caros).

Com exceção dos modelos do C5.0, os custos de classificação errada não serão aplicados ao escorar um modelo e não são levados em conta quando classificar ou comparar modelos usando um nó Classificador Automático, gráfico de avaliação, ou nó Análise. Um modelo que inclui custos poderá não produzir menos erros do que aquele que não inclui e poderá não ter uma classificação mais alta em termos de precisão geral, mas provavelmente executará melhor em termos práticos por possuir um viés integrado a favor de erros *menos caros*.

A matriz de custo mostra o custo para cada combinação possível de categoria predita e categoria real. Por padrão, todos os custos de classificação errada são configurados como 1,0. Para inserir valores de custo customizado, selecione **Usar custos de classificação errada** e insira os valores customizados na matriz de custo.

Para alterar um custo de classificação errada, selecione a célula correspondente à combinação desejada de valores preditos e reais, exclua o conteúdo existente da célula e insira o custo desejado para a célula. Os custos não são simétricos automaticamente. Por exemplo, se você configurar o custo de classificação errada de *A* como *B* para 2,0, o custo da classificação errada de *B* como *A* ainda terá o valor padrão de 1,0, a menos que você também o altere explicitamente.

Nota: Apenas o modelo Árvores de Decisão permite que os custos sejam especificados no momento da construção.

Nuggets do Modelo

Figura 19. Nugget do modelo

Um nugget do modelo é um contêiner para um modelo, ou seja, o conjunto de regras, de fórmulas ou de equações que representam os resultados das operações de construção de seu modelo no IBM SPSS Modeler. O principal propósito de um nugget é escorar dados para gerar predições ou permitir análise adicional das propriedades do modelo. Abrir um nugget do modelo na tela permite ver vários detalhes sobre o modelo, como a importância relativa dos campos de entrada na criação do modelo. Para visualizar as predições, é necessário anexar e executar um processo adicional ou nó de saída. Consulte o tópico "Utilizando Nuggets do Modelo em Fluxos" na página 49 para obter mais informações.

Figura 20. Ligação de modelo do nó de modelagem para o nugget do modelo

Ao executar com sucesso um nó de modelagem, um nugget do modelo correspondente é colocado na tela de fluxo, na qual ele é representado por um ícone dourado em forma de losango (daí o nome "nugget"). Na tela de fluxo, o nugget é mostrado com uma conexão (linha sólida) com o nó adequado mais próximo antes do nó de modelagem, e um link (linha pontilhada) com o próprio nó de modelagem.

O nugget também é colocado na paleta Modelos no canto superior direito da janela do IBM SPSS Modeler. Em qualquer local, os nuggets podem ser selecionados e procurados para visualizar detalhes do modelo.

Os nuggets são sempre colocados na paleta Modelos quando um nó de modelagem é executado com sucesso. É possível configurar uma opção do usuário para controlar se o nugget também é colocado na tela de fluxo.

Os tópicos a seguir fornecem informações sobre como utilizar nuggets do modelo no IBM SPSS Modeler. Para obter um entendimento detalhado dos algoritmos utilizados, consulte o *Guia de Algoritmos do IBM SPSS Modeler* , disponível na pasta *\Documentation* no DVD para IBM SPSS Modeler.

Ligações de Modelo

Por padrão, um nugget é mostrado na tela com uma ligação para o nó de modelagem que o criou. Isso é útil principalmente em fluxos complexos com vários nuggets, permitindo identificar o nugget que será atualizado por cada nó de modelagem. Cada ligação contém um símbolo para indicar se o modelo é substituído quando o nó de modelagem é executado. Consulte o tópico "Substituindo um Modelo" na página 40 para obter mais informações.

Definindo e Removendo Ligações de Modelo

É possível definir e remover ligações manualmente na tela. Quando estiver definindo uma nova ligação, o cursor muda para o cursor de ligação.

Figura 21. Cursor de ligação

Definindo uma nova ligação (menu de contexto)

1. Clique com o botão direito no nó de modelagem no qual você deseja que a ligação inicie.
2. Escolha **Definir Ligação de Modelo** no menu de contexto.
3. Clique no nugget no qual deseja que a ligação termine.

Definindo uma nova ligação (menu principal)

1. Clique no nó de modelagem a partir do qual deseja que a ligação inicie.
2. No menu principal, escolha:
   Editar > **Nó** > **Definir Ligação de Modelo**
3. Clique no nugget no qual deseja que a ligação termine.

Removendo uma ligação existente (menu de contexto)

1. Clique com o botão direito no nugget no término da ligação.
2. Escolha **Remover Ligação de Modelo** no menu de contexto.

Como alternativa:

1. Clique com o botão direito no símbolo no meio da ligação.
2. Escolha **Remover Ligação** no menu de contexto.

Removendo uma ligação existente (menu principal)

1. Clique no nó de modelagem ou no nugget do qual deseja remover a ligação.
2. No menu principal, escolha:
   Editar > **Nó** > **Remover Ligação de Modelo**

Copiando e Colando Ligações de Modelo

Se você copiar um nugget vinculado, sem seu nó de modelagem, e colá-lo no mesmo fluxo, o nugget será colado com uma ligação para o nó de modelagem. Uma nova ligação possui o mesmo status de substituição de modelo (consulte "Substituindo um Modelo" na página 40) como a ligação original.

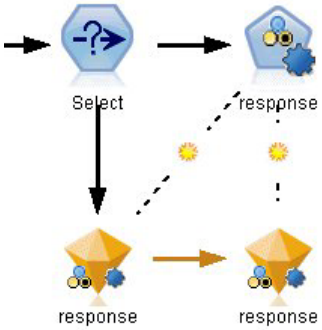


Figura 22. Copiando e colando um nugget vinculado

Se você copiar e colar um nugget com seu nó de modelagem vinculado, a ligação será mantida se os objetos forem colados no mesmo fluxo ou em um novo fluxo.

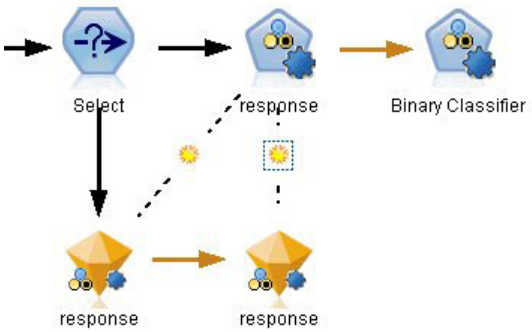


Figura 23. Copiando e colando um nugget vinculado

Nota: se você copiar um nugget vinculado, sem seu nó de modelagem, e colar o nugget em um novo fluxo (ou em um SuperNode que não contém o nó de modelagem), a ligação será quebrada e apenas o nugget será colado.

Ligações de Modelo e SuperNodes

Se você definir um SuperNode para incluir o nó de modelagem ou o nugget do modelo de um modelo vinculado (mas não ambos), a ligação será quebrada. Expandir o SuperNode não restaura a ligação; isso poderá ser feito apenas desfazendo a criação do SuperNode.

Substituindo um Modelo

É possível escolher se deseja substituir (ou seja, atualizar) um nugget existente na reexecução do nó de modelagem que criou o nugget. Se a opção de substituição for desativada, um novo nugget será criado quando reexecutar o nó de modelagem.

Nota: substituir um modelo é diferente de atualizar um modelo, pois isso refere-se a atualizar um modelo em um cenário.

Cada ligação do nó de modelagem com o nugget contém um símbolo para indicar se o modelo é substituído quando o nó de modelagem é reexecutado.

Figura 24. Ligação de modelo com substituição de modelo ativada

A ligação é inicialmente mostrada com a substituição de modelo ativada, representada por um pequeno símbolo de sol na ligação. Nesse estado, reexecutar o nó de modelagem em uma extremidade da ligação apenas atualizará o nugget na outra extremidade.

Figura 25. Ligação de modelo com substituição de modelo desativada

Se a substituição de modelo estiver desativada, o símbolo de ligação será substituído por um ponto cinza. Nesse estado, reexecutar o nó de modelagem em uma extremidade da ligação incluirá uma nova versão atualizada do nugget na tela.

Em qualquer caso, na paleta de Modelos, o nugget existente é atualizado ou um novo nugget é incluído, dependendo da configuração da opção de sistema **Substituir modelo anterior**.

Ordem de Execução

Ao executar um fluxo com diversas ramificações contendo nuggets do modelo, o fluxo é primeiro avaliado para assegurar que uma ramificação com a substituição de modelo ativada seja executada antes de qualquer ramificação que utiliza o nugget do modelo resultante.

Se seus requisitos forem mais complexos, será possível configurar a ordem de execução manualmente por meio de scripts.

Alterando a Configuração de Substituição de Modelo

Para alterar a configuração da substituição de modelo:

1. Clique com o botão direito no símbolo na ligação.
2. Escolha **Ativar/Desativar Substituição de Modelo**, conforme desejado.

Nota: a configuração de substituição de modelo em uma ligação de modelo substitui a configuração na guia Notificações do diálogo Opções do Usuário (Ferramentas > Opções > Opções do Usuário).

A Paleta de Modelos

A paleta de modelos (na guia Modelos na janela de gerenciadores) permite utilizar, examinar e modificar nuggets do modelo de várias maneiras.

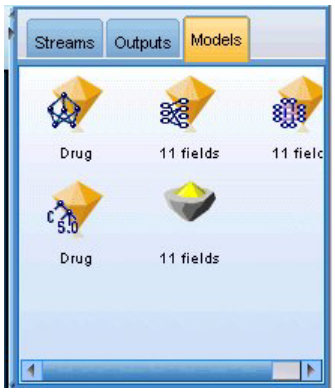


Figura 26. Paleta de Modelos

Clicar com o botão direito em um nugget do modelo na paleta de modelos abre um menu de contexto com as opções a seguir:

- **Incluir No Fluxo.** Inclui o nugget do modelo no fluxo atualmente ativo. Se houver um nó selecionado no fluxo, o nugget do modelo será conectado ao nó selecionado quando essa conexão for possível ou, caso contrário, ao nó mais próximo possível. O nugget é exibido com uma ligação para o nó de modelagem que criou o modelo, se esse nó ainda estiver no fluxo.
- **Procurar.** Abre o navegador de modelo para o nugget.
- **Renomear e Anotar.** Permite renomear o nugget do modelo e/ou modificar a anotação do nugget.

- **Gerar Nó de Modelagem.** Se você tiver um nugget do modelo que deseja modificar ou atualizar e o fluxo utilizado para criar o modelo não estiver disponível, será possível usar essa opção para recriar um nó de modelagem com as mesmas opções usadas para criar o modelo original.
- **Salvar Modelo, Salvar Modelo Como.** Salva o nugget do modelo em um arquivo binário de modelo gerado externo (.gm).
- **Armazenar Modelo.** Armazena o nugget do modelo no IBM SPSS Collaboration and Deployment Services Repository.
- **Exportar PMML.** Exporta o nugget do modelo como linguagem de marcações de modelo preditivo (PMML), que pode ser utilizado para escorar novos dados fora do IBM SPSS Modeler. A opção **Exportar PMML** está disponível para todos os nós de modelo gerados.
- **Incluir no Projeto.** Salva o nugget do modelo e o inclui no projeto atual. Na guia Classes, o nugget será incluído na pasta Modelos Gerados. Na guia CRISP-DM, ele será incluído na fase do projeto padrão.
- **Excluir.** Exclui o nugget do modelo da paleta.

Clicar com o botão direito em uma área não ocupada na paleta de modelos abre um menu de contexto com as opções a seguir:

- **Abrir Modelo.** Carrega um nugget do modelo criado anteriormente no IBM SPSS Modeler.
- **Recuperar Modelo.** Recupera um modelo armazenado a partir de um repositório do IBM SPSS Collaboration and Deployment Services.
- **Carregar Paleta.** Carrega uma paleta de modelos salva a partir de um arquivo externo.
- **Recuperar Paleta.** Recupera uma paleta de modelos armazenada a partir de um repositório do IBM SPSS Collaboration and Deployment Services.
- **Salvar Paleta.** Salva o conteúdo inteiro da paleta de modelos em um arquivo de paleta de modelos gerada (.gen) externo.
- **Armazenar Paleta.** Armazena o conteúdo inteiro da paleta de modelos em um repositório do IBM SPSS Collaboration and Deployment Services.
- **Limpar Paleta.** Exclui todos os nuggets da paleta.
- **Incluir Paleta No Projeto.** Salva a paleta de modelos e a inclui no projeto atual. Na guia Classes, o nugget será incluído na pasta Modelos Gerados. Na guia CRISP-DM, ele será incluído na fase do projeto padrão.
- **Importar PMML.** Carrega um modelo a partir de um arquivo externo. É possível abrir, navegar e escorar modelos PMML criados pelo IBM SPSS Statistics ou outros aplicativos que suportam este formato. Consulte o tópico “Importando e exportando modelos como PMML” na página 50 para obter mais informações.

Procurando Nuggets do Modelo

Os navegadores de nugget do modelo permitem examinar e utilizar os resultados de seus modelos. No navegador, é possível salvar, imprimir ou exportar o modelo gerado, examinar a sumarização do modelo e visualizar ou editar anotações para o modelo. Para alguns tipos de nugget do modelo, também é possível gerar novos nós, como nós Filtro ou nós Conjunto de Regras. Para alguns modelos, também é possível visualizar os parâmetros do modelo, como regras ou centros do cluster. Para alguns tipos de modelos (modelos baseados em árvore e modelos de cluster), é possível visualizar uma representação gráfica da estrutura do modelo. Os controles para utilizar os navegadores de nugget do modelo são descritos a seguir.

Menus

menu Arquivo. Todos os nuggets do modelo possuem um menu Arquivo contendo um subconjunto das opções a seguir:

- **Salvar Nó.** Salva o nugget do modelo em um arquivo de nó (.nod).

- **Armazenar Nó.** Armazena o nugget do modelo em um repositório do IBM SPSS Collaboration and Deployment Services.
- **Cabeçalho e Rodapé.** Permite editar o cabeçalho e o rodapé da página para impressão a partir do nugget.
- **Configuração da Página.** Permite alterar a configuração da página para impressão a partir do nugget.
- **Visualização de Impressão.** Exibe uma visualização de como será a aparência do nugget quando impresso. Selecione as informações que você deseja visualizar a partir do submenu.
- **Imprimir.** Imprime o conteúdo do nugget. Selecione as informações que você deseja imprimir a partir do submenu.
- **Imprimir Visualização.** Imprime a visualização atual ou todas as visualizações.
- **Exportar Texto.** Exporta o conteúdo do nugget em um arquivo de texto. Selecione as informações que você deseja exportar a partir do submenu.
- **Exportar HTML.** Exporta o conteúdo do nugget em um arquivo HTML. Selecione as informações que você deseja exportar a partir do submenu.
- **Exportar PMML.** Exporta o modelo como linguagem de marcações de modelo preditivo (PMML), que poderá ser utilizado com outros softwares compatíveis com PMML. Consulte o tópico "Importando e exportando modelos como PMML" na página 50 para obter mais informações.
- **Exportar SQL.** Exporta o modelo como Linguagem de Consulta Estruturada (SQL), que poderá ser editado e usado com outros bancos de dados.

  Nota: a Exportação SQL está disponível apenas a partir dos seguintes modelos: C5, C&RT, CHAID, QUEST, Regressão Linear, Regressão Logística, Rede Neural, PCA/Fator e Lista de Decisão.
- **Publicar no Server Scoring Adapter.** Publica o modelo em um banco de dados que possui um adaptador de escoragem instalado, permitindo que a escoragem de modelo seja executada no banco de dados. Consulte o tópico "Modelos de Publicação para um Adaptador de Escoragem" na página 52 para obter mais informações.

Menu Gerar. A maioria dos nuggets do modelo também possui um menu Gerar, permitindo gerar novos nós com base no nugget do modelo. As opções disponíveis neste menu dependerão do tipo de modelo que você está procurando. Consulte o tipo de nugget do modelo específico para obter detalhes sobre o que é possível gerar a partir de um modelo específico.

Menu Visualização. Na guia Modelo de um nugget, esse menu permite exibir ou ocultar as várias barras de ferramentas de visualização que estiverem disponíveis no modo atual. Para tornar o conjunto completo de ferramentas disponível, selecione Modo de Edição (o ícone de pincel) na barra de ferramentas Geral.

Botão Visualizar. Alguns nuggets do modelo possuem um botão Visualizar que permite ver uma amostra dos dados do modelo, incluindo campos extras criados pelo processo de modelagem. O número padrão de linhas exibidas é 10, entretanto, é possível alterar isso nas propriedades do fluxo.

Botão Incluir no Projeto Atual. Salva o nugget do modelo e o inclui no projeto atual. Na guia Classes, o nugget será incluído na pasta Modelos Gerados. Na guia CRISP-DM, ele será incluído na fase do projeto padrão.

Sumarização / Informações do Nugget do Modelo

A guia Sumarização ou a visualização Informações de um nugget do modelo exibe informações sobre os campos, sobre as configurações de construção e sobre o processo de estimação do modelo. Os resultados são exibidos em uma visualização em árvore que pode ser expandida ou reduzida ao clicar nos itens específicos.

Análise. Exibe informações sobre o modelo. Os detalhes específicos variam por tipo de modelo e são abordados na seção de cada nugget do modelo. Além disso, se você tiver executado um nó Análise anexado a este nó de modelagem, as informações dessa análise também serão exibidas nesta seção.

Campos. Lista os campos utilizados como o destino e as entradas na construção do modelo. Para modelos de divisão, lista também os campos que determinarem as divisões.

Configurações / Opções de Criação. Contém informações sobre as configurações utilizadas na construção do modelo.

Sumarização do Treinamento. Mostra o tipo de modelo, o fluxo utilizado para criá-lo, o usuário que o criou, quando ele foi construído e o tempo decorrido para construir o modelo. Observe que o tempo decorrido para construir o modelo está disponível apenas na guia Sumarização, não na visualização Informações.

Importância do preditor

Geralmente você desejará focar seus esforços de modelagem nos campos preditores que forem mais importantes e considerar descartar ou ignorar aqueles que forem menos importantes. O gráfico de importância do preditor ajuda a fazer isso ao indicar a importância relativa de cada preditor na estimativa do modelo. Como os valores são relativos, a soma dos valores para todos os preditores na tela é 1,0. A importância do preditor não tem relação com a precisão do modelo. Ela está relacionada apenas com a importância de cada preditor em fazer uma predição, e não se a predição é precisa ou não.

A importância do preditor está disponível para modelos que produzem uma medida estatística adequada de importância, incluindo modelos de redes neurais, árvores de decisão (C& Tree R, C5.0, CHAID e QUEST), redes bayesianas, discriminantes, modelos SVM e SLRM, de regressão logística e linear, lineares generalizados e de vizinho mais próximo (KNN). Para a maioria desses modelos, a importância do preditor pode ser ativada na guia Análise no nó de modelagem. Consulte o tópico “Opções de Análise do Nó de Modelagem” na página 35 para obter mais informações. Para modelos KNN, consulte “Vizinhos” na página 333.

Nota: a importância do preditor não é suportada para modelos de divisão. As configurações da importância do preditor são ignoradas durante a construção de modelos de divisão. Consulte o tópico “Construindo Modelos de Divisão” na página 28 para obter mais informações.

Calcular a importância do preditor pode demorar muito mais tempo do que construir o modelo, principalmente quando utilizar conjuntos de dados grandes. O cálculo demora mais para SVM e regressão de logística do que para outros modelos e está desativado para esses modelos por padrão. Se estiver usando um conjunto de dados com um número grande de preditores, a triagem inicial usando um nó de Seleção de Variável poderá fornecer resultados mais rápidos (veja abaixo).

- A importância do preditor é calculada a partir da partição de teste, se disponível. Caso contrário, os dados de treinamento serão utilizados.
- Para modelos SLRM, a importância do preditor está disponível, mas é calculada pelo algoritmo SLRM. Consulte o tópico “Nuggets do Modelo SLRM” na página 322 para obter mais informações.
- É possível utilizar as ferramentas do gráfico no IBM SPSS Modeler para interagir, editar e salvar o gráfico.
- Opcionalmente, é possível gerar um nó Filtro com base nas informações no gráfico de importância do preditor. Consulte o tópico “Filtrando Variáveis com Base em Importância” na página 45 para obter mais informações.

Importância do Preditor e Seleção de Variável

O gráfico de importância do preditor exibido em um nugget do modelo pode parecer fornecer resultados semelhantes ao nó Seleção em alguns casos. Embora a seleção de variável classifique cada campo de

entrada com base na intensidade de seu relacionamento com a resposta especificada, independentemente de outras entradas, o gráfico de importância do preditor indica a importância relativa de cada entrada com *este* modelo específico. Assim, a seleção de variável será mais conservadora na triagem de entradas. Por exemplo, se o *título do emprego* e a *categoria do emprego* estiverem fortemente relacionados ao salário, a seleção de variável indicaria que ambos são importantes. Mas em modelagem, as interações e correlações também são levadas em consideração. Portanto, você pode achar que apenas uma das duas entradas será utilizada se ambas duplicarem grande parte as mesmas informações. Na prática, a seleção de variável é mais útil para triagem preliminar, particularmente quando lidar com grandes conjuntos de dados com grandes números de variáveis, e a importância do preditor é mais útil para um ajuste preciso do modelo.

Diferenças de Importância do Preditor Entre Modelos Únicos e Nós de Modelagem Automatizados

Dependendo se você estiver criando um único modelo a partir de um nó individual ou utilizando um nó de modelagem automatizado para produzir resultados, poderá haver pequenas diferenças na importância do preditor. Essas diferenças na implementação ocorrem devido a algumas restrições de engenharia.

Por exemplo, com classificadores únicos, como CHAID, o cálculo aplica uma regra de parada e utiliza valores de probabilidade quando calcular valores de importância. Em contraste, o Classificador Automático não utiliza uma regra de parada e utiliza rótulos preditos diretamente no cálculo. Essas diferenças podem significar que se você produzir um modelo único utilizando Classificador Automático, o valor da importância poderá ser considerado como uma estimativa aproximada, em comparação com o valor calculado para um classificador único. Para obter os valores de importância do preditor mais precisos, recomenda-se utilizar um nó único ao invés dos nós de modelagem automatizados.

Filtrando Variáveis com Base em Importância

Opcionalmente, é possível gerar um nó Filtro com base nas informações no gráfico de importância do preditor.

Marque os preditores que desejar incluir no gráfico, se aplicável, e nos menus escolha:

Gerar > **Nó Filtro (Importância do Preditor)**

OR

> **Seleção de Campo (Importância do Preditor)**

Número máximo de variáveis. Inclui ou exclui os preditores mais importantes até o número especificado.

Importância maior que. Inclui ou exclui todos os preditores com importância relativa maior que o valor especificado.

Visualizador de Combinação

Modelos para Combinações

O modelo para uma combinação fornece informações sobre os modelos de componente na combinação e sobre o desempenho da combinação como um todo.

A barra de ferramentas principal (visualização independente) permite escolher se deseja utilizar a combinação ou um modelo de referência para escoragem. Se a combinação for utilizada para escoragem, também é possível selecionar a regra de combinação. Essas mudanças não requerem uma reexecução do modelo; no entanto, estas opções são salvas no modelo (nugget) para escoragem e/ou avaliação de modelo de recebimento de dados. Elas também afetam o PMML exportado do visualizador de combinação.

Combinando Regra. Ao escorar uma combinação, essa é a regra utilizada para combinar os valores preditos a partir dos modelos base para calcular o valor de escore de combinação.

- Os valores preditos de combinação para **variável resposta categórica** podem ser combinados utilizando votação, probabilidade mais alta ou probabilidade média mais alta. **Votação** seleciona a categoria que tem a probabilidade mais alta e mais frequente nos modelos base. **Probabilidade mais alta** seleciona a categoria que atinge a única probabilidade mais alta em todos os modelos base. **Probabilidade média mais alta** Seleciona a categoria com o valor mais alto quando a média das probabilidades da categoria é calculada entre os modelos base.
- Os valores preditos de combinação para **variáveis resposta contínua** podem ser combinados utilizando a média ou mediana dos valores preditos a partir dos modelos base.

O padrão é obtido a partir das especificações feitas durante a construção de modelo. Alterar a regra de combinação recalcula a precisão do modelo e atualiza todas as visualizações de precisão do modelo. O gráfico de Importância do Preditor também é atualizado. Este controle será desativado se o modelo de referência for selecionado para escoragem.

Mostrar Todas as Regras de Combinação. Quando selecionada, os resultados de todas as regras de combinação disponíveis são mostrados no gráfico de qualidade do modelo. O gráfico de Precisão de Modelo de Componente também é atualizado para mostrar linhas de referência para cada método de votação.

Sumarização do Modelo: A visualização Sumarização do Modelo é uma captura instantânea ou uma sumarização rápida da qualidade e da diversidade da combinação.

Qualidade. O gráfico exibe a precisão do modelo final, em comparação com um modelo de referência e com um modelo Naive. A precisão é apresentada em um formato maior e melhor e o "melhor" modelo terá a precisão mais alta. Para uma variável resposta categórica, a precisão é simplesmente a porcentagem de registros para a qual o valor predito corresponde ao valor observado. Para uma variável resposta contínua, a precisão é 1 menos a razão da média de erro absoluto na predição (a média dos valores absolutos dos valores preditos menos os valores observados) com o intervalo de valores preditos (o valor máximo predito menos o valor mínimo predito).

Para combinações de bagging, o modelo de referência é um modelo padrão construído na partição de treinamento inteira. Para combinações impulsionadas, o modelo de referência é o primeiro modelo de componente.

O modelo Naive representará a precisão se nenhum modelo tiver sido construído e designa todos os registros para a categoria modal. O modelo Naive não é calculado para variáveis resposta contínuas.

Diversidade. O gráfico exibe a "diversidade de opiniões" entre os modelos de componente utilizados para construir a combinação, apresentada em um formato maior e mais diversificado. É uma medida do quanto as predições variam entre os modelos base. A diversidade não está disponível para modelos de combinação impulsionados, nem é mostrada para variáveis resposta contínuas.

Importância do Preditor: Geralmente você desejará focar seus esforços de modelagem nos campos preditores que forem mais importantes e considerar descartar ou ignorar aqueles que forem menos importantes. O gráfico de importância do preditor ajuda a fazer isso ao indicar a importância relativa de cada preditor na estimativa do modelo. Como os valores são relativos, a soma dos valores para todos os preditores na tela é 1,0. A importância do preditor não tem relação com a precisão do modelo. Ela está relacionada apenas com a importância de cada preditor em fazer uma predição, e não se a predição é precisa ou não.

A importância do preditor não está disponível para todos os modelos de combinação. O conjunto de preditores pode variar entre os modelos de componente, mas a importância poderá ser calculada para preditores usados em pelo menos um modelo de componente.

Frequência do Preditor: O conjunto de preditores pode variar entre os modelos de componente devido à escolha do método de modelagem ou da seleção do preditor. O gráfico Frequência do Preditor é um gráfico de pontos que mostra a distribuição de preditores nos modelos de componente na combinação. Cada ponto representa um ou mais modelos de componente contendo o preditor. Os preditores são representados no eixo y e são classificados em ordem decrescente de frequência, assim, o preditor mais acima é aquele que é utilizado no maior número de modelos de componente e o preditor mais baixo foi o menos utilizado. Os 10 primeiros preditores são mostrados.

Os preditores que aparecem com mais frequência geralmente são os mais importantes. Esse gráfico não é útil para métodos nos quais o conjunto de preditores não pode variar entre os modelos de componente.

Precisão de Modelo de Componente: O gráfico é um gráfico de pontos de precisão preditiva para modelos de componente. Cada ponto representa um ou mais modelos de componente com o nível de precisão representado no eixo y. Passe o mouse sobre qualquer ponto para obter informações sobre o modelo de componente individual correspondente.

Linhas de referência. O gráfico exibe linhas codificadas por cor para a combinação e também para o modelo de referência e modelos Naive. Um visto aparece ao lado da linha correspondente ao modelo que será utilizado para escoragem.

Interatividade. O gráfico será atualizado se você alterar a combinação da regra.

Combinações impulsionadas. Um gráfico de linha é exibido para combinações impulsionadas.

Detalhes de Modelo de Componente: A tabela exibe informações sobre modelos de componente, listados por linha. Por padrão, os modelos de componente são classificados em ordem crescente de número do modelo. É possível classificar as linhas em ordem crescente ou decrescente pelos valores de qualquer coluna.

Modelo. Um número que representa a ordem sequencial na qual o modelo de componente foi criado.

Precisão. Precisão geral formatada como uma porcentagem.

Método. O método de modelagem.

Preditores. O número de preditores utilizados no modelo de componente.

Tamanho do Modelo. O tamanho do modelo depende do método de modelagem: para árvores, é o número de nós na árvore; para modelos lineares, é o número de coeficientes; para redes neurais, é o número de sinapses.

Registros. O número ponderado de registros de entrada na amostra de treinamento.

Preparação Automática de Dados:

Essa visualização mostra informações sobre quais campos foram excluídos e como os campos transformados foram derivados no passo de preparação automática de dados (ADP). Para cada campo que foi transformado ou excluído, a tabela lista o nome do campo, o seu papel na análise e a ação executada pelo passo ADP. Os campos são classificados em ordem alfabética crescente de nomes de campo.

A ação **Aparar valores discrepantes**, se mostrada, indica que os valores de preditores contínuos que estiverem além de um valor de corte (3 desvios padrão da média) foram configurados para o valor de corte.

Nuggets do Modelo para Modelos de Divisão

O nugget do modelo para um modelo de divisão fornece acesso a todos os modelos separados criados pelas divisões.

Um nugget do modelo de divisão contém:

- uma lista de todos os modelos de divisão criados, com um conjunto de estatísticas sobre cada modelo
- informações sobre o modelo geral

Na lista de modelos de divisão, é possível abrir modelos individuais para examiná-los adicionalmente.

Visualizador de Modelo de Divisão

A guia Modelo lista todos os modelos contidos no nugget e fornece estatísticas em vários formatos sobre os modelos de divisão. Ela possui duas formas gerais, dependendo do nó de modelagem.

Ordenar por. Utilize essa lista para escolher a ordem na qual os modelos são listados. É possível ordenar a lista com base nos valores de qualquer uma das colunas de exibição, em ordem crescente ou decrescente. Como alternativa, clique em um título da coluna para ordenar a lista por essa coluna. O padrão é ordem decrescente de precisão geral.

Mostrar/ocultar menu de colunas. Clique neste botão para exibir um menu a partir do qual é possível escolher colunas individuais para mostrar ou ocultar.

Visualizar. Se estiver utilizando particionamento, será possível optar por visualizar os resultados para os dados de treinamento ou dados de teste.

Para cada divisão, os detalhes mostrados são os seguintes:

Gráfico. Uma miniatura que indica a distribuição de dados para este modelo. Quando o nugget estiver na tela, clique duas vezes na miniatura para abrir o gráfico em tamanho integral.

Modelo. Um ícone do tipo de modelo. Clique duas vezes no ícone para abrir o nugget do modelo para esta divisão específica.

Campos de divisão. Os campos designados no nó de modelagem como campos de divisão, com seus vários valores possíveis.

Nº de Registros na Divisão. O número de registros envolvidos nessa divisão específica.

Nº de Campos Utilizados. Classifica os modelos de divisão com base no número de campos de entrada utilizados.

Precisão Geral (%). A porcentagem de registros que é predita corretamente pelo modelo de divisão com relação ao número total de registros nessa divisão.

Divisão. O título da coluna mostra um ou mais campos utilizados para criar divisões, e as células são os valores da divisão. Clique duas vezes em qualquer divisão para abrir o Visualizador de Modelo para o modelo construído para essa divisão.

Precisão. Precisão geral formatada como uma porcentagem.

Tamanho do Modelo. O tamanho do modelo depende do método de modelagem: para árvores, é o número de nós na árvore; para modelos lineares, é o número de coeficientes; para redes neurais, é o número de sinapses.

Registros. O número ponderado de registros de entrada na amostra de treinamento.

Utilizando Nuggets do Modelo em Fluxos

Os nuggets do modelo são colocados nos fluxos para poder escorar novos dados e gerar novos nós. A **escoragem de dados** permite utilizar as informações obtidas a partir da construção de modelo para criar predições para novos registros. Para ver os resultados da escoragem, é necessário anexar um nó terminal (ou seja, um nó de processamento ou de saída) ao nugget e executar o nó terminal.

Para alguns modelos, os nuggets do modelo também podem fornecer informações adicionais sobre a qualidade da predição, como valores de confiança ou distâncias de centros do cluster. Gerar novos nós permite criar facilmente novos nós com base na estrutura do modelo gerado. Por exemplo, a maioria dos modelos que executam a seleção de campo de entrada permite gerar nós Filtros que transmitirão apenas os campos de entrada que o modelo identificou como importantes.

Nota: Poderá haver pequenas diferenças nos escores designados a um determinado caso por um modelo específico quando executados em diferentes versões do IBM SPSS Modeler. Este é geralmente o resultado de aprimoramentos no software entre versões.

Para Utilizar um Nugget do Modelo para Escoragem dos Dados

1. Conecte o nugget do modelo a uma origem de dados ou fluxo que transmitirá dados para ele.
2. Inclua ou conecte um ou mais nós de processamento ou de saída (como um nó de Tabela ou Análise) ao nugget do modelo.
3. Execute um dos nós de recebimento de dados do nugget do modelo.

Nota: não é possível utilizar o nó Regra Não Refinada para escoragem de dados. Para escorar dados com base em um modelo de regra de associação, use o nó Regra Não Refinada para gerar um nugget Conjunto de Regra e use o nugget Conjunto de Regras para escoragem. Consulte o tópico "Gerando um Conjunto de Regras a partir de um Nugget do Modelo de Associação" na página 257 para obter mais informações.

Para Utilizar um Nugget do Modelo para Gerar Nós de Processamento

1. Na paleta, procure o modelo ou, na tela de fluxo, edite o modelo.
2. Selecione o tipo de nó desejado no menu Gerar da janela do navegador de nugget do modelo. As opções disponíveis variam, dependendo do tipo de nugget do modelo. Consulte o tipo de nugget do modelo específico para obter detalhes sobre o que é possível gerar a partir de um modelo específico.

Gerando novamente um Nó de Modelagem

Se você tiver um nugget do modelo que deseja modificar ou atualizar e o fluxo utilizado para criar o modelo não estiver disponível, será possível gerar novamente um nó de modelagem com as mesmas opções usadas para criar o modelo original.

Para reconstruir um modelo, clique com o botão direito no modelo na paleta de modelos e escolha **Gerar Nó de Modelagem**.

Como alternativa, quando procurar qualquer modelo, escolha **Gerar Nó de Modelagem** a partir do menu Gerar.

O nó de modelagem gerado novamente deve ser funcionalmente idêntico ao que foi utilizado para criar o modelo original na maioria dos casos.

- Para modelos de Árvore de Decisão, configurações adicionais especificadas durante a sessão interativa também podem ser armazenadas com o nó, e a opção **Usar diretivas de árvore** será ativada no nó de modelagem gerado novamente.

- Para modelos de Lista de Decisão, a opção **Usar informações da sessão interativa salva** será ativada. Consulte o tópico "Opções do Modelo da Lista de Decisão" na página 146 para obter mais informações.
- Para modelos de Séries Temporais, a opção **Continuar a estimação usando modelo(s) existente(s)** está ativada, que permite gerar novamente o modelo anterior com os dados atuais. Consulte o tópico "Opções do Modelo de Série Temporal" na página 285 para obter mais informações.

Importando e exportando modelos como PMML

PMML, ou Predictive Model Markup Language, é um formato XML para descrever modelos estatísticos e mineração de dados, incluindo entradas para os modelos, transformações usadas para preparar dados para mineração de dados e os parâmetros que definem os modelos em si. IBM SPSS Modeler pode importar e exportar PMML, possibilitando o compartilhamento de modelos com outros aplicativos que suportem esse formato, como IBM SPSS Statistics.

Para obter mais informações sobre PMML, consulte o website Grupo de Mineração de Dados (*http://www.dmg.org*).

Para Exportar um Modelo

A exportação de PMML é suportada para a maioria dos tipos de modelo gerados no IBM SPSS Modeler. Consulte o tópico "Tipos de modelo que suportam PMML" na página 51 para obter mais informações.

1. Clique com o botão direito em um nugget do modelo na paleta de modelos. (Alternativamente, clique duas vezes em um nugget do modelo na tela e selecione o menu Arquivo.)
2. No menu, clique em **Exportar PMML**.
3. Na caixa de diálogo Exportar (ou Salvar), especifique um diretório de resposta e um nome exclusivo para o modelo.

Nota: é possível alterar opções para exportação de PMML na caixa de diálogo Opções de Usuário. No menu principal, clique em:

Ferramentas > **Opções** > **Opções de Usuário**

e clique na guia PMML.

Para Importar um Modelo Salvo como PMML

Modelos exportados como PMML do IBM SPSS Modeler ou ouro aplicativo podem ser importados na paleta de modelos. Consulte o tópico "Tipos de modelo que suportam PMML" na página 51 para obter mais informações.

1. Na paleta de modelos, clique com o botão direito do mouse na paleta e selecione **Importar PMML** no menu.
2. Selecione o arquivo para importar e especifique opções para rótulos de variáveis conforme necessário.
3. Clique em **Abrir**.

Usar rótulos de variáveis se presentes no modelo. O PMML pode especificar nomes de variáveis e rótulos de variáveis (como ID do Referenciador para *RefID*) para variáveis no dicionário de dados. Selecione essa opção para usar rótulos de variáveis se eles estiverem presentes no PMML exportado originalmente.

Se você tiver selecionado a opção de rótulo de variáveis, mas não houver rótulos de variáveis no PMML, os nomes de variáveis serão usados como normalmente.

Tipos de modelo que suportam PMML

Exportação de PMML

Modelos do IBM SPSS Modeler. Os modelos a seguir criados no IBM SPSS Modeler podem ser exportados como PMML 4.0:

- Árvore C e R
- QUEST
- CHAID
- Regressão linear
- Rede neural
- C5.0
- Regressão Logística
- Genlin
- SVM
- a priori
- Carma
- K-Médias
- Kohonen
- TwoStep
- GLMM (suporte somente para modelos GLMM Somente de Efeito Fixo)
- Lista de Decisão
- Cox
- Sequência (escoragem para modelos PMML de Sequência não é suportada)
- Modelo Statistics

Modelos nativos de banco de dados. Para modelos gerados usando algoritmos nativos do banco de dados, a exportação de PMML está disponível somente para modelos IBM InfoSphere Warehouse. Modelos criados usando Serviços de Análise da Microsoft ou Oracle Data Miner não podem ser exportados. Além disso, observe que os modelos da IBM exportados como PMML não podem ser importados de volta no IBM SPSS Modeler.

Importação de PMML

IBM SPSS Modeler pode importar e escorar modelos PMML gerados pelas versões atuais de todos os produtos IBM SPSS Statistics, incluindo modelos exportados do IBM SPSS Modeler, bem como PMML de transformação ou modelo gerado pelo IBM SPSS Statistics 17.0 ou posterior. Essencialmente, isso significa qualquer PMML que o mecanismo de escoragem pode escorar com as exceções a seguir:

- A Priori, CARMA, Detecção de Anomalias, Sequência e Modelos de Regra de Associação não podem ser importados.
- Os modelos PMML podem não ser procurados após importados no IBM SPSS Modeler, mesmo se puderem ser usados na escoragem. (Observe que isso inclui modelos que foram exportados do IBM SPSS Modeler com os quais começar. Para evitar essa limitação, exporte o modelo como um arquivo de modelo gerado [**.gm*] ao invés de PMML).
- Modelos do IBM InfoSphere Warehouse exportados como PMML não podem ser importados.
- A validação limitada ocorre na importação, mas a validação integral é executada na tentativa de escorar o modelo. Assim, é possível que a importação seja bem-sucedida, mas a escoragem falhará ou produzirá resultados incorretos.

Nota: Para PMML de terceiros importados no IBM SPSS Modeler, IBM SPSS Modeler tentará escorar PMML válido que pode ser reconhecido e escorado. Mas não é garantido que todo PMML escorará ou que escorará da mesma maneira que o aplicativo que o gerou.

Modelos de Publicação para um Adaptador de Escoragem

É possível publicar modelos em um servidor de banco de dados que tenha um adaptador de escoragem instalado. Um adaptador de escoragem permite que a escoragem de modelo seja executada no banco de dados utilizando os recursos da função definida pelo usuário (UDF) do banco de dados. Executar escoragem no banco de dados evita a necessidade de extrair os dados antes da escoragem. Publicar em um adaptador de escoragem também gera algum exemplo de SQL para executar a UDF.

Para publicar em um adaptador de escoragem:

1. Clique duas vezes no nugget do modelo para abri-lo.
2. No menu de nugget do modelo, escolha:
   Arquivo > **Publicar no Adaptador de Escoragem do Servidor**
3. Preencha os campos relevantes na caixa de diálogo e clique em **OK**.

Conexão com o banco de dados. Os detalhes de conexão com o banco de dados que você deseja utilizar para o modelo.

ID de Publicação. (somente banco de dados DB2 for z/OS) Um identificador para o modelo. Se você reconstruir o mesmo modelo e utilizar o mesmo ID de publicação, o SQL gerado permanecerá o mesmo, o que permite reconstruir um modelo sem precisar alterar o aplicativo que utiliza o SQL gerado anteriormente. (Para outros bancos de dados, o SQL gerado é exclusivo para o modelo).

Gerar SQL de Exemplo. Se selecionado, gera o SQL de exemplo no arquivo especificado no campo **Arquivo**.

Modelos Não Refinados

Um modelo não refinado contém informações extraídas dos dados, mas não é projetado para gerar predições diretamente. Isso significa que ele não pode ser incluído nos fluxos. Modelos não refinados são exibidos como “losangos em bruto” na paleta de modelos gerados.

Figura 27. Ícone de modelo não refinado

Para ver informações sobre o modelo de regra não refinado, clique com o botão direito no modelo e escolha **Procurar** no menu de contexto. Assim como outros modelos gerados no IBM SPSS Modeler, as várias guias fornecem informações de sumarização e de regras sobre o modelo criado.

Gerando nós. O menu Gerar permite criar novos nós com base nas regras.

- **Nó Seleção.** Gera um nó de Seleção para selecionar registros aos quais a regra selecionada atualmente se aplica. Esta opção estará desativada se nenhuma regra estiver selecionada.
- **Conjunto de regras.** Gera um nó Conjunto de Regras para prever valores para um campo de destino único. Consulte o tópico “Gerando um Conjunto de Regras a partir de um Nugget do Modelo de Associação” na página 257 para obter mais informações.

Capítulo 4. Modelos de Triagem

Rastreando Campos e Registros

Vários nós de modelagem podem ser utilizados durante os estágios preliminares de uma análise para localizar campos e registros que forem de maior interesse para modelagem. É possível utilizar o nó Seleção de Variável para verificar e classificar campos por importância e usar o nó Detecção de Anomalias para localizar registros incomuns que não estiverem em conformidade com os padrões conhecidos de dados "normais".

O nó Seleção de Variável verifica campos de entrada para remoção com base em um conjunto de critérios (como a porcentagem de valores omissos) e, em seguida, classifica a importância das entradas restantes com relação a um destino especificado. Por exemplo, dado um conjunto de dados com centenas de possíveis entradas, quais delas mais poderão ser úteis para modelar os resultados do paciente?

O nó Detecção de Anomalia identifica casos incomuns, ou valores discrepantes, que não estiverem em conformidade com os padrões de dados “normais”. Com este nó, é possível identificar valores discrepantes, mesmo se eles não se ajustarem a nenhum dos padrões anteriormente conhecidos e mesmo se você não souber exatamente o que procura.

Observe que a detecção de anomalias identifica registros ou casos incomuns por meio de análise de cluster com base no conjunto de campos selecionados no modelo sem levar em consideração nenhum campo de destino específico (dependente) e independentemente se esses campos forem relevantes ao padrão que você está tentando prever. Por essa razão, você pode querer utilizar a detecção de anomalias em combinação com a seleção de variável ou outra técnica para verificar e classificar campos. Por exemplo, é possível utilizar a seleção de variável para identificar os campos mais importantes com relação a um destino específico e, em seguida, usar a detecção de anomalias para localizar os registros que forem mais incomuns com relação a esses campos. (Uma abordagem alternativa seria a construção de um modelo de árvore de decisão e, em seguida, examinar quaisquer registros que forem classificados incorretamente como possíveis anomalias. No entanto, este método seria mais difícil para replicar ou automatizar em grande escala).

Nó Seleção de Variável

Os problemas de mineração de dados podem envolver centenas, ou até mesmo milhares, de campos que podem potencialmente ser usados como entradas. Como resultado, muito tempo e esforço poderão ser gastos examinando quais campos ou variáveis deverão ser incluídos no modelo. Para limitar as opções, o algoritmo Seleção de Variável pode ser utilizado para identificar os campos que forem mais importantes para uma determinada análise. Por exemplo, se estiver tentando prever resultados de um paciente com base em diversos fatores, quais desses fatores poderão ser os mais importantes?

A seleção de variável consiste em três passos:

- **Triagem.** Remove entradas e registros não significativos e problemáticos, ou casos como campos de entrada com muitos valores omissos ou com uma variação muita alta ou muito baixa de serem úteis.
- **Ranqueamento.** Classifica o restante das entradas e designa classificações com base na importância.
- **Seleção.** Identifica o subconjunto de variáveis a serem utilizadas em modelos subsequentes, por exemplo, ao preservar apenas as entradas mais importantes e filtrar ou excluir todas as outras.

Em uma era em que muitas organizações estão lotadas de dados, os benefícios da seleção de variável para simplificar e acelerar o processo de modelagem podem ser substanciais. Ao focar sua atenção direto nos campos de maior importância, é possível reduzir a quantia de cálculo necessária, localizar mais

facilmente relacionamentos pequenos, porém importantes que poderiam de outra forma passarem despercebidos e, por fim, obter modelos mais simples, mais precisos e mais facilmente explicáveis. Ao reduzir o número de campos utilizados no modelo, você pode achar que é possível reduzir os tempos de escoragem e também a quantia de dados coletados em iterações futuras.

Exemplo. Uma empresa telefônica possui um armazém de dados contendo informações sobre respostas de uma promoção especial fornecidas por 5.000 clientes dessa empresa. Os dados incluem um grande número de campos contendo idades, profissão, renda e estatísticas de uso de telefone do cliente. Três campos de destino mostram se o cliente respondeu ou não a cada uma das três ofertas. A empresa deseja utilizar estes dados para ajudar a prever quais clientes têm maior probabilidade de responder a ofertas semelhantes no futuro.

Requisitos. Um campo de destino único (um com seu papel configurado para *Destino*), junto com diversos campos de entrada que você deseja verificar ou classificar com relação ao destino. Ambos os campos de destino e de entrada podem ter um nível de medição de *Contínuo* (intervalo numérico) ou *Categórico*.

Configurações do Modelo de Seleção de Variável

As configurações na guia Modelo incluem opções de modelo padrão com configurações que permitem fazer um ajuste preciso dos critérios de triagem de campos de entrada.

Nome do modelo É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Triagem de Campos de Entrada

A triagem envolve remover entradas ou casos que não incluírem nenhuma informação útil com relação ao relacionamento de entrada/destino. As opções de triagem baseiam-se em atributos do campo em questão sem considerar o poder preditivo com relação ao campo de destino selecionado. Os campos verificados são excluídos dos cálculos utilizados para classificar entradas e, opcionalmente, podem ser filtrados ou removidos dos dados utilizados na modelagem.

Os campos podem ser verificados com base nos critérios a seguir:

- **Porcentagem máxima de valores omissos**Verifica campos com muitos valores omissos, expressos como uma porcentagem do número total de registros. Os campos com uma porcentagem grande de valores omissos fornecem informações pouco preditivas.
- **Porcentagem máxima de registros em uma única categoria.** Verifica campos que tiverem registros caindo na mesma categoria com relação ao número total de registros. Por exemplo, se 95% dos clientes no banco de dados dirigirem o mesmo tipo de carro, incluir essa informação não é útil para distinguir um cliente do outro. Quaisquer campos que excederem o máximo especificado são verificados. Essa opção se aplica apenas a campos categóricos.
- **Número máximo de categorias como uma porcentagem de registros.** Verifica campos com muitas categorias com relação ao número total de registros. Se uma porcentagem grande das categorias contiver apenas um único caso, o campo poderá ser de uso limitado. Por exemplo, se cada cliente usar um chapéu diferente, essa informação não deverá ser útil para modelar padrões do comportamento. Essa opção se aplica apenas a campos categóricos.
- **Coeficiente mínimo de variação.** Verifica campos com um coeficiente de variância menor ou igual ao mínimo especificado. Esta medida é a razão entre o desvio padrão do campo de entrada com a média do campo de entrada. Se este valor estiver próximo a zero, não haverá muita variabilidade nos valores da variável. Essa opção se aplica apenas a campos contínuos (intervalo numérico).
- **Desvio padrão mínimo.** Verifica campos com um desvio padrão menor ou igual ao mínimo especificado. Essa opção se aplica apenas a campos contínuos (intervalo numérico).

Registros com dados omissos. Os registros ou casos que tiverem valores omissos para o campo de destino, ou valores omissos para todas as entradas, são excluídos automaticamente de todos os cálculos utilizados nos ranqueamentos.

Opções de Seleção de Variável

A guia Opções permite especificar as configurações padrão para selecionar ou excluir campos de entrada no nugget do modelo. Em seguida, é possível incluir o modelo em um fluxo para selecionar um subconjunto de campos para uso em esforços de construção de modelo subsequentes. Como alternativa, é possível substituir essas configurações ao selecionar ou cancelar seleção de campos adicionais no navegador do modelo após gerar o modelo. No entanto, as configurações padrão permitem aplicar o nugget do modelo sem mudanças adicionais, o que pode ser particularmente útil para propósitos de script.

Consulte o tópico "Resultados do Modelo de Seleção de Variável" na página 56 para obter mais informações.

As opções a seguir estão disponíveis:

Todos os campos classificados. Seleciona campos com base em seu ranqueamento, como *importante*, *marginal* ou *não importante*. É possível editar o rótulo de cada ranqueamento e também os valores de corte utilizados para designar registros para um ranqueamento ou outro.

Número máximo de campos. Seleciona os *n* principais campos com base na importância.

Importância maior que. Seleciona todos os campos com uma importância maior que o valor especificado.

O campo de destino é sempre preservado, independentemente da seleção.

Opções de Classificação de Importância

Tudo categórico. Quando todas as entradas e o destino forem categóricos, a importância poderá ser classificada com base em qualquer uma das quatro medidas:

- **Qui-quadrado de Pearson.** Testa a independência do destino e da entrada sem indicar a intensidade ou a direção de qualquer relacionamento existente.
- **Qui-quadrado de razão de verossimilhança.** Semelhante ao qui-quadrado de Pearson, mas também testa a independência do destino e da entrada.
- **V de Cramer.** Uma medida de associação com base na estatística qui-quadrado de Pearson. Os valores variam de 0, que indica nenhuma associação, a 1, que indica uma associação perfeita.
- **Lambda.** Uma medida de associação que reflete a redução proporcional em erro quando a variável é utilizada para prever o valor de destino. Um valor de 1 indica que o campo de entrada prediz o destino perfeitamente, ao passo que um valor de 0 significa que a entrada não fornece nenhuma informação útil sobre o destino.

Alguns categóricos. Quando algumas entradas, não todas, forem categóricas e o destino também for categórico, a importância poderá ser classificada com base no qui-quadrado de Pearson ou de razão de verossimilhança. (O *V* de Cramer e o lambda não estão disponíveis, a menos que todas as entradas sejam categóricas).

Categórico versus Contínuo. Ao classificar uma entrada categórica com relação a um destino contínuo ou vice-versa (um ou outro será categórico, mas não ambos), a estatística de *F* será utilizada.

Ambos contínuos. Ao classificar uma entrada contínua com relação a um destino contínuo, a estatística de *t* baseada no coeficiente de correlação será utilizada.

Nuggets do Modelo de Seleção de Variável

Os nuggets do modelo de Seleção de Variável exibem a importância de cada entrada com relação a um destino selecionado, conforme classificadas pelo nó Seleção de Variável. Todos os campos que tiverem sido verificados antes da classificação também são listados. Consulte o tópico "Nó Seleção de Variável" na página 53 para obter mais informações.

Ao executar um fluxo contendo um nugget do modelo de Seleção de Variável, o modelo atua como um filtro que preserva apenas as entradas selecionadas, conforme indicado pela seleção atual na guia Modelo. Por exemplo, é possível selecionar todos os campos classificados como importantes (uma das opções padrão) ou selecionar manualmente um subconjunto de campos na guia Modelo. O campo de destino também é preservado, independentemente da seleção. Todos os outros campos são excluídos.

A filtragem baseia-se apenas no nome do campo, por exemplo, se você selecionar *idade* e *renda*, qualquer campo que corresponder a um desses nomes será preservado. O modelo não atualiza classificações de campo com base em novos dados, ele apenas filtra os campos com base nos nomes selecionados. Por essa razão, deve-se tomar cuidado ao aplicar o modelo em dados novos ou atualizados. Na dúvida, é recomendado gerar o modelo novamente.

Resultados do Modelo de Seleção de Variável

A guia Modelo de um nugget do modelo Seleção de Variável exibe a classificação e a importância de todas as entradas na área de janela superior e permite selecionar os campos para filtragem utilizando as caixas de seleção na coluna à esquerda. Ao executar o fluxo, apenas os campos selecionados são preservados, e os outros campos são descartados. As seleções padrão baseiam-se nas opções especificadas no nó de construção de modelo, e também é possível selecionar ou cancelar seleção de campos adicionais conforme necessário.

A área de janela inferior lista as entradas que tiverem sido excluídas da classificação com base na porcentagem de valores omissos ou em outros critérios especificados no nó de modelagem. Assim como os campos classificados, é possível incluir ou descartar esses campos utilizando as caixas de seleção na coluna à esquerda. Consulte o tópico "Configurações do Modelo de Seleção de Variável" na página 54 para obter mais informações.

- Para ordenar a lista por classificação, nome do campo, importância ou qualquer uma das outras colunas exibidas, clique no cabeçalho da coluna. Ou, para usar a barra de ferramentas, selecione o item desejado na lista Ordenar Por e utilize as setas para cima e para baixo para alterar a direção da ordenação.
- É possível utilizar a barra de ferramentas para marcar ou desmarcar todos os campos e acessar a caixa de diálogo Verificar Campos, que permite selecionar os campos por classificação ou importância. Também é possível pressionar as teclas Shift e Ctrl enquanto clica nos campos para estender a seleção e utilizar a barra de espaço para ativar ou desativar um grupo de campos selecionados. Consulte o tópico "Selecionando Campos por Importância" para obter mais informações.
- Os valores limite para classificar entradas como importantes, marginais ou não importantes são exibidos na legenda abaixo da tabela. Esses valores são especificados no nó de modelagem. Consulte o tópico "Opções de Seleção de Variável" na página 55 para obter mais informações.

Selecionando Campos por Importância

Ao escorar dados usando um nugget do modelo Seleção de Variável, todos os campos selecionados na lista de campos classificados ou verificados -- conforme indicado pelas caixas de seleção na coluna à esquerda -- serão preservados. Outros campos serão descartados. Para alterar a seleção, é possível utilizar a barra de ferramentas para acessar a caixa de diálogo Verificar Campos, que permite selecionar os campos para classificação ou importância.

Todos os campos marcados. Seleciona todos os campos marcados como importantes, marginais ou não importantes.

Número máximo de campos. Permite selecionar um máximo de *n* campos com base na importância.

Importância maior que. Seleciona todos os campos com uma importância maior que o limite especificado.

Gerando um Filtro a partir de um Modelo de Seleção de Variável

Com base nos resultados de um modelo de Seleção de Variável, é possível utilizar Gerar Filtro na caixa de diálogo Variável para gerar um ou mais nós Filtro que incluem ou excluem subconjuntos de campos com base na importância relativa do destino especificado. Ao passo que o nugget do modelo também pode ser utilizado como um filtro, isso oferece a flexibilidade de experimentar subconjuntos diferentes de campos sem copiar ou modificar o modelo. O campo de destino é sempre preservado pelo filtro, independentemente se a inclusão ou exclusão estiver selecionada.

Incluir/Excluir. É possível escolher incluir ou excluir campos, por exemplo, para incluir os 10 principais campos ou excluir todos os campos marcados como não importantes.

Campos selecionados. Inclui ou exclui todos os campos selecionados atualmente na tabela.

Todos os campos marcados. Seleciona todos os campos marcados como importantes, marginais ou não importantes.

Número máximo de campos. Permite selecionar um máximo de *n* campos com base na importância.

Importância maior que. Seleciona todos os campos com uma importância maior que o limite especificado.

Nó de Detecção de Anomalias

Os modelos de detecção de anomalias são utilizados para identificar valores discrepantes, ou casos incomuns, nos dados. Diferentemente de outros métodos de modelagem que armazenam regras sobre casos incomuns, os modelos de detecção de anomalias armazenam informações sobre como é um comportamento normal. Isso permite identificar valores discrepantes mesmo se eles não corresponderem a nenhum padrão conhecido, e pode ser particularmente útil em aplicativos, como detecção de fraude, em que novos padrões surgem constantemente. A detecção de anomalias é um método não supervisionado, o que significa que ele não requer que um conjunto de dados de treinamento contendo casos conhecidos de fraude seja utilizado como um ponto de início.

Ao passo que métodos tradicionais de identificação de valores discrepantes geralmente examinam uma ou duas variáveis por vez, a detecção de anomalias pode examinar grandes números de campos para identificar grupos de clusters ou de peers nos quais registros semelhantes se enquadram. Em seguida, cada registro pode ser comparado com outros em seu grupo de peers para identificar possíveis anomalias. Quanto mais distante um caso estiver do centro normal, mais provavelmente incomum ele deverá ser. Por exemplo, o algoritmo pode agrupar registros em três clusters distintos e sinalizar aqueles que estiverem longe do centro de qualquer cluster.

Cada registro é designado a um índice de anomalias, que é a razão do índice de desvio de grupo com a sua média sobre o cluster ao qual o caso pertence. Quanto maior for o valor deste índice, mais desvios o caso terá além da média. Sob circunstâncias usuais, casos com valores de índice de anomalia menores que 1 ou até mesmo 1,5 não seriam considerados como anomalias porque o desvio é quase o mesmo ou um pouco mais que a média. No entanto, casos com um valor de índice maior que 2 podem ser bons candidatos a anomalias porque o desvio é pelo menos o dobro da média.

A detecção de anomalias é um método exploratório projetado para detecção rápida de casos ou registros incomuns que devem ser candidatos a uma análise adicional. Esses devem ser considerados como anomalias *suspeitas* que, sob uma análise mais detalhada, podem ou não vir a ser reais. Você pode achar

que um registro é perfeitamente válido, mas opta por verificá-lo a partir dos dados para propósitos de construção de modelo. Como alternativa, se o algoritmo descobrir repetidamente falsas anomalias, isso poderá apontar para um erro ou artefato no processo de coleção de dados.

Observe que a detecção de anomalias identifica registros ou casos incomuns por meio de análise de cluster com base no conjunto de campos selecionados no modelo sem levar em consideração nenhum campo de destino específico (dependente) e independentemente se esses campos forem relevantes ao padrão que você está tentando prever. Por essa razão, você pode querer utilizar a detecção de anomalias em combinação com a seleção de variável ou outra técnica para verificar e classificar campos. Por exemplo, é possível utilizar a seleção de variável para identificar os campos mais importantes com relação a um destino específico e, em seguida, usar a detecção de anomalias para localizar os registros que forem mais incomuns com relação a esses campos. (Uma abordagem alternativa seria a construção de um modelo de árvore de decisão e, em seguida, examinar quaisquer registros que forem classificados incorretamente como possíveis anomalias. No entanto, este método seria mais difícil para replicar ou automatizar em grande escala).

Exemplo. Na triagem de concessões de desenvolvimento agrícola para eventuais casos de fraude, a detecção de anomalias pode ser utilizada para descobrir desvios da norma, destacando os registros que forem anormais e que valem a pena realizar uma investigação adicional. Você está particularmente interessado em conceder aplicativos que parecem requerer muito (ou pouco) dinheiro de acordo com o tipo e tamanho da propriedade.

Requisitos. Um ou mais campos de entrada. Observe que apenas os campos com o papel configurado para **Entrada** usando uma origem ou nó Tipo podem ser utilizados como entradas. Os campos de destino com o papel configurado para **Destino** ou **Ambos** são ignorados.

Intensidades. Ao sinalizar casos que *não* estiverem em conformidade com um conjunto conhecido de regras ao invés daqueles que estão, os modelos de Detecção de Anomalias podem identificar casos excepcionais, mesmo quando eles não seguirem padrões conhecidos anteriormente. Quando utilizada em combinação com a seleção de variável, a detecção de anomalias permite verificar grandes quantias de dados para identificar os registros de maior interesse de modo relativamente rápido.

Opções do Modelo de Detecção de Anomalias

Nome do modelo É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Determinar valor de corte para anomalias com base em. Especifica o método utilizado para determinar o valor de corte para sinalizar anomalias. As opções a seguir estão disponíveis:

- **Nível de índice de anomalia mínimo.** Especifica o valor mínimo de corte para sinalizar anomalias. Os registros que atendem ou excedem esse limite são sinalizados.
- **Percentual dos registros mais anômalos nos dados de treinamento.** Configura automaticamente o limite em um nível que sinaliza a porcentagem especificada de registros nos dados de treinamento. O corte resultante é incluído como um parâmetro no modelo. Observe que esta opção determina como o valor de corte é configurado, *não* a porcentagem real de registros a serem sinalizados durante a escoragem. Os resultados da escoragem reais podem variar dependendo dos dados.
- **Número de registros mais anômalos nos dados de treinamento.** Configura automaticamente o limite em um nível que sinaliza o número especificado de registros nos dados de treinamento. O limite resultante é incluído como um parâmetro no modelo. Observe que esta opção determina como o valor de corte é configurado, *não* o número específico de registros a serem sinalizados durante a escoragem. Os resultados da escoragem reais podem variar dependendo dos dados.

Nota: independentemente de como o valor de corte é determinado, ele não afeta o valor de índice de anomalia subjacente relatado para cada registro. Ele simplesmente especifica o limite para sinalizar registros como anômalos durante a estimativa ou escoragem do modelo. Se você desejar posteriormente

examinar um número de registros maior ou menor, será possível usar um nó Seleção para identificar um subconjunto de registros com base no valor do índice de anomalia (`$O-AnomalyIndex > X`).

Número de campos de anomalias para relatório. Especifica o número de campos para relatar como uma indicação do porquê um determinado registro está sinalizado como uma anomalia. Os campos mais anômalos são relatados, definidos como aqueles que mostram o maior desvio da norma de campo para o cluster ao qual o registro foi designado.

Opções Avançadas de Detecção de Anomalias

Para especificar opções para valores omissos e outras configurações, configure o modo para **Especialista** na guia Especialista.

Coeficiente de ajustamento. Valor utilizado para balancear a ponderação relativa fornecida para campos contínuos (intervalo numérico) e categóricos no cálculo da distância. Valores maiores aumentam a influência dos campos contínuos. Este deve ser um valor diferente de zero.

Calcular automaticamente o número de grupos de peers. A detecção de anomalias pode ser utilizada para analisar rapidamente um grande número de soluções possíveis para escolher o número ideal de grupos de peers para os dados de treinamento. É possível ampliar ou restringir o intervalo ao configurar o número mínimo e máximo de grupos de peers. Valores maiores permitem que o sistema explore uma variedade mais ampla de soluções possíveis, mas à custa de um tempo de processamento maior.

Especificar o número de grupos de peers. Se você souber quantos clusters serão incluídos em seu modelo, selecione esta opção e insira o número de grupos de peers. Selecionar essa opção geralmente resulta em melhor desempenho.

Nível e razão de ruído. Essas configurações determinam como os valores discrepantes são tratados durante o armazenamento em cluster de dois estágios. No primeiro estágio, uma árvore de variável do cluster (CF) é utilizada para condensar os dados de um número muito grande de registros individuais para um número gerenciável de clusters. A árvore é construída com base em medidas de similaridade e, quando um nó da árvore obtém muitos registros, ele é dividido em nós filhos. No segundo estágio, o armazenamento em cluster hierárquico inicia nos nós terminais da árvore CF. O tratamento de ruído é ativado na primeira passagem de dados e desativado na segunda passagem de dados. Os casos no cluster de ruído na primeira passagem de dados são designados aos clusters regulares na segunda passagem de dados.

- **Nível de ruído.** Especifique um valor entre 0 e 0,5. Essa configuração será relevante apenas se a árvore CF for preenchida durante a fase de crescimento, significando que ela não poderá aceitar mais nenhum caso em um nó folha e que nenhum nó folha poderá ser dividido.

  Se a árvore CF for preenchida e o nível de ruído for configurado como 0, o limite será aumentado e a árvore CF crescerá novamente com todos os casos. Após o armazenamento em cluster final, os valores que não puderem ser designados a um cluster são valores discrepantes rotulados. O cluster de valor discrepante recebe um número de identificação de -1. O cluster de valor discrepante não é incluído na contagem do número de clusters; ou seja, se você especificar *n* clusters e o tratamento de ruído, o algoritmo gerará *n* clusters e um cluster de ruído. Na prática, aumentar este valor fornece ao algoritmo mais latitude para ajustar registros incomuns na árvore ao invés de designá-los para um cluster separado de valores discrepantes.

  Se a árvore CF for preenchida e o nível de ruído for maior que 0, a árvore CF crescerá novamente após colocar quaisquer dados em folhas esparsas em sua própria folha de ruído. Uma folha será considerada esparsa se a razão do número de casos na folha esparsa com o número de casos na maior folha for menor que o nível de ruído. Após a árvore crescer, os valores discrepantes serão colocados na árvore de CF se possível. Caso contrário, os valores discrepantes serão descartados para a segunda fase do armazenamento em cluster.
- **Razão de ruído.** Especifica a parte da memória alocada para o componente que deve ser utilizada para o armazenamento em buffer de ruído. Este valor varia entre 0,0 e 0,5. Se a inserção de um caso

específico em uma folha da árvore resultar em uma tensão menor que o limite, a folha não será dividida. Se a tensão exceder o limite, a folha será dividida, incluindo outro cluster pequeno na árvore CF. Na prática, aumentar essa configuração pode fazer com que o algoritmo se mova mais rapidamente para uma árvore mais simples.

Imputar valores omissos. Para campos contínuos, substitui a média de campo ao invés de quaisquer valores omissos. Para campos categóricos, as categorias omissas são combinadas e tratadas como uma categoria válida. Se essa opção estiver desmarcada, quaisquer registros com valores omissos serão excluídos da análise.

Nuggets do modelo de Detecção de Anomalias

Os nuggets do modelo de Detecção de Anomalias contêm todas as informações capturadas pelo modelo de Detecção de Anomalias e também informações sobre os dados de treinamento e sobre o processo de estimação.

Ao executar um fluxo que contém um nugget do modelo de Detecção de Anomalias, um número de novos campos é incluído no fluxo, conforme determinado pelas seleções feitas na guia Configurações na nugget do modelo. Consulte o tópico "Configurações do Modelo de Detecção de Anomalias" na página 61 para obter mais informações. Os novos nomes de campo baseiam-se no nome do modelo, precedidos por *$O*, conforme sumarizado na tabela a seguir.

Tabela 6. Geração de novo nome do campo.

Nome do campo	Descrição
$O-Anomaly	Campo de flag que indica se o registro é anômalo ou não.
$O-AnomalyIndex	O valor de índice de anomalia para o registro.
$O-PeerGroup	Especifica o grupo de peers ao qual o registro é designado.
$O-Field-n	Nome do *n*ésimo campo mais anômalo em termos de desvio da norma de cluster.
$O-FieldImpact-n	Índice de desvio de variável para o campo. Esse valor mede o desvio da norma de campo para o cluster ao qual o registro é designado.

Opcionalmente, é possível suprimir escores de registros não anômalos para tornar os resultados mais fáceis de ler. Consulte o tópico "Configurações do Modelo de Detecção de Anomalias" na página 61 para obter mais informações.

Detalhes do Modelo de Detecção de Anomalias

A guia Modelo de um modelo de Detecção de Anomalias gerado exibe informações sobre os grupos de peers no modelo.

Observe que os tamanhos e as estatísticas de grupo de peers relatados são estimativas baseadas nos dados de treinamento e podem diferir um pouco dos resultados da escoragem real, mesmo se executada nos mesmos dados.

Sumarização do Modelo de Detecção de Anomalias

A guia Sumarização de um nugget do modelo de Detecção de Anomalias exibe informações sobre os campos, sobre as configurações de construção e sobre o processo de estimação. O número de grupos de peers também é mostrado, junto com o valor de corte usado para sinalizar registros como anômalos.

Configurações do Modelo de Detecção de Anomalias

Use a guia Configurações para especificar opções para escorar o nugget do modelo.

Indicar registros anômalos com Especifica como os registros anômalos são tratados na saída.

- **Flag e índice** Cria um campo de flag que é configurado para *True* para todos os registros que excederem o valor de corte incluído no modelo. O índice de anomalia também é relatado para cada registro em um campo separado. Consulte o tópico "Opções do Modelo de Detecção de Anomalias" na página 58 para obter mais informações.
- **Apenas Flag** Cria um campo de flag, mas sem relatar o índice de anomalia para cada registro.
- **Apenas Índice** Relata o índice de anomalia sem criar um campo de flag.

Número de campos de anomalia para relatório Especifica o número de campos para relatar como uma indicação de por que um registro específico será sinalizado como uma anomalia. Os campos mais anômalos são relatados, definidos como aqueles que mostram o maior desvio da norma de campo para o cluster ao qual o registro foi designado.

Descartar registros Selecione esta opção para descartar todos os registros **Não anômalos** do fluxo, facilitando focar em possíveis anomalias em qualquer um dos nós de recebimento de dados. Como alternativa, é possível escolher descartar todos os registros **Anômalos** para limitar a análise subsequente para os registros que não estiverem sinalizados como possíveis anomalias com base no modelo.

Nota: Devido às mínimas diferenças no arredondamento, o número real de registros sinalizados durante a escoragem pode não ser idêntico ao número de registros sinalizados durante o treinamento do modelo, mesmo se executado nos mesmos dados.

Gerar SQL para este modelo Ao usar dados de um banco de dados, código SQL pode ser enviado por push de volta para o banco de dados para execução, fornecendo desempenho superior para muitas operações.

Selecione uma das opções a seguir para especificar como a geração de SQL é executada.

- **Padrão: Escorar usando o Server Scoring Adapter (se instalado) no processo** Se conectado a um banco de dados com um adaptador de escoragem instalado, gera a SQL utilizando o adaptador de escoragem e funções definidas pelo usuário (UDF) associadas e escora seu modelo no banco de dados. Quando nenhum adaptador de escoragem estiver disponível, essa opção busca seus dados novamente a partir do banco de dados e os escora no SPSS Modeler.
- **Escorar fora do Banco de dados** Se selecionada, esta opção busca seus dados novamente a partir do banco de dados e os escora no SPSS Modeler.

Capítulo 5. Nós de Modelagem Automatizados

Os nós de modelagem automatizados estimam e comparam um número de métodos de modelagem diferentes, permitindo experimentar uma variedade de abordagens em uma única execução de modelagem. É possível selecionar os algoritmos de modelagem para utilizar e as opções específicas para cada um deles, incluindo combinações que, de outra forma, seriam mutuamente exclusivas. Por exemplo, ao invés de escolher entre métodos rápidos, dinâmicos ou de poda para uma Rede Neural, é possível experimentar todos eles. O nó explora cada combinação possível de opções, classifica cada modelo candidato com base na medida que você especificar e salva o melhor modelo para uso na escoragem ou análise adicional.

É possível escolher entre três nós de modelagem automatizados, dependendo das necessidades de sua análise:

O nó Classificador Automático cria e compara um número de modelos diferentes de resultados binários (sim ou não, rotatividade ou não rotatividade, e assim por diante), permitindo escolher a melhor abordagem para uma análise específica. Diversos algoritmos de modelagem são suportados, possibilitando selecionar os métodos que deseja utilizar, as opções específicas para cada um deles e os critérios para comparar os resultados. O nó gera um conjunto de modelos com base nas opções especificadas e classifica os melhores candidatos de acordo com os critérios que você especificar.

O nó Previsor Contínuo Automático estima e compara modelos de resultados de intervalo numérico contínuos utilizando um número de métodos diferentes. O nó funciona da mesma maneira que o nó Previsor Categórico Automático, permitindo escolher os algoritmos a serem utilizados e experimentá-los com as diversas combinações de opções em uma única passagem de modelagem. Os algoritmos suportados incluem redes neurais, Árvore C&R, CHAID, regressão linear, regressão linear generalizada e Support Vector Machines (SVMs). Os modelos podem ser comparados com base na correlação, no erro relativo ou no número de variáveis utilizadas.

O nó Cluster Automático estima e compara modelos de armazenamento em cluster que identificam grupos de registros que possuem características semelhantes. O nó funciona da mesma maneira que outros nós de modelagem automatizados, permitindo experimentá-los com as diversas combinações de opções em uma única passagem de modelagem. Os modelos podem ser comparados utilizando medidas básicas com as quais é possível tentar filtrar e classificar a utilidade dos modelos de cluster e fornecer uma medida com base na importância de campos específicos.

Os melhores modelos são salvos em um nugget do modelo composto único, permitindo procurá-los, compará-los e escolher quais modelos usar na escoragem.

- Apenas para respostas binárias, nominais e numéricas, é possível selecionar diversos modelos de escoragem e combinar os escores em uma única combinação de modelo. Ao combinar predições a partir de diversos modelos, as limitações em modelos individuais podem ser evitadas, resultando normalmente em uma precisão geral maior do que pode ser obtida a partir de qualquer um dos modelos.
- Opcionalmente, é possível escolher se deseja realizar drill down dos resultados e gerar nós de modelagem ou nuggets do modelo para qualquer um dos modelos individuais que desejar usar ou explorar ainda mais.

Modelos e Tempo de Execução

Dependendo do conjunto de dados e do número de modelos, os nós de modelagem automatizados demoram horas ou até mesmo mais tempo para executar. Ao selecionar opções, preste atenção no número de modelos que estão sendo produzidos. Quando prático, você pode querer planejar execuções de modelagem durante a noite ou finais de semana quando os recursos do sistema estiverem menos propensos à demanda.

- Se necessário, um nó Partição ou Amostra pode ser usado para reduzir o número de registros incluídos na passagem de treinamento inicial. Após limitar as opções para alguns modelos candidatos, o conjunto de dados integral poderá ser restaurado.
- Para reduzir o número de campos de entrada, use a Seleção de Variável. Consulte o tópico "Nó Seleção de Variável" na página 53 para obter mais informações. Como alternativa, é possível utilizar as execuções de sua modelagem inicial para identificar campos e opções que valem a pena explorar ainda mais. Por exemplo, se os seus modelos de melhor desempenho parecerem usar os mesmos três campos, essa é uma forte indicação de que vale a pena manter esses campos.
- Opcionalmente, é possível limitar a quantia de tempo gasto estimando qualquer modelo específico e especificar as medidas de avaliação usadas para verificar e classificar modelos.

Configurações do Algoritmo do Nó de Modelagem Automatizado

Para cada tipo de modelo, é possível utilizar as configurações padrão ou escolher opções para cada tipo de modelo. As opções específicas são semelhantes àquelas disponíveis nos nós de modelagem separados, com a diferença de que, ao invés de escolher uma configuração ou outra, é possível escolher quantas desejar para aplicar na maioria dos casos. Por exemplo, se comparar modelos de Rede Neural, será possível escolher vários métodos de treinamento diferentes e experimentar cada método com e sem uma semente aleatória. Todas as combinações possíveis das opções selecionadas serão utilizadas, o que torna muito fácil gerar diversos modelos diferentes em uma única transmissão. No entanto, deve-se tomar cuidado, pois escolher diversas configurações pode fazer com que o número de modelos se multiplique muito rapidamente.

Para escolher opções para cada tipo de modelo

1. No nó de modelagem automatizada, selecione a guia **Especialista**.
2. Clique na coluna **Parâmetros de modelo** para o tipo de modelo.
3. No menu suspenso, escolha **Especificar**.
4. No diálogo **Configurações de algoritmo**, selecione as opções na coluna **Opções**.

Nota: opções adicionais estão disponíveis na guia Especialista do diálogo **Configurações de algoritmo**.

Regras de Parada do Nó de Modelagem Automatizada

As regras de parada especificadas para nós de modelagem automatizados estão relacionadas à execução do nó geral, e não à parada de modelos individuais construídos pelo nó.

Restringir o tempo de execução geral. (apenas modelos Rede Neural, K-Médias, Kohonen, TwoStep, SVM, KNN, Rede Bayesiana e Árvore C&R) Para a execução após um determinado número de horas). Todos os modelos gerados até esse ponto serão incluídos no nugget do modelo, mas nenhum modelo adicional será produzido.

Parar assim que modelos válidos forem produzidos. Para a execução quando um modelo transmitir todos os critérios especificados na guia Descartar (para o nó Classificador Automático ou Cluster Automático) ou na guia Modelo (para o nó Numeração Automática). Consulte o tópico "Opções de Descarte do Nó Classificador Automático" na página 70 para obter mais informações. Consulte o tópico "Opções de Descarte do Nó Cluster Automático" na página 77 para obter mais informações.

Nó Classificador Automático

O nó Classificador Automático estima e compara modelos com respostas nominais (conjunto) ou binárias (sim/não) utilizando diversos métodos diferentes, permitindo experimentar uma variedade de abordagens em uma única execução de modelagem. É possível selecionar os algoritmos a serem utilizados e experimentar com diversas combinações de opções. Por exemplo, ao invés de escolher entre métodos rápidos, dinâmicos ou de poda para uma Rede Neural, é possível experimentar todos eles. O nó explora cada combinação possível de opções, classifica cada modelo candidato com base na medida que você especificar e salva os melhores modelos para uso na escoragem ou análise adicional. Consulte o tópico Capítulo 5, "Nós de Modelagem Automatizados", na página 63 para obter mais informações.

Exemplo. Uma empresa de varejo possui dados históricos que rastreiam as ofertas feitas para clientes específicos em campanhas passadas. A empresa agora deseja obter resultados mais rentáveis ao corresponder a oferta certa para cada cliente.

Requisitos. Um campo de destino com um nível de medição de *Nominal* ou *Flag* (com o pepel configurado como **Destino**) e pelo menos um campo de entrada (com o papel configurado como **Entrada**). Para um campo de flag, o valor *True* definido para a resposta será assumido para representar uma ocorrência quando calcular o lucro, a elevação e estatísticas relacionadas. Os campos de entrada podem ter um nível de medição *Contínuo* ou *Categórico*, com a limitação de que algumas entradas podem não ser apropriadas para alguns tipos de modelo. Por exemplo, os campos ordinais utilizados como entradas em modelos Árvore C&R, CHAID e QUEST devem ter armazenamento numérico (não sequência de caracteres) e serão ignorados por estes modelos se especificado de outra forma. Da mesma forma, os campos de entrada contínuos podem ser categorizados em alguns casos. Os requisitos são os mesmos que quando utilizar os nós de modelagem individuais, por exemplo, um modelo de Rede Bayesiana funciona da mesma forma, independentemente se for gerado a partir do nó Rede Bayesiana ou do nó Classificador Automático.

Frequência e campos de ponderação. A frequência e a ponderação são utilizadas para dar importância extra para alguns registros sobre outros porque, por exemplo, quando o usuário sabe que o conjunto de dados de construção sub-representa uma parte da população pai (Ponderação) ou porque um registro representa um número de casos idênticos (Frequência). Se especificado, um campo de frequência poderá ser utilizado pelos modelos de Árvore C&R, CHAID, QUEST, Lista de Decisão e Rede Bayesiana. Um campo de ponderação pode ser utilizado pelos modelos C&RT, CHAID e C5.0 Outros tipos de modelo ignorarão esses campos e construirão os modelos de qualquer maneira. Os campos de frequência e de ponderação são utilizados apenas para construção de modelo e não são considerados ao avaliar ou escorar os modelos. Consulte o tópico "Usando Campos de Frequência e de Ponderação" na página 33 para obter mais informações.

Prefixos. Se anexar um nó de tabela ao nugget do Nó Classificador Automático ou do Nó Numeração Automática, haverá várias novas variáveis na tabela com nomes que começam com um dos seguintes prefixos: *$G*, *$GE*, *$R*, *$XR* e *$XRE*.

Por convenção, os nomes dos campos gerados durante a escoragem baseiam-se no campo de destino, mas com um prefixo padrão. Os prefixos $G e $GE são gerados pelo Modelo Linear Generalizado, $R é o prefixo utilizado para a predição gerada pelo modelo CHAID nesse caso, $X normalmente é gerado utilizando uma combinação e $XR, $XS e $XF são utilizados como prefixos nos casos em que o campo de destino for um campo Contínuo, Categórico ou de Flag, respectivamente. Tipos de modelo diferentes utilizam conjuntos diferentes de prefixos.

Tipos de Modelo Suportados

Os tipos de modelos suportados incluem Rede Neural, Árvore C&R, QUEST, CHAID, C5.0, Regressão Logística, Lista de Decisão, Rede Bayesiana, Discriminante, Vizinho Mais Próximo e SVM. Consulte o tópico "Opções Avançadas do Nó Classificador Automático" na página 67 para obter mais informações.

Opções de Modelo do Nó Classificador Automático

A guia Modelo do nó Classificador Automático permite especificar o número de modelos a serem criados, com os critérios utilizados para comparar os modelos.

Nome do modelo É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Utilizar dados particionados. Se um campo de partição for definido, essa opção assegurará que apenas os dados da partição de treinamento sejam utilizados para construir o modelo.

Criar modelos de divisão. Constrói um modelo separado para cada valor possível de campos de entrada que são especificados como campos de divisão. Consulte o tópico "Construindo Modelos de Divisão" na página 28 para obter mais informações.

Classificar modelos por. Especifica os critérios utilizados para comparar e classificar os modelos. As opções incluem a precisão geral, a área sob a curva ROC, o lucro, a elevação e o número de campos. Observe que todas estas medidas estarão disponíveis no relatório sumarização, independentemente do que estiver selecionado aqui.

Nota: para um destino nominal (conjunto), o ranqueamento é restrito a uma **Precisão geral** ou **Número de Campos**.

Ao calcular lucros, elevação e estatísticas relacionadas, o valor *True* definido para o campo de destino é assumido para representar uma ocorrência.

- **Precisão geral** A porcentagem de registros corretamente predita pelo modelo com relação ao número total de registros.
- **Área na curva ROC** A curva ROC fornece um índice para o desempenho de um modelo. Quanto mais distante a curva estiver acima da linha de referência, mais preciso será o teste.
- **Lucro (Acumulativo)** A soma de lucros em percentuais acumulativos (ordenados em termos de confiança para a predição), conforme calculado com base nos critérios de custo, renda e ponderação especificados. Geralmente, o lucro começa perto de 0 para o percentil superior, aumenta de forma constante e, em seguida, diminui. Para obter um bom modelo, os lucros mostrarão um pico bem-definido, que é relatado juntamente ao percentil no qual ele ocorre. Para um modelo que não fornece nenhuma informação, a curva de lucro será relativamente reta e poderá aumentar, diminuir ou nivelar, dependendo da estrutura de custo/renda aplicada.
- **Elevação (Acumulativo)** A razão de ocorrências em quantis acumulativos com relação à amostra geral (em que os quantis são ordenados em termos de confiança para a predição). Por exemplo, um valor de elevação de 3 para o quantil superior indica uma taxa de acerto três vezes maior que a amostra geral. Para obter um bom modelo, a elevação deve iniciar bem acima de 1,0 para os quantis superiores e, em seguida, cair acentuadamente para 1,0 para os quantis inferiores. Para um modelo que não fornece nenhuma informação, a elevação ficará em torno de 1,0.
- **Número de campos** Modelos de ranqueamento com base no número de campos de entrada utilizados.

Classificar modelos usando. Se uma partição estiver em uso, será possível especificar se os ranqueamentos baseiam-se no conjunto de dados de treinamento ou no conjunto de testes. Com grandes conjuntos de dados, utilizar uma partição para triagem preliminar de modelos poderá melhorar enormemente o desempenho.

Número de modelos a serem utilizados. Especifica o número máximo de modelos a serem listados no nugget do modelo produzido pelo nó. Os modelos com maior classificação são listados de acordo com o critério de classificação especificado. Observe que aumentar esse limite poderá diminuir o desempenho. O valor máximo permitido é 100.

Calcular a importância do preditor. Para modelos que produzem uma medida apropriada de importância, é possível exibir um gráfico que indica a importância relativa de cada preditor na estimativa do modelo. Geralmente, você deseja concentrar seus esforços de modelagem nos preditores que forem de maior importância e considerar eliminar ou ignorar aqueles que forem de menor importância. Observe que a importância do preditor poderá estender o tempo necessário para calcular alguns modelos, e não será recomendado se você simplesmente desejar uma comparação geral em muitos modelos diferentes. Isso é mais útil quando você tiver restringido sua análise para uma quantia de modelos que você desejar explorar detalhadamente. Consulte o tópico "Importância do preditor" na página 44 para obter mais informações.

Critérios de Lucro. *Nota*. Apenas para respostas flags. Lucro é igual à renda de cada registro, menos o custo do registro. Os lucros para um quantil são simplesmente a soma dos lucros de todos os registros no quantil. Considera-se que os lucros se aplicam apenas às ocorrências e que os custos se aplicam a todos os registros.

- **Custos.** Especifique o custo associado a cada registro. É possível selecionar custos **Fixo** ou **Variável**. Para custos fixos, especifique o valor do custo. Para custos variáveis, clique no botão Seletor de Campo para selecionar um campo como o campo de custo. (**Custos** não está disponível para gráficos ROC).
- **Renda.** Especifique a renda associada a cada registro que representa uma ocorrência. É possível selecionar custos **Fixo** ou **Variável**. Para rendas fixas, especifique o valor da renda. Para renda variável, clique no botão Seletor de Campo para selecionar um campo como o campo de renda. (**Renda** não está disponível para gráficos ROC).
- **Ponderação.** Se os registros em seus dados representarem mais de uma unidade, será possível usar ponderações de frequência para ajustar os resultados. Especifique a ponderação associada a cada registro usando as ponderações **Fixa** ou **Variável**. Para ponderações fixas, especifique o valor da ponderação (o número de unidades por registro). Para ponderações variáveis, clique no botão Seletor de Campo para selecionar um campo como o campo de ponderação. (**Ponderação** não está disponível para gráficos ROC).

Critérios de Elevação. *Nota*. Apenas para respostas flags. Especifica o percentil a ser utilizado para cálculos de elevação. Observe que também é possível alterar este valor quando comparar os resultados. Consulte o tópico "Nuggets do Modelo Automatizado" na página 77 para obter mais informações.

Opções Avançadas do Nó Classificador Automático

A guia Especialista do nó Classificador Automático permite aplicar uma partição (se disponível), selecionar os algoritmos a serem utilizados e especificar as regras de parada.

Selecionar modelos. Por padrão, todos os modelos são selecionados para serem construídos, entretanto, se você tiver o Analytic Server, será possível optar por restringir os modelos para aqueles no qual eles podem ser executados.Analytic Server e pré-configurá-los para que eles construam modelos de divisão ou que estejam prontos para processar conjuntos de dados muito grandes.

Modelos usados. Use as caixas de seleção na coluna à esquerda para selecionar os tipos de modelo (algoritmo) para incluir na comparação. Quanto mais tipos você selecionar, mais modelos serão criados e mais demorado o processamento será.

Tipo de modelo. Lista os algoritmos disponíveis (consulte abaixo).

Parâmetro de modelo. Para cada tipo de modelo, é possível utilizar as configurações padrão ou selecionar **Especificar** para escolher opções para cada tipo de modelo. As opções específicas são semelhantes àquelas disponíveis nos nós de modelagem separados, com a diferença de que diversas opções ou combinações podem ser selecionadas. Por exemplo, se comparar os modelos de Rede Neural, ao invés de escolher um dos seis métodos de treinamento, é possível escolher todos eles para treinar seis modelos em um único passo.

Número de modelos. Lista o número de modelos produzidos para cada algoritmo com base nas configurações atuais. Ao combinar opções, o número de modelos pode aumentar rapidamente, portanto, é altamente recomendado prestar atenção nesse número, principalmente quando usar grandes conjuntos de dados.

Tempo máximo restrito gasto construindo um único modelo. (apenas modelos K-Médias, Kohonen, TwoStep, SVM, KNN, Bayes Net e Lista de Decisão) Configura um limite de tempo máximo para qualquer modelo. Por exemplo, se um modelo específico requerer um longo tempo inesperado para treinar devido a alguma interação complexa, você provavelmente não vai querer que isso atrase toda a execução de modelagem.

Nota: se o destino for um campo nominal (conjunto), a opção Lista de Decisão estará indisponível.

Algoritmos Suportados

O nó Rede Neural utiliza um modelo simplificado da maneira com que o cérebro humano processa informações. Ele funciona ao simular um grande número de unidades de processamento simples interconectadas que lembram versões de neurônios abstratas. As redes neurais são estimadores de função geral poderosos que requerem conhecimento mínimo em estatística ou matemática para treinamento ou aplicação.

O nó C5.0 constrói uma árvore de decisão ou um conjunto de regras. O modelo funciona dividindo a amostra com base no campo que fornece o ganho máximo de informações em cada nível. O campo de destino deve ser categórico. Diversas divisões em mais de dois subgrupos são permitidas.

O nó Árvore de Classificação e Regressão (C&R) gera uma árvore de decisão que permite prever ou classificar observações futuras. O método utiliza particionamento recursivo para dividir os registros de treinamento em segmentos ao minimizar a impureza em cada passo, em que um nó na árvore será considerado "puro" se 100% dos casos no nó caírem em uma categoria específica do campo de destino. Os campos de destino e de entrada podem ser intervalos numéricos ou categóricos (nominal, ordinal ou sinalizadores) e todas as divisões são binárias (somente dois subgrupos).

O nó QUEST fornece um método de classificação binária para construir árvores de decisão, projetado para reduzir o tempo de processamento necessário para grandes análises de Árvore C&R enquanto também reduz a tendência localizada nos métodos de árvore de classificação para favorecer entradas que permitem mais divisões. Os campos de entrada podem ser intervalos numéricos (conínuo), ao passo que o campo de destino deve ser categórico. Todas as divisões são binárias.

O nó CHAID gera árvores de decisão usando estatísticas qui-quadrado para identificar divisões ideais. Diferentemente dos nós Árvore C&R e QUEST, o CHAID pode gerar árvores não binárias, o que significa que algumas divisões possuem mais de duas ramificações. Os campos de destino e de entrada podem ser um intervalo numérico (contínuo) ou categóricos. Um Exhaustive CHAID é uma modificação de CHAID que executa uma tarefa mais completa de examinar todas as possíveis divisões, mas demora mais tempo para calcular.

A regressão logística é uma técnica estatística para classificar registros com base em valores de campos de entrada. Ela é semelhante a uma regressão linear, mas usa um campo de destino categórico ao invés de um intervalo numérico.

O nó Lista de Decisão identifica os subgrupos ou segmentos, que mostram uma probabilidade maior ou menor de um resultado binário fornecido com relação à população geral. Por exemplo, é possível procurar por clientes que forem menos propensos a migrarem para o concorrente ou que responderão favoravelmente a uma campanha. É possível incorporar o conhecimento dos negócios no modelo ao incluir seus próprios segmentos customizados e visualizar modelos alternativos lado a lado para comparar os resultados. Os modelos de Lista de Decisão consistem em uma lista de regras em que cada regra possui uma condição e um resultado. As regras são aplicadas na ordem, e a primeira regra que corresponder determina o resultado.

O nó Rede Bayesiana permite construir um modelo de probabilidade ao combinar evidências observada e registrada com conhecimento do mundo real para estabelecer a probabilidade das ocorrências. O nó foca nas redes Tree Augmented Naïve Bayes (TAN) e Markov Blanket que são utilizadas principalmente para classificação.

A análise discriminante faz suposições mais rígidas do que a regressão logística, mas pode ser uma alternativa ou um complemento poderoso para uma análise de regressão logística quando essas suposições forem atendidas.

O nó *k*-Nearest Neighbor (KNN) associa um novo caso à categoria ou valor dos *k* objetos mais próximos a ele no espaço do preditor, em que *k* é um número inteiro. Os casos semelhantes estão próximos uns dos outros e os casos dissimilares estão distantes.

O nó Support Vector Machine (SVM) permite classificar dados em um dos dois grupos sem super ajuste. O SVM funciona bem com conjuntos de dados grandes, como aqueles com um número muito grande de campos de entrada.

Custos de classificação errada

Em alguns contextos, determinados tipos de erros são mais caros que outros. Por exemplo, pode ser mais caro classificar um solicitante de crédito de alto risco como baixo risco (um tipo de erro) do que classificar um solicitante de baixo risco como alto risco (um tipo diferente de erro). Os custos de classificação errada permitem especificar a importância relativa de diferentes tipos de erros de predição.

Os custos de classificação errada são basicamente ponderações aplicadas a resultados específicos. Essas ponderações são fatoradas no modelo e podem, na realidade, alterar a predição (como uma forma de proteger contra erros caros).

Com exceção dos modelos do C5.0, os custos de classificação errada não serão aplicados ao escorar um modelo e não são levados em conta quando classificar ou comparar modelos usando um nó Classificador Automático, gráfico de avaliação, ou nó Análise. Um modelo que inclui custos poderá não produzir menos erros do que aquele que não inclui e poderá não ter uma classificação mais alta em termos de precisão geral, mas provavelmente executará melhor em termos práticos por possuir um viés integrado a favor de erros *menos caros*.

A matriz de custo mostra o custo para cada combinação possível de categoria predita e categoria real. Por padrão, todos os custos de classificação errada são configurados como 1,0. Para inserir valores de custo customizado, selecione **Usar custos de classificação errada** e insira os valores customizados na matriz de custo.

Para alterar um custo de classificação errada, selecione a célula correspondente à combinação desejada de valores preditos e reais, exclua o conteúdo existente da célula e insira o custo desejado para a célula. Os

custos não são simétricos automaticamente. Por exemplo, se você configurar o custo de classificação errada de *A* como *B* para 2,0, o custo da classificação errada de *B* como *A* ainda terá o valor padrão de 1,0, a menos que você também o altere explicitamente.

Opções de Descarte do Nó Classificador Automático

A guia Descarte do nó Classificador Automático permite descartar automaticamente modelos que não atenderem a determinados critérios. Esses modelos não serão listados no relatório sumarização.

É possível especificar um limite mínimo para a precisão geral e um limite máximo para o número de variáveis utilizadas no modelo. Além disso, para respostas de flag, é possível especificar um limite mínimo para elevação, lucro e área sob a curva; a elevação e o lucro são determinados conforme especificado na guia Modelo. Consulte o tópico "Opções de Modelo do Nó Classificador Automático" na página 66 para obter mais informações.

Opcionalmente, é possível configurar o nó para parar a execução na primeira vez em que um modelo que atender a todos os critérios especificados for gerado. Consulte o tópico "Regras de Parada do Nó de Modelagem Automatizada" na página 64 para obter mais informações.

Opções de Configurações do Nó Classificador Automático

A guia Configurações do nó Classificador Automático permite pré-configurar as opções de escoragem de tempo que estão disponíveis no nugget.

Filtrar campos gerados por modelos combinados. Remove da saída todos os campos adicionais gerados pelos modelos individuais que são alimentados no nó Combinação. Marque esta caixa de seleção se você estiver interessado apenas no escore combinado de todos os modelos de entrada. Assegure-se de que esta opção esteja desmarcada se, por exemplo, desejar utilizar um nó Análise ou nó Avaliação para comparar a precisão do escore combinado com a precisão de cada um dos modelos de entrada individuais.

Nó Numeração Automática

O nó Numeração Automática estima e compara modelos para resultados de intervalo numérico contínuo utilizando um número de métodos diferentes, permitindo experimentar uma variedade de abordagens em uma única execução de modelagem. É possível selecionar os algoritmos a serem utilizados e experimentar com diversas combinações de opções. Por exemplo, é possível prever valores domésticos utilizando modelos de rede neural, regressão linear, C&RT e CHAID para ver quais deles executam melhor, bem como experimentar diferentes combinações de métodos de regressão stepwise, forward e backward. O nó explora cada combinação possível de opções, classifica cada modelo candidato com base na medida que você especificar e salva o melhor modelo para uso na escoragem ou análise adicional. Consulte o tópico Capítulo 5, "Nós de Modelagem Automatizados", na página 63 para obter mais informações.

Exemplo. Um município deseja estimar com mais precisão impostos de imóveis e ajustar valores para propriedades específicas conforme necessário sem precisar inspecionar cada propriedade. Utilizando o nó Numeração Automática, o analista pode gerar e comparar um número de modelos que preveem valores de propriedade com base no tipo de construção, localização, tamanho e outros fatores conhecidos.

Requisitos. Um campo de destino único (com o papel configurado como **Destino**) e pelo menos um campo de entrada (com o papel configurado como **Entrada**). O destino deve ser um campo contínuo (intervalo numérico), como *idade* ou *renda*. Os campos de entrada podem ser contínuos ou categóricos, com a limitação de que algumas entradas podem não ser apropriadas para alguns tipos de modelo. Por exemplo, os modelos de Árvore C&R podem utilizar campos de sequência de caracteres categóricos como entradas, ao passo que os modelos de regressão linear não podem utilizar esses campos e os ignorará, se especificado. Os requisitos são os mesmos que quando utilizar nós de modelagem individuais. Por exemplo, um modelo CHAID funciona da mesma forma, independentemente se gerado a partir do nó CHAID ou do nó Numeração Automática.

Frequência e campos de ponderação. A frequência e a ponderação são utilizadas para dar importância extra para alguns registros sobre outros porque, por exemplo, quando o usuário sabe que o conjunto de dados de construção sub-representa uma parte da população pai (Ponderação) ou porque um registro representa um número de casos idênticos (Frequência). Se especificado, um campo de frequência poderá ser utilizado pelos algoritmos Árvore C&R e CHAID. Um campo de ponderação pode ser utilizado pelos algoritmos C&RT, CHAID, Regressão e GenLin Outros tipos de modelo ignorarão esses campos e construirão os modelos de qualquer maneira. Os campos Frequência e Ponderação são utilizados apenas para construção de modelo e não são considerados ao avaliar ou escorar os modelos. Consulte o tópico "Usando Campos de Frequência e de Ponderação" na página 33 para obter mais informações.

Prefixos. Se anexar um nó de tabela ao nugget do Nó Classificador Automático ou do Nó Numeração Automática, haverá várias novas variáveis na tabela com nomes que começam com um dos seguintes prefixos: *$G*, *$GE*, *$R*, *$XR* e *$XRE*.

Por convenção, os nomes dos campos gerados durante a escoragem baseiam-se no campo de destino, mas com um prefixo padrão. Os prefixos $G e $GE são gerados pelo Modelo Linear Generalizado, $R é o prefixo utilizado para a predição gerada pelo modelo CHAID nesse caso, $X normalmente é gerado utilizando uma combinação e $XR, $XS e $XF são utilizados como prefixos nos casos em que o campo de destino for um campo Contínuo, Categórico ou de Flag, respectivamente. Tipos de modelo diferentes utilizam conjuntos diferentes de prefixos.

Tipos de Modelo Suportados

Os tipos de modelos suportados incluem Rede Neural, Árvore C&R, CHAID, Regressão, GenLin, Vizinho Mais Próximo e SVM. Consulte o tópico "Opções Avançadas do Nó Numeração Automática" na página 72 para obter mais informações.

Opções de Modelo do Nó Numeração Automática

A guia Modelo do nó Numeração Automática permite especificar o número de modelos a serem salvos, com os critérios utilizados para comparar os modelos.

Nome do modelo É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Utilizar dados particionados. Se um campo de partição for definido, essa opção assegurará que apenas os dados da partição de treinamento sejam utilizados para construir o modelo.

Criar modelos de divisão. Constrói um modelo separado para cada valor possível de campos de entrada que são especificados como campos de divisão. Consulte o tópico "Construindo Modelos de Divisão" na página 28 para obter mais informações.

Classificar modelos por. Especifica os critérios utilizados para comparar modelos.

- **Correlação.** A correlação de Pearson entre o valor observado para cada registro e o valor predito pelo modelo. A correlação é uma medida de associação linear entre duas variáveis, em que os valores mais próximos de 1 indicam um relacionamento mais forte. (Os valores de correlação variam entre -1, para um relacionamento negativo perfeito, e 1 para um relacionamento positivo perfeito. Um valor de 0 indica que não há relacionamento linear, ao passo que um modelo com uma correlação negativa teria um ranqueamento menor de todos).
- **Número de campos.** O número de campos utilizados como preditores no modelo. Escolhendo os modelos que utilizam menos campos pode aperfeiçoar a preparação de dados e melhorar o desempenho em alguns casos.
- **Erro relativo.** O erro relativo é a razão da variância dos valores observados a partir daqueles preditos pelo modelo com a variância dos valores observados na média. Na prática, ele compara o grau de desempenho do modelo com um modelo **nulo** ou **intercepto** que simplesmente retorna o valor médio

do campo de destino como a predição. Para obter um bom modelo, esse valor deve ser menor que 1, indicando que o modelo é mais preciso que o modelo nulo. Um modelo com um erro relativo maior que 1 é menos preciso do que o modelo nulo e, portanto, não é útil. Para modelos de regressão linear, o erro relativo é igual ao quadrado da correlação e não inclui novas informações. Para modelos não lineares, o erro relativo não está relacionado à correlação e fornece uma medida adicional para avaliar o desempenho do modelo.

Classificar modelos usando. Se uma partição estiver em uso, será possível especificar se os ranqueamentos baseiam-se na partição de treinamento ou na partição de testes. Com grandes conjuntos de dados, utilizar uma partição para triagem preliminar de modelos poderá melhorar enormemente o desempenho.

Número de modelos a serem utilizados. Especifica o número máximo de modelos a serem mostrados no nugget do modelo produzido pelo nó. Os modelos com maior classificação são listados de acordo com o critério de classificação especificado. Aumentar este limite permite comparar os resultados de mais modelos, mas pode diminuir o desempenho. O valor máximo permitido é 100.

Calcular a importância do preditor. Para modelos que produzem uma medida apropriada de importância, é possível exibir um gráfico que indica a importância relativa de cada preditor na estimativa do modelo. Geralmente, você deseja concentrar seus esforços de modelagem nos preditores que forem de maior importância e considerar eliminar ou ignorar aqueles que forem de menor importância. Observe que a importância do preditor poderá estender o tempo necessário para calcular alguns modelos, e não será recomendado se você simplesmente desejar uma comparação geral em muitos modelos diferentes. Isso é mais útil quando você tiver restringido sua análise para uma quantia de modelos que você desejar explorar detalhadamente. Consulte o tópico “Importância do preditor” na página 44 para obter mais informações.

Não manter modelos se. Especifica valores do limite para correlação, erro relativo e número de campos utilizados. Os modelos que não atenderem a nenhum desses critérios serão descartados e não serão listados no relatório sumarização.

- **Correlação menor que.** A correlação mínima (em termos de valor absoluto) para um modelo a ser incluído no relatório sumarização.
- **Número de campos utilizados é maior que.** O número máximo de campos a serem utilizados por qualquer modelo a ser incluído.
- **Erro relativo é maior que.** O erro máximo relativo para qualquer modelo a ser incluído.

Opcionalmente, é possível configurar o nó para parar a execução na primeira vez em que um modelo que atender a todos os critérios especificados for gerado. Consulte o tópico “Regras de Parada do Nó de Modelagem Automatizada” na página 64 para obter mais informações.

Opções Avançadas do Nó Numeração Automática

A guia Especialista do nó Numeração Automática permite selecionar os algoritmos e as opções a serem utilizados e especificar as regras de parada.

Selecionar modelos. Por padrão, todos os modelos são selecionados para serem construídos, entretanto, se você tiver o Analytic Server, será possível optar por restringir os modelos para aqueles no qual eles podem ser executados.Analytic Server e pré-configurá-los para que eles construam modelos de divisão ou que estejam prontos para processar conjuntos de dados muito grandes.

Modelos usados. Use as caixas de seleção na coluna à esquerda para selecionar os tipos de modelo (algoritmo) para incluir na comparação. Quanto mais tipos você selecionar, mais modelos serão criados e mais demorado o processamento será.

Tipo de modelo. Lista os algoritmos disponíveis (consulte abaixo).

Parâmetro de modelo. Para cada tipo de modelo, é possível utilizar as configurações padrão ou selecionar **Especificar** para escolher opções para cada tipo de modelo. As opções específicas são semelhantes àquelas disponíveis nos nós de modelagem separados, com a diferença de que diversas opções ou combinações podem ser selecionadas. Por exemplo, se comparar os modelos de Rede Neural, ao invés de escolher um dos seis métodos de treinamento, é possível escolher todos eles para treinar seis modelos em um único passo.

Número de modelos. Lista o número de modelos produzidos para cada algoritmo com base nas configurações atuais. Ao combinar opções, o número de modelos pode aumentar rapidamente, portanto, é altamente recomendado prestar atenção nesse número, principalmente quando usar grandes conjuntos de dados.

Tempo máximo restrito gasto construindo um único modelo. (apenas modelos K-Médias, Kohonen, TwoStep, SVM, KNN, Bayes Net e Lista de Decisão) Configura um limite de tempo máximo para qualquer modelo. Por exemplo, se um modelo específico requerer um longo tempo inesperado para treinar devido a alguma interação complexa, você provavelmente não vai querer que isso atrase toda a execução de modelagem.

Algoritmos Suportados

O nó Rede Neural utiliza um modelo simplificado da maneira com que o cérebro humano processa informações. Ele funciona ao simular um grande número de unidades de processamento simples interconectadas que lembram versões de neurônios abstratas. As redes neurais são estimadores de função geral poderosos que requerem conhecimento mínimo em estatística ou matemática para treinamento ou aplicação.

O nó Árvore de Classificação e Regressão (C&R) gera uma árvore de decisão que permite prever ou classificar observações futuras. O método utiliza particionamento recursivo para dividir os registros de treinamento em segmentos ao minimizar a impureza em cada passo, em que um nó na árvore será considerado "puro" se 100% dos casos no nó caírem em uma categoria específica do campo de destino. Os campos de destino e de entrada podem ser intervalos numéricos ou categóricos (nominal, ordinal ou sinalizadores) e todas as divisões são binárias (somente dois subgrupos).

O nó CHAID gera árvores de decisão usando estatísticas qui-quadrado para identificar divisões ideais. Diferentemente dos nós Árvore C&R e QUEST, o CHAID pode gerar árvores não binárias, o que significa que algumas divisões possuem mais de duas ramificações. Os campos de destino e de entrada podem ser um intervalo numérico (contínuo) ou categóricos. Um Exhaustive CHAID é uma modificação de CHAID que executa uma tarefa mais completa de examinar todas as possíveis divisões, mas demora mais tempo para calcular.

A regressão linear é uma técnica estatística comum para sumarizar dados e fazer predições ao ajustar uma linha ou superfície reta que minimiza as discrepâncias entre os valores de saída preditos e reais.

O modelo Linear Generalizado expande o modelo linear geral para que a variável dependente seja linearmente relacionada aos fatores e covariáveis por meio de uma função de ligação especificada. Além disso, o modelo permite à variável dependente ter uma distribuição não normal. Ele cobre a funcionalidade de um grande número de modelos estatísticos, incluindo regressão linear, regressão logística, modelos de log-linear para dados de contagem e modelos de sobrevivência censurados por intervalo.

O nó *k*-Nearest Neighbor (KNN) associa um novo caso à categoria ou valor dos *k* objetos mais próximos a ele no espaço do preditor, em que *k* é um número inteiro. Os casos semelhantes estão próximos uns dos outros e os casos dissimilares estão distantes.

O nó Support Vector Machine (SVM) permite classificar dados em um dos dois grupos sem super ajuste. O SVM funciona bem com conjuntos de dados grandes, como aqueles com um número muito grande de campos de entrada.

Os modelos de regressão lineares preveem uma variável resposta contínua com base em relacionamentos lineares entre o destino e um ou mais preditores.

Opções de Configurações do Nó Numeração Automática

A guia Configurações do nó Numeração Automática permite pré-configurar as opções de escoragem de tempo que estão disponíveis no nugget.

Filtrar campos gerados por modelos combinados. Remove da saída todos os campos adicionais gerados pelos modelos individuais que são alimentados no nó Combinação. Marque esta caixa de seleção se você estiver interessado apenas no escore combinado de todos os modelos de entrada. Assegure-se de que esta opção esteja desmarcada se, por exemplo, desejar utilizar um nó Análise ou nó Avaliação para comparar a precisão do escore combinado com a precisão de cada um dos modelos de entrada individuais.

Calcular erro padrão. Para uma variável resposta contínua (intervalo numérico), um cálculo de erro padrão é executado por padrão para calcular a diferença entre os valores medidos ou estimados e os valores reais e para mostrar quão próximo essas estimativas corresponderam.

Nó Cluster Automático

O nó Cluster Automático estima e compara modelos de armazenamento em cluster que identificam grupos de registros com características semelhantes. O nó funciona da mesma maneira que outros nós de modelagem automatizados, permitindo experimentá-los com as diversas combinações de opções em uma única passagem de modelagem. Os modelos podem ser comparados utilizando medidas básicas com as quais é possível tentar filtrar e classificar a utilidade dos modelos de cluster e fornecer uma medida com base na importância de campos específicos.

Os modelos de armazenamento em cluster geralmente são utilizados para identificar grupos que podem ser usados como entradas em análises subsequentes. Por exemplo, você pode querer almejar grupos de clientes com base em características demográficas, como rendimento, ou com base nos serviços que eles compraram no passado. Isso pode ser feito sem conhecer previamente os grupos e suas características -- você pode não saber quantos grupos procurar ou quais variáveis utilizar na definição deles. Os modelos de armazenamento em cluster são muitas vezes referidos como modelos de aprendizado não supervisionado, uma vez que eles não usam um campo de destino e não retornam uma predição específica que possa ser avaliada como verdadeira ou falsa. O valor de um modelo de armazenamento em cluster é determinado por sua capacidade de capturar agrupamentos de interesse nos dados e fornecer descrições úteis desses agrupamentos. Consulte Capítulo 11, “Modelos de Armazenamento em Cluster”, na página 223 para obter mais informações.

Requisitos. Um ou mais campos que definem as características de interesse. Os modelos de cluster não utilizam campos de destino da mesma maneira que outros modelos porque eles não fazem predições específicas que possam ser avaliadas como verdadeiras ou falsas. Ao invés disso, eles são utilizados para identificar grupos de casos que possam estar relacionados. Por exemplo, não é possível utilizar um

modelo de cluster para prever se um determinado cliente irá migrar para o concorrente ou responder a uma oferta. No entanto, é possível utilizar um modelo de cluster para designar clientes aos grupos com base na tendência que eles têm para fazer essas coisas. Os campos de ponderação e de frequência não são utilizados.

Campos de avaliação. Embora nenhum destino seja utilizado, é possível, opcionalmente, especificar um ou mais campos de avaliação a serem utilizados na comparação de modelos. A utilidade de um modelo de cluster pode ser avaliada por medir quão bem (ou mal) os clusters diferenciam esses campos.

Tipos de Modelo Suportados

Os tipos de modelos suportados incluem TwoStep, K-Médias e Kohonen.

Opções de Modelo do Nó Cluster Automático

A guia Modelo do nó Cluster Automático permite especificar o número de modelos a serem salvos, com os critérios utilizados para comparar os modelos.

Nome do modelo É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Utilizar dados particionados. Se um campo de partição for definido, essa opção assegurará que apenas os dados da partição de treinamento sejam utilizados para construir o modelo.

Classificar modelos por. Especifica os critérios utilizados para comparar e classificar os modelos.

- **Silhueta.** Um índice que mede a coesão e a separação do cluster. Consulte *Medida de Ranqueamento por Silhueta* abaixo para obter mais informações.
- **Número de clusters.** O número de clusters no modelo.
- **Tamanho do menor cluster.** O menor tamanho de cluster.
- **Tamanho do maior cluster.** O maior tamanho de cluster.
- **Menor/maior cluster.** A razão do tamanho do menor cluster com o maior cluster.
- **Importância.** A importância do campo **Avaliação** na guia **Campos**. Observe que isso poderá ser calculado apenas se um campo **Avaliação** tiver sido especificado.

Classificar modelos usando. Se uma partição estiver em uso, será possível especificar se os ranqueamentos baseiam-se no conjunto de dados de treinamento ou no conjunto de testes. Com grandes conjuntos de dados, utilizar uma partição para triagem preliminar de modelos poderá melhorar enormemente o desempenho.

Número de modelos a serem mantidos. Especifica o número máximo de modelos a serem listados no nugget produzido pelo nó. Os modelos com maior classificação são listados de acordo com o critério de classificação especificado. Observe que aumentar esse limite poderá diminuir o desempenho. O valor máximo permitido é 100.

Medida de Ranqueamento por Silhueta

A medida de ranqueamento padrão, Silhueta, possui um valor padrão de 0 porque um valor menor que 0 (ou seja, negativo) indica que a distância média entre um caso e os pontos em seu cluster designado é maior que a distância mínima média dos pontos em outro cluster. Portanto, os modelos com uma Silhueta negativa poderão seguramente ser descartados.

A medida de ranqueamento é, na realidade, um coeficiente da Silhueta modificado que combina os conceitos de coesão de cluster (preferencialmente modelos que contêm clusters altamente coesos) e de

separação de cluster (preferencialmente modelos que contêm clusters altamente separados). O coeficiente de Silhueta médio é simplesmente a média de todos os casos do cálculo a seguir para cada caso individual:

$(B - A) / \max(A, B)$

em que A é a distância do caso até o centroide do cluster ao qual o caso pertence e B é a distância mínima do caso até o centroide de cada um dos demais clusters.

O coeficiente de Silhueta (e sua média) varia entre -1 (indica um modelo muito pobre) e 1 (indicando um modelo excelente). A média pode ser realizada no nível do total de casos (que produz a Silhueta total) ou no nível de clusters (que produz Silhueta de cluster). As distâncias podem ser calculadas utilizando distâncias euclidianas.

Opções Avançadas do Nó Cluster Automático

A guia Especialista do nó Cluster Automático permite aplicar uma partição (se disponível), selecionar os algoritmos a serem utilizados e especificar as regras de parada.

Modelos usados. Use as caixas de seleção na coluna à esquerda para selecionar os tipos de modelo (algoritmo) para incluir na comparação. Quanto mais tipos você selecionar, mais modelos serão criados e mais demorado o processamento será.

Tipo de modelo. Lista os algoritmos disponíveis (consulte abaixo).

Parâmetro de modelo. Para cada tipo de modelo, é possível utilizar as configurações padrão ou selecionar **Especificar** para escolher opções para cada tipo de modelo. As opções específicas são semelhantes àquelas disponíveis nos nós de modelagem separados, com a diferença de que diversas opções ou combinações podem ser selecionadas. Por exemplo, se comparar os modelos de Rede Neural, ao invés de escolher um dos seis métodos de treinamento, é possível escolher todos eles para treinar seis modelos em um único passo.

Número de modelos. Lista o número de modelos produzidos para cada algoritmo com base nas configurações atuais. Ao combinar opções, o número de modelos pode aumentar rapidamente, portanto, é altamente recomendado prestar atenção nesse número, principalmente quando usar grandes conjuntos de dados.

Tempo máximo restrito gasto construindo um único modelo. (apenas modelos K-Médias, Kohonen, TwoStep, SVM, KNN, Bayes Net e Lista de Decisão) Configura um limite de tempo máximo para qualquer modelo. Por exemplo, se um modelo específico requerer um longo tempo inesperado para treinar devido a alguma interação complexa, você provavelmente não vai querer que isso atrase toda a execução de modelagem.

Algoritmos Suportados

O nó K-Médias armazena em cluster o conjunto de dados em grupos distintos (ou clusters). O método define um número fixo de clusters, designa iterativamente registros para clusters e ajusta os centros do cluster até que o refinamento adicional não consiga mais melhorar o modelo. Ao invés de tentar prever um resultado, o *k-médias* utiliza um processo conhecido como aprendizado não supervisionado para descobrir padrões no conjunto de campos de entrada.

O nó Kohonen gera um tipo de rede neural que pode ser utilizado para armazenar em cluster o conjunto de dados em grupos distintos. Quando a rede for totalmente treinada, os registros similares deverão estar próximos no mapa de saída, ao passo que registros que forem diferentes estarão distantes. É possível examinar o número de observações capturadas por cada unidade no nugget do modelo para identificar as unidades fortes. Isto poderá dar uma ideia do número apropriado de clusters.

O nó TwoStep utiliza um método de clusterização em dois passos. O primeiro passo faz uma única passagem através dos dados para compactar os dados de entrada brutos em um conjunto gerenciável de subclusters. O segundo passo usa um método de armazenamento em cluster hierárquico para mesclar progressivamente os subclusters em clusters cada vez maiores. O TwoStep tem a vantagem de estimar automaticamente o número ideal de clusters para os dados de treinamento. Ele pode manipular tipos de campo combinados e grandes conjuntos de dados de maneira eficiente.

Opções de Descarte do Nó Cluster Automático

A guia Descarte do nó Cluster Automático permite descartar automaticamente modelos que não atenderem a determinados critérios. Esses modelos não serão listados no nugget do modelo.

É possível especificar o valor mínimo de silhueta, números de cluster, tamanhos de cluster e a importância do campo de avaliação utilizados no modelo. A silhueta e o número e o tamanho dos clusters são determinados conforme especificado no nó de modelagem. Consulte o tópico “Opções de Modelo do Nó Cluster Automático” na página 75 para obter mais informações.

Opcionalmente, é possível configurar o nó para parar a execução na primeira vez em que um modelo que atender a todos os critérios especificados for gerado. Consulte o tópico “Regras de Parada do Nó de Modelagem Automatizada” na página 64 para obter mais informações.

Nuggets do Modelo Automatizado

Quando um nó de modelagem automatizado é executado, o nó estima os modelos candidatos para cada combinação possível de opções, classifica cada modelo candidato com base na medida que você especificar e salva os melhores modelos em um nugget do modelo automatizado composto. Este nugget do modelo contém na realidade um conjunto de um ou mais modelos gerados pelo nó, que podem ser procurados ou selecionados individualmente para uso na escoragem. O tipo de modelo e o tempo de construção são listados para cada modelo, com diversas outras medidas conforme apropriado para o tipo de modelo. É possível ordenar a tabela em qualquer uma dessas colunas para identificar rapidamente os modelos de maior interesse.

- Para procurar qualquer um dos nuggets do modelo individuais, clique duas vezes no ícone do nugget. Em seguida, é possível gerar um nó de modelagem desse modelo na tela de fluxo, ou uma cópia do nugget do modelo na paleta de modelos.
- O gráfico miniatura fornece uma avaliação visual rápida de cada tipo de modelo, conforme sumarizado a seguir. É possível clicar duas vezes em uma miniatura para gerar um gráfico em tamanho integral. O gráfico em tamanho integral mostra até 1000 pontos e se baseará em uma amostra se o conjunto de dados contiver mais. (Apenas para gráficos de dispersão, como o gráfico é gerado novamente toda vez que for exibido, quaisquer mudanças nos dados anteriores, como atualização de uma amostra aleatória ou partição se **Configurar Semente Aleatória** não estiver selecionada, poderão ser refletidas toda vez que um gráfico de dispersão for gerado novamente).
- Utilize a barra de ferramentas para mostrar ou ocultar colunas específicas na guia Modelo ou para alterar a coluna utilizada para ordenar a tabela. (Também é possível alterar a ordem clicando nos cabeçalhos da coluna).
- Utilize o botão Excluir para remover permanentemente quaisquer modelos não utilizados.
- Para reordenar colunas, clique em um cabeçalho da coluna e arraste a coluna para o local desejado.

- Se uma partição estiver em uso, será possível optar por visualizar os resultados para o treinamento ou teste de partição, conforme aplicável.

As colunas específicas dependem dos tipos de modelos que estão sendo comparados, conforme detalhado abaixo.

Respostas Binárias

- Para modelos binários, o gráfico miniatura mostra a distribuição de valores reais, sobreposta com os valores preditos, para fornecer uma indicação visual rápida de quantos registros foram preditos corretamente em cada categoria.
- Os critérios de ranqueamento correspondem às opções no nó de modelagem Classificador Automático. Consulte o tópico "Opções de Modelo do Nó Classificador Automático" na página 66 para obter mais informações.
- Para obter o máximo de lucro, o percentil no qual o máximo ocorre também é relatado.
- Para elevação acumulativa, é possível alterar o percentil selecionado utilizando a barra de ferramentas.

Respostas Nominais

- Para modelos nominais (conjunto), o gráfico miniatura mostra a distribuição de valores reais, sobreposta com os valores preditos, para fornecer uma indicação visual rápida de quantos registros foram preditos corretamente em cada categoria.
- Os critérios de ranqueamento correspondem às opções no nó de modelagem Classificador Automático. Consulte o tópico "Opções de Modelo do Nó Classificador Automático" na página 66 para obter mais informações.

Variáveis Resposta Contínua

- Para os modelos contínuos (intervalo numérico), o gráfico representa valores preditos versus valores observados para cada modelo, fornecendo uma indicação visual rápida da correlação entre eles. Para obter um bom modelo, os pontos deverão estar agrupados ao longo da diagonal, ao invés de estarem dispersos aleatoriamente no gráfico.
- Os critérios de ranqueamento correspondem às opções no nó de modelagem Numeração Automática. Consulte o tópico "Opções de Modelo do Nó Numeração Automática" na página 71 para obter mais informações.

Respostas do Cluster

- Para modelos de cluster, o gráfico representa contagens com relação a clusters de cada modelo, fornecendo uma indicação visual rápida de distribuição de cluster.
- Os critérios de ranqueamento correspondem às opções no nó de modelagem Cluster Automático. Consulte o tópico "Opções de Modelo do Nó Cluster Automático" na página 75 para obter mais informações.

Selecionando Modelos para Escoragem

A coluna **Utilizar?** permite selecionar os modelos a serem usados na escoragem.

- Para respostas binárias, nominais e numéricas, é possível selecionar diversos modelos de escoragem e combinar os escores no único nugget do modelo combinado. Ao combinar predições a partir de diversos modelos, as limitações em modelos individuais podem ser evitadas, resultando normalmente em uma precisão geral maior do que pode ser obtida a partir de qualquer um dos modelos.
- Para modelos de cluster, apenas um modelo de escoragem pode ser selecionado por vez. Por padrão, o modelo que tiver a melhor classificação será selecionado primeiro.

Gerando Nós e Modelos

É possível gerar uma cópia do nugget do modelo automatizado composto ou do nó de modelagem automatizado a partir do qual ele foi construído. Por exemplo, isso poderá ser útil se você não tiver o fluxo original a partir do qual o nugget do modelo automatizado foi construído. Como alternativa, é possível gerar um nugget ou nó de modelagem para qualquer um dos modelos individuais listados no nugget do modelo automatizado.

Nugget de Modelagem Automatizado

No menu Gerar, selecione **Modelo para Paleta** para incluir o nugget do modelo automatizado na paleta Modelos. O modelo gerado pode ser salvo ou utilizado no estado em que se encontra sem executar novamente o fluxo.

Como alternativa, é possível selecionar **Gerar Nó de Modelagem** a partir do menu Gerar para incluir o nó de modelagem na tela de fluxo. Este nó pode ser usado para reestimar os modelos selecionados sem repetir a execução da modelagem inteira.

Nugget de Modelagem Individual

1. No menu **Modelo**, clique duas vezes no nugget individual necessário. Uma cópia desse nugget é aberta em um novo diálogo.
2. No menu Gerar do novo diálogo, selecione **Modelo para Paleta** para incluir o nugget de modelagem individual na paleta Modelos.
3. Como alternativa, é possível selecionar **Gerar Nó de Modelagem** a partir do menu Gerar no novo diálogo para incluir o nó de modelagem individual na tela de fluxo.

Gerando Gráficos de Avaliação

Apenas para modelos binários, é possível gerar gráficos de avaliação que oferecem uma maneira visual para avaliar e comparar o desempenho de cada modelo. Os gráficos de avaliação não estão disponíveis para modelos gerados pelos nós Numeração Automática e Cluster Automático.

1. Na coluna *Usar?* no nugget do modelo automatizado do Classificador Automático, selecione os modelos que deseja avaliar.
2. No menu Gerar, escolha **Gráfico(s) de Avaliação**. A caixa de diálogo Gráfico de Avaliação é exibida.
3. Selecione o tipo de gráfico e outras opções conforme desejado.

Gráfico de Avaliação

Na guia Modelo do nugget do modelo automatizado, é possível realizar drill down para exibir gráficos individuais para cada um dos modelos mostrados. Para os nuggets Classificador Automático e Numeração Automática, a guia Gráfico exibe um gráfico e a importância do preditor que refletem os resultados de todos os modelos combinados. Consulte o tópico “Importância do preditor” na página 44 para obter mais informações.

Para o Classificador Automático, um gráfico de distribuição é mostrado, ao passo que um multigráfico (também conhecido como um gráfico de dispersão) é mostrado para a Numeração Automática.

Capítulo 6. Árvores de Decisão

Modelos de Árvore de Decisão

Utilize os modelos de árvore de decisão para desenvolver sistemas de classificação que prevejam ou classifiquem observações futuras com base em um conjunto de regras de decisão. Se você tiver dados divididos em classes de seu interesse (por exemplo, empréstimos de alto versus baixo risco, assinantes versus não assinantes, votantes versus não votantes, ou tipos de bactérias), seus dados poderão ser usados para construir regras que possam ser utilizadas para classificar casos novos ou antigos com máxima precisão. Por exemplo, é possível construir uma árvore que classifica o risco de crédito ou intenções de compra com base na idade e em outros fatores.

Esta abordagem, às vezes conhecida como *indução de regra*, tem várias vantagens. Primeiro, o processo de raciocínio por trás do modelo é evidente quando navegar pela árvore. Isso está em contraste com outras técnicas de modelagem *caixa preta* em que a lógica interna pode ser difícil de trabalhar.

Segundo, o processo inclui automaticamente em sua regra apenas os atributos que realmente forem importantes para uma tomada de decisão. Os atributos que não contribuírem com a precisão da árvore são ignorados. Isso pode produzir informações muito úteis sobre os dados e também ser utilizado para reduzir os dados para campos relevantes antes de treinar outra técnica de aprendizado, como uma rede neural.

Os nuggets do modelo de árvore de decisão podem ser convertidos em uma coleção de regras 'if-then' (um *conjunto de regras*) que, em muitos casos, mostram as informações em formato mais compreensível. A apresentação da árvore de decisão é útil quando desejar ver como os atributos nos dados podem *dividir*, ou *particionar*, a população em subconjuntos relevantes ao problema. A saída do nó Árvore do AS é diferente dos outros nós de Árvore de Decisão porque ela inclui uma lista de regras diretamente no nugget sem precisar criar um conjunto de regras. A apresentação do conjunto de regras será útil se você desejar ver como grupos específicos de itens se relacionam com uma conclusão específica. Por exemplo, a regra a seguir fornece um *perfil* de um grupo de carros que valem a pena comprar:

```
IF tested = 'yes'
AND mileage = 'low'
THEN -> 'BUY'.
```

Algoritmos de Construção de Árvore

Cinco algoritmos estão disponíveis para executar análise de classificação e segmentação. Todos esses algoritmos fazem basicamente a mesma coisa, eles examinam todos os campos de seu conjunto de dados para localizar o que fornece a melhor classificação ou predição ao dividir os dados em subgrupos. O processo é aplicado recursivamente, dividindo subgrupos em unidades cada vez menores até que a árvore seja concluída (conforme definido por determinados critérios de parada). Os campos de destino e de entrada utilizados na construção da árvore podem ser contínuos (intervalo numérico) ou categóricos, dependendo do algoritmo utilizado. Se uma variável resposta contínua for utilizada, uma árvore de regressão será gerada; se uma variável resposta categórica for utilizada, uma árvore de classificação será gerada.

O nó Árvore de Classificação e Regressão (C&R) gera uma árvore de decisão que permite prever ou classificar observações futuras. O método utiliza particionamento recursivo para dividir os registros de treinamento em segmentos ao minimizar a impureza em cada passo, em que um nó na árvore será considerado “puro” se 100% dos casos no nó caírem em uma categoria específica do campo de destino. Os campos de destino e de entrada podem ser intervalos numéricos ou categóricos (nominal, ordinal ou sinalizadores) e todas as divisões são binárias (somente dois subgrupos).

O nó CHAID gera árvores de decisão usando estatísticas qui-quadrado para identificar divisões ideais. Diferentemente dos nós Árvore C&R e QUEST, o CHAID pode gerar árvores não binárias, o que significa que algumas divisões possuem mais de duas ramificações. Os campos de destino e de entrada podem ser um intervalo numérico (contínuo) ou categóricos. Um Exhaustive CHAID é uma modificação de CHAID que executa uma tarefa mais completa de examinar todas as possíveis divisões, mas demora mais tempo para calcular.

O nó QUEST fornece um método de classificação binária para construir árvores de decisão, projetado para reduzir o tempo de processamento necessário para grandes análises de Árvore C&R enquanto também reduz a tendência localizada nos métodos de árvore de classificação para favorecer entradas que permitem mais divisões. Os campos de entrada podem ser intervalos numéricos (conínuo), ao passo que o campo de destino deve ser categórico. Todas as divisões são binárias.

O nó C5.0 constrói uma árvore de decisão ou um conjunto de regras. O modelo funciona dividindo a amostra com base no campo que fornece o ganho máximo de informações em cada nível. O campo de destino deve ser categórico. Diversas divisões em mais de dois subgrupos são permitidas.

O nó Árvore-AS estará disponível apenas se você tiver uma conexão com o IBM SPSS Analytic Server. Este nó é semelhante ao nó CHAID existente, no entanto, o nó Árvore-AS é projetado para processar Big Data para criar uma árvore única e exibe o modelo resultante no visualizador de saída que foi incluído no SPSS Modeler versão 17. O nó gera uma árvore de decisão usando estatísticas qui-quadrado (CHAID) para identificar divisões ideais. Essa utilização do CHAID pode gerar árvores não binárias, o que significa que algumas divisões possuem mais de duas ramificações. Os campos de destino e de entrada podem ser um intervalo numérico (contínuo) ou categóricos. Um Exhaustive CHAID é uma modificação de CHAID que executa uma tarefa mais completa de examinar todas as possíveis divisões, mas demora mais tempo para calcular.

Usos Gerais de Análises Baseadas em Árvore

A seguir estão alguns usos gerais de análise baseada em árvore:

Segmentação Identificar pessoas que podem ser membros de uma classe específica.

Estratificação Designar casos em uma das várias categorias, como grupos de alto, médio e baixo risco.

Predição Criar regras e utilizá-las para prever eventos futuros. A predição também pode significar tentativas de relacionar os atributos preditivos aos valores de uma variável contínua.

Redução de dados e triagem de variável Selecionar um subconjunto útil de preditores a partir de um grande conjunto de variáveis a serem utilizadas na construção de um modelo paramétrico formal.

Identificação de interação Identificar relacionamentos que pertencerem apenas a subgrupos específicos e especificá-los em um modelo paramétrico formal.

Mesclagem de categoria e unificação de variáveis contínuas Recodificar e agrupar categorias do preditor e variáveis contínuas com perda mínima de informações.

O Construtor de Árvore Interativo

É possível gerar um modelo de árvore automaticamente, no qual o algoritmo decide a melhor divisão em cada nível, ou é possível usar o construtor de árvore interativo para assumir o controle, aplicando seu conhecimento de negócios para refinar ou simplificar a árvore antes de salvar o nugget do modelo.

1. Crie um fluxo e inclua um dos nós de árvore de decisão Árvore C&R, CHAID ou QUEST.

   Nota: A construção de árvore interativa não é suportada para as árvores Árvore do AS ou C5.0.
2. Abra o nó e, na guia Campos, selecione os campos de destino e de preditor e especifique as opções de modelo adicionais, conforme necessário. Para obter instruções específicas, consulte a documentação para cada nó de construção de árvore.
3. No painel Objetivos da guia Opções de Criação, selecione **Ativar sessão interativa**.
4. Clique em **Executar** para ativar o construtor de árvore.

A árvore atual é exibida, começando com o nó raiz. É possível editar e podar a árvore nível por nível e acessar os ganhos, riscos e informações relacionadas antes de gerar um ou mais modelos.

Comentários

- Com os nós Árvore C&R, CHAID e QUEST, todos os campos ordinais utilizados no modelo devem ter armazenamento numérico (não sequência de caracteres). Se necessário, o nó Reclassificar pode ser utilizado para convertê-los.
- Opcionalmente, é possível utilizar um campo de partição para separar os dados em amostras de treinamento e de teste.
- Como uma alternativa ao uso do construtor de árvore, também é possível gerar um modelo diretamente a partir do nó de modelagem assim como outros modelos do IBM SPSS Modeler. Consulte o tópico “Compilando um Modelo de Árvore Diretamente” na página 94 para obter mais informações.

Crescendo e Podando a Árvore

A guia Visualizador no construtor de árvore permite visualizar a árvore atual, começando com o nó raiz.

1. Para crescer a árvore, nos menus escolha:

   Árvore > **Crescer Árvore**

   O sistema constrói a árvore, dividindo recursivamente cada ramificação até que um ou mais critérios de parada sejam atendidos. Em cada divisão, o melhor preditor é selecionado automaticamente com base no método de modelagem utilizado.
2. Como alternativa, selecione **Crescer Árvore em Um Nível** para incluir um único nível.
3. Para incluir uma ramificação abaixo de um nó específico, selecione o nó e selecione **Crescer Ramificação**.
4. Para escolher o preditor usado para uma divisão, selecione o nó desejado e selecione **Crescer Ramificação com Divisão Customizada**. Consulte o tópico “Definindo Divisões Customizadas” na página 84 para obter mais informações.
5. Para podar uma ramificação, selecione um nó e selecione **Remover Ramificação** para limpar o nó selecionado.
6. Para remover o nível inferior da árvore, selecione **Remover Um Nível**.
7. Apenas para as árvores Árvore C&R e QUEST, selecione **Crescer Árvore e Podar** para podar com base em um algoritmo de complexidade de custo que ajusta a estimativa de risco com base no número de nós terminais, geralmente resultando em uma árvore mais simples. Consulte o tópico “Nó Árvore C&R” na página 95 para obter mais informações.

Lendo Regras de Divisão na guia Visualizador

Ao visualizar regras de divisão na guia Visualizador, os colchetes indicam que o valor adjacente é incluído no intervalo, ao passo que os parênteses indicam que o valor adjacente é excluído do intervalo. A expressão (23,37], portanto, significa de 23 exclusivos para 37 inclusivos, ou seja, exatamente acima de 23 para 37. Na guia Modelo, a mesma condição seria exibida como:

```
Age > 23 and Age <= 37
```

Interrompendo o crescimento da árvore. Para interromper uma operação de crescimento da árvore (se estiver demorando mais tempo do que o esperado, por exemplo), clique no botão Parar Execução na barra de ferramentas.

Figura 28. Botão Parar Execução

O botão é ativado somente durante o crescimento da árvore. Ele para a operação de crescimento atual em seu ponto atual, deixando quaisquer nós que já tiverem sido incluídos, sem salvar as mudanças ou fechar a janela. O construtor de árvore permanece aberto, permitindo gerar um modelo, atualizar as diretivas ou exportar a saída no formato apropriado, conforme necessário.

Definindo Divisões Customizadas

A caixa de diálogo Definir Divisão permite selecionar o preditor e especificar condições para cada divisão.

1. No construtor de árvore, selecione um nó na guia Visualizador e nos menus escolha:
   Árvore > **Crescer Ramificação com Divisão Customizada**
2. Selecione o preditor desejado na lista suspensa ou clique no botão **Preditores** para visualizar detalhes de cada preditor. Consulte o tópico "Visualizando Detalhes do Preditor" para obter mais informações.
3. É possível aceitar as condições padrão para cada divisão ou selecionar **Customizado** para especificar condições para a divisão conforme apropriado.
   - Para preditores contínuos (intervalo numérico), é possível utilizar os campos **Editar valores do intervalo** para especificar o intervalo de valores que caem em cada novo nó.
   - Para preditores categóricos, é possível utilizar os campos **Editar valores do conjunto** ou **Editar valores ordinais** para determinar os valores específicos (ou intervalo de valores no caso de um preditor ordinal) que são mapeados para cada novo nó.
4. Selecione **Crescer** para crescer novamente a ramificação utilizando o preditor selecionado.

Em geral, a árvore pode ser dividida utilizando qualquer preditor, independentemente das regras de parada. As únicas exceções são quando o nó é puro (significando que 100% dos casos caem na mesma classe de destino, não restando nada para dividir) ou quando o preditor escolhido é constante (não há nada com relação ao qual dividir).

Valores omissos em. Apenas para árvores CHAID, se os valores omissos estiverem disponíveis para um determinado preditor, você terá a opção para designá-los a um nó filho específico quando definir uma divisão customizada. (Com a Árvore C&R e QUEST, os valores omissos são manipulados utilizando substitutos conforme definido no algoritmo. Consulte o tópico "Detalhes e Substitutos da Divisão" na página 85 para obter mais informações).)

Visualizando Detalhes do Preditor

A caixa de diálogo Selecionar Preditor exibe estatísticas sobre preditores disponíveis (ou "concorrentes" conforme eles são às vezes chamados) que podem ser utilizadas para a divisão atual.

- Para CHAID e Exhaustive CHAID, a estatística qui-quadrado é listada para cada preditor categórico; se um preditor for um intervalo numérico, a estatística de *F* será mostrada. A estatística qui-quadrado é uma medida do quão independente o campo de destino é do campo de divisão. Uma estatística qui-quadrado alta geralmente está relacionada a uma probabilidade inferior, o que significa que há

menos chance de que os dois campos sejam independentes – uma indicação de que essa é uma divisão ideal. Os graus de liberdade também são incluídos porque eles levam em consideração o fato de que é mais fácil para uma divisão de três vias ter uma estatística grande e uma probabilidade pequena do que para uma divisão de duas vias.

- Para Árvore C&R e QUEST, a melhoria para cada preditor é exibida. Quanto maior o aprimoramento, maior será a redução de impureza entre os nós pai e filho se esse preditor for utilizado. (Um nó puro é aquele em que todos os casos caem em uma única categoria de destino; quanto menor a impureza na árvore, melhor o modelo ajusta os dados). Em outras palavras, uma figura de melhoria alta geralmente indica uma divisão útil para esse tipo de árvore. A medida de impureza utilizada é especificada no nó de construção de árvore.

Detalhes e Substitutos da Divisão

É possível selecionar qualquer nó na guia Visualizador e selecionar o botão de informações de divisão no lado direito da barra de ferramentas para visualizar detalhes sobre a divisão para esse nó. A regra de divisão utilizada é exibida, com estatísticas relevantes. Para árvores categóricas da Árvore C&R, a melhoria e a associação estão exibidas. A associação é uma medida da correspondência entre um substituto e o campo de divisão primário, em que o "melhor" substituto geralmente é aquele que mais se aproxima do campo de divisão. Para Árvore C&R e QUEST, quaisquer substitutos utilizados no lugar do preditor primário também são listados.

Para editar a divisão para o nó selecionado, é possível clicar no ícone no lado esquerdo do painel de substitutos para abrir a caixa de diálogo Definir Divisão. (Como um atalho, é possível selecionar um substituto na lista antes de clicar no ícone para selecioná-lo como o campo de divisão primário).

Substitutos. Quando aplicável, quaisquer substitutos para o campo de divisão primário são mostrados para o nó selecionado. Os substitutos são campos alternativos utilizados se o valor do preditor primário estiver omisso para um determinado registro. O número máximo de substitutos permitidos para uma determinada divisão é especificado no nó de construção de árvore, mas o número real dependerá dos dados de treinamento. Em geral, quanto mais houver dados omissos, mais substitutos deverão ser utilizados. Para outros modelos de árvore de decisão, essa guia estará vazia.

Nota: Para serem incluídos no modelo, os substitutos devem ser identificados durante a fase de treinamento. Se a amostra de treinamento não possuir nenhum valor omisso, então nenhum substituto será identificado e quaisquer registros com valores omissos encontrados durante o teste ou escoragem serão incluídos automaticamente no nó filho com o maior número de registros. Se valores omissos forem esperados durante os testes ou escoragens, assegure-se de que os valores estejam omissos também na amostra de treinamento. Substitutos não estão disponíveis para as árvores de CHAID.

Embora os substitutos não sejam utilizados para árvores CHAID, ao definir uma divisão customizada, você tem a opção de designá-los para um nó-filho específico. Consulte o tópico “Definindo Divisões Customizadas” na página 84 para obter mais informações.

Customizando a Visualização em Árvore

A guia Visualizador no construtor de árvore exibe a árvore atual. Por padrão, todas as ramificações na árvore são expandidas, mas é possível expandir e reduzir as ramificações e customizar outras configurações, conforme necessário.

- Clique no sinal de menos (**-**) no canto inferior direito de um nó pai para ocultar todos os seus nós filhos. Clique no sinal de mais (**+**) no canto inferior direito de um nó pai para exibir seus nós filhos.
- Utilize o menu Visualizar ou a barra de ferramentas para alterar a orientação da árvore (de cima para baixo, da esquerda para a direita ou da direita para a esquerda).
- Clique no botão "Exibir rótulos de campo e de valor" na barra de ferramentas principal para mostrar ou ocultar rótulos de campo e de valor.
- Utilize os botões de lupa para aumentar ou diminuir o zoom da visualização ou clique no botão de mapa da árvore no lado direito da barra de ferramentas para visualizar um diagrama da árvore inteira.

- Se um campo de partição estiver em uso, será possível trocar a visualização em árvore entre as partições de treinamento e de teste (**Visualizar > Partição**). Quando a amostra de teste é exibida, a árvore poderá ser visualizada, mas não editada. (A partição atual é exibida na barra de status no canto inferior direito da janela).
- Clique no botão de informações de divisão (o botão "i" à extrema direita da barra de ferramentas) para visualizar os detalhes sobre a divisão atual. Consulte o tópico “Detalhes e Substitutos da Divisão” na página 85 para obter mais informações.
- Exibir estatísticas, gráficos, ou ambos, dentro de cada nó (veja abaixo).

Exibindo Estatísticas e Gráficos

Estatísticas do nó. Para um campo de destino categórico, a tabela em cada nó mostra o número e a porcentagem de registros em cada categoria e a porcentagem da amostra inteira que o nó representa. Para um campo de destino contínuo (intervalo numérico), a tabela mostra a média, o desvio padrão, o número de registros e o valor predito do campo de destino.

Gráficos do nó. Para um campo de destino categórico, é um gráfico de barras de porcentagens em cada categoria do campo de destino. Antes de cada linha da tabela há uma amostra de cor que corresponde à cor que representa cada uma das categorias do campo de destino nos gráficos do nó. Para um campo de destino contínuo (intervalo numérico), o gráfico mostra um histograma do campo de destino para os registros no nó.

Ganhos

A guia Ganhos exibe estatísticas de todos os nós terminais na árvore. Os ganhos fornecem uma medida de até que ponto a média ou a proporção em um determinado nó difere da média geral. Em geral, quanto maior essa diferença, mais útil a árvore será como uma ferramenta para tomada de decisões. Por exemplo, um índice ou um valor de "elevação" de 148% para um nó indica que os registros no nó têm quase uma e meia mais chances de cair na categoria de destino do que no conjunto de dados como um todo.

Para os nós Árvore C&R e QUEST nos quais um conjunto de prevenção ao super ajuste é especificado, dois conjuntos de estatísticas são exibidos:

- Conjunto de crescimento de árvore – a amostra de treinamento com o conjunto de prevenção ao super ajuste removido
- conjunto de prevenção ao superajuste

Para outras árvores interativas Árvore C&R e QUEST, e para todas as árvores interativas CHAID, apenas as estatísticas do conjunto de crescimento de árvore são exibidas.

A guia Ganhos permite:

- Exibir estatísticas nó por nó, acumulativas ou de quantis.
- Exibir ganhos ou lucros.
- Alternar a visualização entre as tabelas e gráficos.
- Selecionar a categoria de destino (apenas variáveis resposta categóricas).
- Classificar a tabela em ordem crescente ou decrescente com base na porcentagem do índice. Se as estatísticas para diversas partições forem exibidas, as classificações serão sempre aplicadas na amostra de treinamento ao invés de na amostra de teste.

Em geral, as seleções feitas na tabela de ganhos serão atualizadas na visualização em árvore e vice-versa. Por exemplo, se você selecionar uma linha na tabela, o nó correspondente será selecionado na árvore.

Ganhos de Classificação

Para árvores de classificação (aquelas com uma variável resposta categórica), a porcentagem de índice de ganho informa o quanto maior a proporção de uma determinada categoria de destino em cada nó difere da proporção geral.

Estatísticas de Nó por Nó

Nessa visualização, a tabela exibe uma linha para cada nó terminal. Por exemplo, se a resposta geral de sua campanha por mala direta foi de 10%, e 20% dos registros que caírem no nó X responderam positivamente, a porcentagem de índice para o nó seria de 200%, indicando que a probabilidade de os respondentes desse grupo comprarem é duas vezes maior que a população geral.

Para os nós Árvore C&R e QUEST nos quais um conjunto de prevenção ao super ajuste é especificado, dois conjuntos de estatísticas são exibidos:

- Conjunto de crescimento de árvore – a amostra de treinamento com o conjunto de prevenção ao super ajuste removido
- conjunto de prevenção ao superajuste

Para outras árvores interativas Árvore C&R e QUEST, e para todas as árvores interativas CHAID, apenas as estatísticas do conjunto de crescimento de árvore são exibidas.

Nós. O ID do nó atual (conforme exibido na guia Visualizador).

Nó: n. O número total de registros nesse nó.

Nó (%). A porcentagem de todos os registros no conjunto de dados que caem nesse nó.

Ganho: n. O número de registros com a categoria de destino selecionada que caem nesse nó. Em outras palavras, de todos os registros no conjunto de dados que caem na categoria de destino, quantos estão nesse nó?

Ganho (%). A porcentagem de todos os registros na categoria de destino, em todo o conjunto de dados, que caem nesse nó.

Resposta (%). A porcentagem de registros no nó atual que caem na categoria de destino. As respostas neste contexto às vezes são referidas como "ocorrências".

Índice (%). A porcentagem de resposta para o nó atual expressa como uma porcentagem do percentual de resposta do conjunto de dados inteiro. Por exemplo, um valor de índice de 300% indica que os registros nesse nó têm três vezes mais chances de cair na categoria de destino do que no conjunto de dados como um todo.

Estatísticas Acumulativas

Na visualização acumulativa, a tabela exibe um nó por linha, e as estatísticas são acumulativas, classificadas em ordem crescente ou decrescente por porcentagem de índice. Por exemplo, se uma ordem decrescente for aplicada, o nó com a porcentagem de índice mais alta é listado primeiro, e as estatísticas nas linhas que se seguem são acumulativas dessa linha para cima.

A porcentagem de índice acumulativo diminui linha por linha conforme nós com porcentagens de resposta cada vez mais inferiores forem incluídos. O índice acumulativo para a linha final é sempre 100% visto que, neste ponto, o conjunto de dados inteiro é incluído.

Quantis

Nesta visualização, cada linha na tabela representa um quantil ao invés de um nó. Os quantis são quartis, quintis (quintos), decis (décimos), vingtiles (vigésimos) ou percentis (centésimos). Diversos nós poderão ser listados em um único quantil se mais de um nó for necessário para atingir essa porcentagem (por exemplo, se quartis forem exibidos, mas os dois principais nós contiverem menos de 50% de todos os casos). O restante da tabela é acumulativo e pode ser interpretado da mesma maneira que a visualização acumulativa.

Lucros e o ROI de Classificação

Para árvores de classificação, as estatísticas de ganhos também podem ser exibidas em termos de lucro e de ROI (retorno sobre o investimento). A caixa de diálogo Definir Lucros permite especificar a receita e as despesas para cada categoria.

1. Na guia Ganhos, clique no botão Lucro (rotulado como $/$) na barra de ferramentas para acessar a caixa de diálogo.
2. Insira os valores de receita e de despesa para cada categoria do campo de destino.

Por exemplo, se custar $0,48 para enviar uma oferta por correio para cada cliente e a receita de uma resposta positiva for de $9,95 para uma assinatura de três meses, então cada resposta *não* custará $0,48 e cada resposta *sim* terá um ganho de $9,47 (calculado como 9,95-0,48).

Na tabela de ganhos, o **lucro** é calculado como a soma das receitas menos as despesas para cada um dos registros em um nó terminal. O **ROI** é o lucro total dividido pelo total de despesas em um nó.

Comentários

- Os valores de lucro afetam apenas os valores médios de lucro e de ROI exibidos na tabela de ganhos, como uma maneira de visualizar as estatísticas em termos mais aplicáveis aos seus resultados. Eles não afetam a estrutura básica do modelo de árvore. Os lucros não devem ser confundidos com os custos de classificação errada, que são especificados no nó de construção de árvore e fatorados no modelo como uma forma de proteção contra erros caros.
- As especificações de lucro não são persistidas entre uma sessão interativa de construção de árvore e a próxima.

Ganhos de Regressão

Para árvores de regressão, é possível escolher entre visualizações nó por nó, nó por nó acumulativo e quantil. Os valores médios são mostrados na tabela. Os gráficos estão disponíveis apenas para quantis.

Gráficos de Ganhos

Os gráficos podem ser exibidos na guia Ganhos como uma alternativa para as tabelas.

1. Na guia Ganhos, selecione o ícone Quantis (terceiro a partir da esquerda na barra de ferramentas). (Os gráficos não estão disponíveis para estatísticas nó por nó ou acumulativas).
2. Selecione o ícone Gráficos.
3. Selecione as unidades exibidas (percentis, decis, e assim por diante) na lista suspensa conforme desejado.
4. Selecione **Ganhos**, **Resposta** ou **Elevação** para alterar a medida exibida.

Gráfico de Ganhos

O gráfico de ganhos representa os valores na coluna *Ganhos (%)* da tabela. Os ganhos são definidos como a proporção de ocorrências em cada incremento com relação ao número total de ocorrências na árvore, utilizando a seguinte equação:

(ocorrência em incremento / número total de ocorrências) x 100%

O gráfico ilustra efetivamente o quanto você ainda precisa efetuar cast da rede para capturar uma determinada percentagem de todas as ocorrências na árvore. A linha diagonal representa a resposta

esperada para a amostra inteira, se o modelo não foi utilizado. Neste caso, a taxa de resposta seria constante, uma vez que a probabilidade de uma pessoa responder é a mesma que a de outra pessoa. Para dobrar seu lucro, você precisaria perguntar para o dobro de pessoas. A curva da linha indica o quanto é possível melhorar sua resposta, incluindo apenas aquelas que se classificam nos percentuais mais altos com base no ganho. Por exemplo, incluir os 50% principais pode render mais de 70% das respostas positivas. Quanto mais íngreme for a curva, maior será o ganho.

Gráfico de Elevação

O gráfico de elevação representa os valores na coluna *Índice (%)* da tabela. Este gráfico compara a porcentagem de registros em cada incremento que forem ocorrências com a porcentagem geral de ocorrências no conjunto de dados de treinamento, utilizando a seguinte equação:

(ocorrências em incremento / registros em incremento) / (número total de ocorrências / número total de registros)

Gráfico de Resposta

O gráfico de resposta representa os valores na coluna *Resposta (%)* da tabela. A resposta é uma porcentagem de registros no incremento que forem ocorrências, utilizando a seguinte equação:

(respostas em incremento / registros em incremento) x 100%

Seleção Baseada em Ganhos

A caixa de diálogo Seleção Baseada em Ganhos permite selecionar automaticamente os nós terminais com os melhores (ou piores) ganhos com base em uma regra ou um limite especificado. Em seguida, é possível gerar um nó Seleção com base na seleção.

1. Na guia Ganhos, selecione a visualização nó por nó ou acumulativa e selecione a categoria de destino na qual deseja basear a seleção. (As seleções baseiam-se na exibição da tabela atual e não estão disponíveis para quantis).
2. Na guia Ganhos, a partir dos menus escolha:

   Editar > **Selecionar Nós Terminais** > **Seleção Baseada em Ganhos**

   Selecionar apenas. É possível selecionar nós correspondentes *ou* nós não correspondentes – por exemplo, para selecionar *quase todos* os principais 100 registros.

   Corresponder por informações ganhos. Corresponde nós com base em estatísticas de ganho para a categoria de destino atual, incluindo:

   - Nós em que o ganho, resposta ou elevação (índice) corresponde a um limite especificado, por exemplo, resposta maior ou igual a 50%.
   - Os *n* principais nós com base no ganho para a categoria de destino.
   - Os principais nós até um número especificado de registros.
   - Os principais nós até uma porcentagem especificada de dados de treinamento.
3. Clique em **OK** para atualizar a seleção na guia Visualizador.
4. Para criar um novo nó Seleção com base na seleção atual na guia Visualizador, escolha **Selecionar Nó** no menu Gerar. Consulte o tópico “Gerando Nós Filtro e Seleção” na página 93 para obter mais informações.

Nota: como na realidade você está selecionando nós ao invés de registros ou porcentagens, uma correspondência perfeita com o critério de seleção nem sempre poderá ser atingida. O sistema seleciona nós completos *até* o nível especificado. Por exemplo, se você selecionar os 12 principais casos e você tiver 10 no primeiro nó e dois no segundo nó, apenas o primeiro nó será selecionado.

Riscos

Os riscos informam as chances de classificação errada em qualquer nível. A guia Riscos exibe uma estimativa de risco pontual e (para saídas categóricas) uma tabela de classificação errada.

- Para predições numéricas, o risco é uma estimativa agrupada da variância em cada um dos nós terminais.
- Para predições categóricas, o risco é a proporção de casos incorretamente classificados, ajustado para quaisquer custos anteriores ou de classificação errada.

Salvando Modelos e Resultados da Árvore

É possível salvar ou exportar os resultados de suas sessões interativas de construção de árvore de várias maneiras, incluindo:

- Gerar um modelo com base na árvore atual (**Gerar > Gerar modelo**).
- Salvar as diretivas utilizadas para expandir a árvore atual. Na próxima vez que o nó de construção de árvore for executado, a árvore atual irá recrescer automaticamente, incluindo todas as divisões customizadas que tiverem sido definidas.
- Exportar informações de modelo, de ganhos e de risco. Consulte o tópico "Exportando Informações de Modelo, Ganho e Risco" na página 92 para obter mais informações.

A partir do construtor de árvore ou de um nugget do modelo de árvore, é possível:

- Gerar um filtro ou selecionar um nó com base na árvore atual. Consulte o tópico "Gerando Nós Filtro e Seleção" na página 93 para obter mais informações.
- Gerar um nugget do Conjunto de Regras que representa a estrutura em árvore como um conjunto de regras que definem as ramificações de terminal da árvore. Consulte o tópico "Gerando um Conjunto de Regras a partir de uma Árvore de Decisão" na página 93 para obter mais informações.
- Além disso, é possível exportar o modelo no formato PMML apenas para nuggets do modelo de árvore. Consulte o tópico "A Paleta de Modelos" na página 41 para obter mais informações. Se o modelo incluir quaisquer divisões customizadas, essas informações não serão preservadas no PMML exportado. (A divisão é preservada, mas o fato de que ela é customizada ao invés de escolhida pelo algoritmo não é).
- Gerar um gráfico com base em uma parte selecionada da árvore atual. *Nota*: isto funcionará apenas para um nugget quando ele estiver anexado a outros nós em um fluxo. Consulte o tópico "Gerando Gráficos" na página 118 para obter mais informações.

Nota: a árvore interativa em si não pode ser salva. Para evitar perda de seu trabalho, gere um modelo e/ou atualize as diretivas de árvore antes de fechar a janela do construtor de árvore.

Gerando um Modelo a partir do Construtor de Árvore

Para gerar um modelo com base na árvore atual, nos menus do construtor de árvore escolha:

Gerar > **Modelo**

Na caixa de diálogo Gerar Novo Modelo, é possível escolher dentre as opções a seguir:

Nome do modelo. É possível especificar um nome customizado ou gerar o nome automaticamente com base no nome do nó de modelagem.

Criar nó em. É possível incluir o nó na **Tela**, na **Paleta GM** ou em **Ambos**.

Incluir diretivas de árvore. Selecione essa caixa para incluir as diretivas da árvore atual no modelo gerado. Isso permite gerar novamente a árvore, se necessário. Consulte o tópico "Diretivas de Crescimento de Árvore" na página 91 para obter mais informações.

Diretivas de Crescimento de Árvore

Para os modelos Árvore C&R, CHAID e QUEST, as diretivas de árvore especificam condições para crescimento da árvore, um nível por vez. As diretivas são aplicadas sempre que o construtor de árvore interativo é ativado a partir do nó.

- As diretivas são mais seguramente utilizadas como uma forma de gerar novamente uma árvore criada durante uma sessão interativa anterior. Consulte o tópico "Atualizar Diretivas de Árvore" na página 92 para obter mais informações. Também é possível editar diretivas manualmente, mas isso deverá ser feito com cuidado.
- As diretivas são altamente específicas para a estrutura da árvore que elas descrevem. Assim, qualquer mudança nos dados subjacentes ou nas opções de modelagem poderá causar falha de um conjunto de diretrizes anteriormente válidas. Por exemplo, se o algoritmo CHAID alterar uma divisão de duas vias para uma divisão de três vias com base em dados atualizados, quaisquer diretivas baseadas na divisão de duas vias anterior falharão.

Nota: se você optar por gerar um modelo diretamente (sem utilizar o construtor de árvore), quaisquer diretivas de árvore serão ignoradas.

Editando Diretivas

1. Para visualizar ou editar diretivas salvas, abra o nó de construção de árvore e selecione o painel Objetivo da guia Opções de Criação.
2. Selecione **Iniciar sessão interativa** para ativar os controles, selecione **Usar diretivas de árvore** e clique em **Diretivas**.

Sintaxe da Diretiva

As diretivas especificam condições para o crescimento da árvore, iniciando com o nó raiz. Por exemplo, para crescer a árvore em um nível:

```
Grow Node Index 0 Children 1 2
```

Como nenhum preditor está especificado, o algoritmo escolhe a melhor divisão.

Observe que a primeira divisão deve sempre estar no nó raiz (`Index 0`) e os valores de índice para os dois filhos devem ser especificados (`1` e `2` neste caso). É inválido especificar `Grow Node Index 2 Children 3 4`, a menos que você primeiro tenha crescido a raiz que criou o Nó 2.

Para crescer a árvore:

```
Grow Tree
```

Para crescer e podar a árvore (apenas Árvore C&R):

```
Grow_And_Prune Tree
```

Para especificar uma divisão customizada para um preditor contínuo:

```
Grow Node Index 0 Children 1 2 Spliton
  ( "EDUCATE", Interval ( NegativeInfinity, 12.5)
     Interval ( 12.5, Infinity ))
```

Para dividir em um preditor nominal com dois valores:

```
Grow Node Index 2 Children 3 4 Spliton
  ( "GENDER", Group( "0.0" )Group( "1.0" ))
```

Para dividir em um preditor nominal com diversos valores:

```
Grow Node Index 6 Children 7 8 Spliton
   ( "ORGS", Group( "2.0","4.0" )
             Group( "0.0","1.0","3.0","6.0" ))
```

Para dividir em um preditor ordinal:

```
Grow Node Index 4 Children 5 6 Spliton
  ( "CHILDS", Interval ( NegativeInfinity, 1.0)
              Interval ( 1.0, Infinity ))
```

Nota: ao especificar divisões customizadas, os nomes e valores de campos (`EDUCATE`, `GENDER`, `CHILDS`, etc.) fazem distinção entre maiúsculas e minúsculas.

Diretivas para Árvores CHAID

As diretivas para árvores CHAID são particularmente sensíveis a mudanças nos dados ou no modelo porque -- ao contrário da Árvore C&R e QUEST -- elas não são restritas a utilizar divisões binárias. Por exemplo, a sintaxe a seguir é perfeitamente válida, mas falhará se o algoritmo dividir o nó raiz em mais de dois filhos:

```
Grow Node Index 0 Children 1 2
Grow Node Index 1 Children 3 4
```

Com o CHAID, é possível que o Nó 0 tenha 3 ou 4 filhos, o que causaria falha na segunda linha da sintaxe.

Utilizando Diretivas em Scripts

As diretivas também podem ser integradas em scripts utilizando aspas triplas.

Atualizar Diretivas de Árvore

Para preservar seu trabalho a partir de uma sessão interativa de construção de árvore, é possível salvar as diretivas utilizadas para gerar a árvore atual. Ao contrário de salvar um nugget do modelo, que não pode ser editado ainda mais, isso permite gerar novamente a árvore em seu estado atual para edição posterior.

Para atualizar as diretivas, nos menus do construtor de árvore escolha:

Arquivo > **Atualizar Diretivas**

As diretivas são salvas no nó de modelagem utilizado para criar a árvore (ou Árvore C&R, QUEST ou CHAID) e podem ser utilizadas para gerar novamente a árvore atual. Consulte o tópico “Diretivas de Crescimento de Árvore” na página 91 para obter mais informações.

Exportando Informações de Modelo, Ganho e Risco

No construtor de árvore, é possível exportar as estatísticas de modelo, de ganhos e de risco em formatos de texto, HTML ou de imagem, conforme apropriado.

1. Na janela do construtor de árvore, selecione a guia ou a visualização que você deseja exportar.
2. Nos menus, escolha:

   Arquivo > **Exportar**
3. Selecione **Texto**, **HTML** ou **Gráfico**, conforme apropriado, e selecione os itens específicos que você deseja exportar no submenu.

Onde aplicável, a exportação baseia-se nas seleções atuais.

Exportando em formatos de Texto ou HTML. É possível exportar estatísticas de ganho ou de risco para a partição de treinamento ou de teste (se definido). A exportação baseia-se nas seleções atuais na guia Ganhos, por exemplo, é possível escolher estatísticas de nó por nó, acumulativas ou quantis.

Exportando gráficos. É possível exportar a árvore atual, conforme exibido na guia Visualizador, ou exportar gráficos de ganhos para a partição de treinamento ou de teste (se definido). Os formatos disponíveis incluem *.JPEG*, *.PNG* e *.BMP*. Para ganhos, a exportação baseia-se nas seleções atuais na guia Ganhos (disponível apenas quando um gráfico é exibido).

Gerando Nós Filtro e Seleção

Na janela do construtor de árvore, ou quando procurar por um nugget do modelo de árvore de decisão, a partir dos menus, escolha:

Gerar > **Nó Filtro**

ou

> **Nó Seleção**

Nó Filtro. Gera um nó que filtra todos os campos não utilizados pela árvore atual. Essa é uma maneira rápida de reduzir o conjunto de dados para incluir apenas os campos que forem selecionados como importantes pelo algoritmo. Se houver um nó Tipo antes deste nó de árvore de decisão, todos os campos com o papel *Destino* serão transmitidos pelo nugget do modelo Filtro.

Nó Seleção. Gera um nó que seleciona todos os registros que caírem no nó atual. Essa opção requer que uma ou mais ramificações da árvore sejam selecionadas na guia Visualizador.

O nugget do modelo é colocado na tela de fluxo.

Gerando um Conjunto de Regras a partir de uma Árvore de Decisão

É possível gerar um nugget do modelo de Conjunto de Regras que representa a estrutura da árvore como um conjunto de regras que definem as ramificações terminais da árvore. Os conjuntos de regras geralmente podem reter a maioria das informações importantes de uma árvore de decisão integral, mas com um modelo menos complexo. A diferença mais importante é que, com um conjunto de regras, mais de uma regra poderá ser aplicada a qualquer registro específico ou nenhuma regra poderá se aplicar. Por exemplo, é possível ver todas as regras que preveem um resultado *não* seguidas por todas as regras que preveem *sim*. Se diversas regras se aplicarem, cada regra receberá um "voto" ponderado com base na confiança que estiver associada a essa regra, e a predição final é decidida combinando os votos ponderados de todas as regras que se aplicarem ao registro em questão. Se nenhuma regra se aplicar, uma predição padrão será designada ao registro.

Nota: Ao escorar um conjunto de regras, é possível observar diferenças na escoragem comparado com a escoragem com relação à árvore, visto que cada ramificação terminal em uma árvore é escorada de modo independente. Uma área onde essa diferença pode se tornar notável é quando houver valores omissos em seus dados.

Os conjuntos de regras podem ser gerados apenas a partir de árvores com campos de destino categóricos (não árvores de regressão).

Na janela do construtor de árvore, ou quando procurar por um nugget do modelo de árvore de decisão, a partir dos menus, escolha:

Gerar > **Conjunto de Regras**

Nome do conjunto de regras Especifique o nome do novo nugget do modelo do Conjunto de Regras.

Criar nó em Controla o local do novo nugget do modelo de Conjunto de Regras. Selecione **Tela**, **Paleta GM** ou **Ambos**.

Mínimo de instâncias Especifique o número mínimo de instâncias (número de registros aos quais a regra se aplica) para preservar no nugget do modelo do Conjunto de Regras. As regras com suporte menor que o valor especificado não são incluídas no novo conjunto de regras.

Mínimo de confiança Especifique a confiança mínima para regras a serem preservadas no nugget do modelo do Conjunto de Regras. As regras com confiança menor que o valor especificado não são incluídas no novo conjunto de regras.

Compilando um Modelo de Árvore Diretamente

Como uma alternativa ao uso do construtor de árvore interativa, é possível compilar um modelo de árvore de decisão diretamente do nó quando o fluxo é executado. Isso é consistente com a maioria dos outros nós de construção de modelo. Para a árvore de C5.0 e modelos de Árvore do AS que não são suportados pelo construtor de árvore interativa, este é o único método que pode ser utilizado.

1. Crie um fluxo e inclua um dos nós de árvore de decisão - Árvore C&R. CHAID, QUEST, C5.0 ou Árvore do AS.
2. Para Árvore C&R, QUEST ou CHAID, no painel do Objetivo da guia Opções de Criação, escolha um dos objetivos principais. Se você escolher **Compilar uma árvore única**, assegure-se de que **Modo** esteja configurado para **Gerar modelo**.

   Para o C5.0, na guia Modelo, configure o **Tipo de Saída** para **Árvore de Decisão**.

   Para Árvore do AS, no painel Configurações Básicas da guia Opções de Criação, selecione o tipo de **Algoritmo de crescimento de árvore**.
3. Selecione os campos de destino e preditor e especifique opções adicionais de modelo, conforme necessário. Para obter instruções específicas, consulte a documentação para cada nó de construção de árvore.
4. Execute o fluxo para gerar o modelo.

Comente sobre a compilação de árvore

- Ao gerar árvores utilizando este método, as diretivas de crescimento de árvore são ignoradas.
- Não importa se for interativo ou direto, ambos os métodos de criação de árvores de decisão geram definitivamente modelos semelhantes. Trata-se apenas de uma questão do nível de controle que você deseja ao longo do caminho.

Nós de Árvore de Decisão

Os nós de Árvore de Decisão no IBM SPSS Modeler fornecem acesso aos algoritmos de construção de árvore a seguir:

- Árvore C&R
- QUEST
- CHAID
- C5.0
- Árvore do AS

Consulte o tópico “Modelos de Árvore de Decisão” na página 81 para obter mais informações.

Os algoritmos são semelhantes pelo fato de que todos eles podem construir uma árvore de decisão recursivamente ao dividir os dados em subgrupos cada vez menores. Entretanto, há algumas diferenças importantes.

Campos de entrada. Os campos de entrada (preditores) podem ser de qualquer um dos tipos a seguir (níveis de medição): contínuo, categóricos, flag, nominal ou ordinal.

Campos de destino. Apenas um campo de destino pode ser especificado. Para Árvore C&R, CHAID e Árvore do AS, o destino pode ser contínuo, categórico, flag, nominal ou ordinal. Para QUEST, pode ser categórico, flag ou nominal. Para o C5.0, o destino pode ser flag, nominal ou ordinal.

Tipo de divisão. A Árvore C&R e QUEST suportam apenas divisões binárias (ou seja, cada nó da árvore pode ser dividido em no máximo duas ramificações). Por outro lado, o CHAID, o C5.0 e a Árvore do AS suportam divisão em mais de duas ramificações por vez.

Método usado para divisão. Os algoritmos diferem nos critérios utilizados para decidir as divisões. Quando a Árvore C&R prediz uma saída categórica, uma medida de dispersão é utilizada (por padrão, o coeficiente de Gini, embora isso possa ser alterado). Para variáveis resposta contínuas, o método de desvio de quadrado mínimo é utilizado. O CHAID e a Árvore do AS utilizam um teste qui-quadrado, ao passo que o QUEST utiliza um teste qui-quadrado para preditores categóricos e análise de variância para entradas contínuas. Para o C5.0, uma medida de teoria da informação é utilizada, que é a razão de ganho de informações.

Tratamento de valor omisso. Todos os algoritmos permitem valores omissos para os campos preditores, embora eles utilizem métodos diferentes para manipulá-los. A Árvore C&R e o QUEST utilizam campos de predição substitutos, quando necessário, para avançar um registro com valores omissos pela árvore durante o treinamento. O CHAID torna os valores omissos uma categoria separada e permite que eles sejam utilizados na construção da árvore. O C5.0 utiliza um método fracionário, que transmite uma parte fracionária de um registro para cada ramificação da árvore a partir de um nó no qual a divisão se baseia em um campo com um valor omisso.

Poda. A Árvore C&R, o QUEST e o C5.0 oferecem a opção para crescer a árvore totalmente e, em seguida, podá-la ao remover divisões de nível inferior que não contribuírem significativamente com a precisão da árvore. No entanto, todos os algoritmos de árvore de decisão permitem controlar o tamanho mínimo do subgrupo, o que ajuda a evitar ramificações com poucos registros de dados.

Construção de árvore interativa. A Árvore C&R, o QUEST e o CHAID fornecem uma opção para ativar uma sessão interativa. Isso permite construir sua árvore um nível por vez, editar as divisões e podar a árvore antes de criar o modelo. O C5.0 e a Árvore do AS não possuem opção interativa.

Probabilidades Anteriores. A Árvore C&R e o QUEST suportam a especificação de probabilidades anteriores para categorias quando prever um campo de destino categórico. As Probabilidades anteriores são estimativas da frequência relativa geral de cada categoria de destino na população a partir da qual os dados de treinamento são obtidos. Em outras palavras, elas são as estimativas de probabilidade que você faria para cada valor de destino possível antes de saber qualquer informação sobre os valores do preditor. O CHAID, C5.0 e a Árvore do AS não suportam especificação de probabilidades anteriores.

Conjuntos de regras. Não disponível para Árvore do AS. Para modelos com campos de destino categóricos, os nós da árvore de decisão fornecem a opção para criar o modelo no formato de um conjunto de regras, que às vezes pode ser mais fácil de interpretar do que uma árvore de decisão complexa. Para Árvores C&R, QUEST e CHAID, é possível gerar um conjunto de regras a partir de uma sessão interativa, e para o C5.0, é possível especificar essa opção no nó de modelagem. Além disso, todos os modelos de árvore de decisão permitem gerar um conjunto de regras a partir do nugget do modelo. Consulte o tópico “Gerando um Conjunto de Regras a partir de uma Árvore de Decisão” na página 93 para obter mais informações.

Nó Árvore C&R

O nó Árvore de Classificação e Regressão (C&R) é um método de classificação e de predição baseado em árvore.Semelhante ao C5.0, esse método utiliza particionamento recursivo para dividir os registros de treinamento em segmentos com valores de campo de saída semelhantes. O nó Árvore C&R é iniciado ao examinar os campos de entrada para localizar a melhor divisão, medida pela redução em um índice impureza resultante da divisão. A divisão define dois subgrupos, em que cada um deles é dividido

posteriormente em mais dois subgrupos, e assim por diante, até que um dos critérios de parada seja acionado. Todas as divisões são binárias (somente dois subgrupos).

Poda

As Árvores C&R fornecem a opção de primeiro crescer a árvore e, em seguida, podá-la com base em um algoritmo de complexidade de custo que ajusta a estimativa de risco com base no número de nós terminais. Este método, que permite que a árvore cresça para um tamanho grande antes da poda com base em critérios mais complexos, pode resultar em árvores menores com melhores propriedades de validação cruzada. O aumento do número de nós terminais geralmente reduz o risco para dados atuais (treinamento), mas o risco real poderá ser maior quando o modelo é generalizado para dados não vistos. Em um caso extremo, suponha que você tenha um nó terminal separado para cada registro no conjunto de treinamento. A estimativa de risco seria de 0% já que cada registro fica em seu próprio nó, mas o risco de classificação errada para dados não vistos (teste) seria quase certamente maior que 0. No entanto, a medida de complexidade de custo tenta compensar isso.

Exemplo. Uma empresa de TV a cabo encomendou um estudo de marketing para determinar quais clientes poderiam assinar um serviço de notícias interativas via cabo. Usando os dados do estudo, é possível criar um fluxo no qual o campo de destino é a intenção de comprar a assinatura e os campos preditores incluem idade, sexo, educação, categoria de renda, horas gastas assistindo televisão por dia e número de filhos. Ao aplicar um nó Árvore C&R no fluxo, você será capaz de prever e classificar as respostas para obter a maior taxa de resposta para sua campanha.

Requisitos. Para treinar um modelo de Árvore C&R, um ou mais campos de *Entrada* e exatamente um campo de *Destino* são necessários. Os campos de destino e de entrada podem ser contínuos (intervalo numérico) ou categóricos. Campos configurados para *Ambos* ou *Nenhum* são ignorados. Os campos usados no modelo devem ter seus tipos totalmente instanciados e quaisquer campos ordinais (conjunto ordenado) utilizados no modelo devem ter armazenamento numérico (não sequência de caracteres). Se necessário, o nó Reclassificar pode ser utilizado para convertê-los.

Intensidades. Os modelos da Árvore C&R são bastante robustos na presença de problemas como dados omissos e grandes números de campos. Eles geralmente não requerem longos tempos de treinamento para estimar. Além disso, os modelos de Árvore C&R tendem a ser mais fáceis de entender do que alguns outros tipos de modelo, já que as regras derivadas do modelo possuem uma interpretação muito clara. Ao contrário do C5.0, a Árvore C&R pode acomodar campos de saída contínuos e também categóricos.

Nó CHAID

O CHAID, ou Chi-squared Automatic Interaction Detection, é um método de classificação para construir árvores de decisão usando estatísticas qui-quadrado para identificar divisões ideais.

O CHAID primeiro examina as tabulações cruzadas entre cada um dos campos de entrada e o resultado, e testa a significância usando um teste de independência qui-quadrado. Se mais de uma dessas relações forem estatisticamente significativas, o CHAID selecionará o campo de entrada que for mais significativo (menor valor de *p*). Se uma entrada tiver mais de duas categorias, elas serão comparadas e as categorias que não mostrarem diferenças no resultado serão reduzidas juntas. Isso é feito ao unir sucessivamente o par de categorias que mostrarem a diferença menos significativa. Esse processo de mesclagem de categoria para quando todas as categorias restantes diferirem no nível de teste especificado. Para campos de entrada nominal, quaisquer categorias podem ser mescladas e, para um conjunto ordinal, apenas categorias contínuas podem ser mescladas.

O Exhaustive CHAID é uma modificação do CHAID que executa uma tarefa mais completa de examinar todas as divisões possíveis para cada preditor, mas leva mais tempo para calcular.

Requisitos. Os campos de destino e de entrada podem ser contínuos ou categóricos e os nós podem ser divididos em dois ou mais subgrupos em cada nível. Todos os campos ordinais utilizados no modelo devem ter armazenamento numérico (não sequência de caracteres). Se necessário, o nó Reclassificar pode ser utilizado para convertê-los.

Intensidades. Diferentemente dos nós Árvore C&R e QUEST, o CHAID pode gerar árvores não binárias, o que significa que algumas divisões possuem mais de duas ramificações. Isso, portanto, tende a criar uma árvore mais larga do que os métodos de crescimento binários. O CHAID funciona para todos os tipos de entradas e aceita ambas ponderações de caso e variáveis de frequência.

Nó QUEST

O QUEST — ou Quick, Unbiased, Efficient Statistical Tree — é um método de classificação binário para construir árvores de decisão. Um dos principais motivos para seu desenvolvimento é reduzir o tempo de processamento necessário para grandes análises da Árvore C&R com muitas variáveis ou com muitos casos. Um segundo objetivo do QUEST é reduzir a tendência localizada nos métodos da árvore de classificação para favorecer entradas que permitem mais divisões, ou seja, campos de entrada contínuos (intervalo numérico) ou aqueles com muitas categorias.

- O QUEST utiliza uma sequência de regras, com base em testes de significância, para avaliar os campos de entrada em um nó. Para fins de seleção, um mínimo de teste poderá precisar ser executado em cada entrada em um nó. Ao contrário da Árvore C&R, todas as divisões não são examinadas e, ao contrário da Árvore C&R e do CHAID, as combinações de categoria não são testadas quando avaliar um campo de entrada para seleção. Isso acelera a análise.
- As divisões são determinadas ao executar uma análise discriminante quadrática usando a entrada selecionada nos grupos formados pelas categorias de destino. Esse método mais uma vez melhora a velocidade sobre uma procura exaustiva (Árvore C&R) para determinar a divisão ideal.

Requisitos. Os campos de entrada podem ser contínuos (intervalos numéricos), mas o campo de destino deve ser categórico. Todas as divisões são binárias. Os campos de ponderação não podem ser utilizados. Quaisquer campos ordinais (conjunto ordenado) utilizados no modelo devem ter armazenamento numérico (não sequência de caracteres). Se necessário, o nó Reclassificar pode ser utilizado para convertê-los.

Intensidades. Assim como o CHAID, mas ao contrário da Árvore C&R, o QUEST utiliza testes estatísticos para decidir se um campo de entrada é utilizado ou não. Ele também separa os problemas da seleção de entrada e da divisão, aplicando critérios diferentes a cada um deles. Isto contrasta com o CHAID, no qual o resultado do teste estatístico que determina que a seleção de variável também produz a divisão. Da mesma forma, a Árvore C&R usa a medida de mudança de impureza para selecionar o campo de entrada e determinar a divisão.

Opções de Campos do Nó Árvore de Decisão

Na guia Campos, você escolhe se deseja utilizar as configurações de papel do campo já definidas em nós de envio de dados ou fazer as designações de campo manualmente.

Usar papéis predefinidos Esta opção utiliza as configurações de papel (destinos, preditores e assim por diante) a partir de um nó Tipo de envio de dados (ou na guia Tipos de um nó de origem de envio de dados).

Usar designações de campo customizadas Para designar manualmente os destinos, preditores e outros papéis, selecione esta opção.

Campos Use os botões de seta para designar itens manualmente dessa lista para os vários campos de papel à direita da tela. Os ícones indicam os níveis de medição válidos para cada campo de papel.

Para selecionar todos os campos na lista, clique no botão **Tudo**, ou clique em um botão de nível de medição individual para selecionar todos os campos com esse nível de medição.

Destino Selecione um campo como o destino para a predição.

Preditores (Entradas). Escolha um ou mais campos como entradas para a predição.

Ponderação de Análise. (somente árvores CHAID, C&RT e Árvore do AS) Para usar um campo como uma ponderação de caso, especifique esse campo aqui. As ponderações de caso são utilizadas para considerar as diferenças na variância nos níveis do campo de saída. Consulte o tópico "Usando Campos de Frequência e de Ponderação" na página 33 para obter mais informações.

Opções de Construção do Nó Árvore de Decisão

A guia Opções de Criação é onde você configura todas as opções para construir o modelo. Obviamente, é possível clicar somente no botão **Executar** para construir um modelo com todas as opções padrão, no entanto, você normalmente deseja customizar a construção para seus próprios propósitos.

A guia contém várias áreas de janela diferentes na quais você configura as customizações que são específicas para seu modelo.

Nós de Árvore de Decisão - Objetivos

Para os nós Árvore C&R, QUEST e CHAID, na área de janela Objetivos da guia Opções de Criação, é possível escolher se deseja construir um novo modelo ou atualizar um modelo existente. Também é possível configurar o objetivo principal do nó: construir um modelo com precisão ou estabilidade melhorada ou construir um para uso com conjuntos de dados muito grandes.

O que você deseja fazer?

Construir novo modelo. (Padrão) Cria um modelo totalmente novo toda vez que executar um fluxo contendo esse nó de modelagem.

Continuar treinando o modelo existente. Por padrão, um modelo completamente novo é criado sempre que um nó de modelagem é executado. Se essa opção for selecionada, o treinamento continua com o último modelo produzido com sucesso pelo nó. Isso permite atualizar ou renovar um modelo existente sem precisar acessar os dados originais e poderá resultar em um desempenho significativamente mais rápido desde que *apenas* os registros novos ou atualizados sejam alimentados no fluxo. Detalhes do modelo anterior são armazenados com o nó de modelagem, o que permite utilizar essa opção mesmo se o nugget do modelo anterior não estiver mais disponível na paleta de fluxo ou de Modelos.

Nota: Essa opção será ativada apenas se você selecionar **Construir uma árvore única** (para Árvore C&R, CHAID e QUEST), **Criar um modelo padrão** (para Rede Neural e Linear) ou **Criar um modelo para conjuntos de dados muito grandes** como o objetivo.

Qual é o seu objetivo principal?

- **Construir uma árvore única.** Cria um único modelo de árvore de decisão padrão. Os modelos padrão são geralmente mais fáceis de interpretar e podem ser mais rápidos de escorar do que modelos construídos utilizando outras opções de objetivo.

  Nota: Para modelos de divisão, para utilizar esta opção com **Continuar treinando modelo existente**, você deverá estar conectado ao Analytic Server.

  Modo. Especifica o método utilizado para construir o modelo. **Gerar modelo** cria um modelo automaticamente quando o fluxo é executado. **Ativar sessão interativa** abre o construtor de árvore, que permite construir uma árvore um nível por vez, editar divisões e podar conforme desejado antes de criar o nugget do modelo.

Usar diretivas de árvore. Selecione esta opção para especificar as diretivas a serem aplicadas ao gerar uma árvore interativa do nó. Por exemplo, é possível especificar divisões de primeiro e segundo nível, que seriam aplicadas automaticamente quando o construtor de árvore fosse ativado. Também é possível salvar diretivas a partir de uma sessão de construção de árvore interativa para recriar a árvore em uma data futura. Consulte o tópico "Atualizar Diretivas de Árvore" na página 92 para obter mais informações.

- **Aprimorar a precisão do modelo (boosting).** Escolha esta opção se desejar utilizar um método especial, conhecido como **boosting**, para melhorar a taxa de precisão do modelo. O boosting funciona ao construir diversos modelos em uma sequência. O primeiro modelo é construído da maneira usual. Em seguida, um segundo modelo é construído para focar nos registros que forem classificados incorretamente pelo primeiro modelo. Em seguida, um terceiro modelo é construído para focar nos erros do segundo modelo, e assim por diante. Por último, os casos são classificados ao aplicar o conjunto inteiro de modelos neles, utilizando um procedimento de votação ponderada para combinar as predições separadas em uma predição geral. O boosting pode melhorar significativamente a precisão de um modelo de árvore de decisão, como também requer um treinamento mais longo.
- **Aprimorar a estabilidade do modelo (bagging).** Escolha esta opção se desejar utilizar um método especial, conhecido como **bagging** (agregação de bootstrap), para melhorar a estabilidade do modelo e evitar super ajuste. Esta opção cria diversos modelos e os combina, a fim de obter predições mais confiáveis. Os modelos obtidos utilizando essa opção podem demorar mais tempo para construir e escorar do que os modelos padrão.
- **Criar um modelo para conjuntos de dados muito grandes.** Escolha esta opção quando estiver trabalhando com conjuntos de dados que forem muito grandes para construir um modelo utilizando qualquer uma das outras opções de objetivo. Esta opção divide os dados em blocos de dados menores e constrói um modelo em cada bloco. Em seguida, os modelos mais precisos são automaticamente selecionados e combinados em um nugget do modelo único. É possível executar atualização do modelo incremental se você selecionar a opção **Continuar treinando modelo existente** nessa tela.

  Nota: Esta opção para conjuntos de dados muito grandes requer uma conexão com o IBM SPSS Modeler Server.

Nós de Árvore de Decisão - Básicos

Especifique as opções básicas sobre como a árvore de decisão deve ser construída.

Algoritmo de crescimento de árvore (apenas CHAID e Árvore do AS) Escolha o tipo de algoritmo **CHAID** que você deseja utilizar. O **Exhaustive CHAID** é uma modificação do CHAID que executa uma tarefa mais completa de examinar todas as divisões possíveis para cada preditor, mas leva mais tempo para calcular.

Profundidade máxima da árvore Especifique o número máximo de níveis abaixo do nó raiz (o número de vezes em que a amostra será dividida recursivamente). O padrão é 5; escolha **Customizado** e insira um valor para especificar um número diferente de níveis.

Poda (apenas C&RT e QUEST)

Podar árvore para evitar super ajuste A poda consiste em remover divisões de nível inferior que não contribuírem significativamente para a precisão da árvore. A poda pode ajudar a simplificar a árvore, tornando-a mais fácil de interpretar e, em alguns casos, melhora a generalização. Se desejar uma árvore integral sem poda, deixe essa opção desmarcada.

- **Configurar diferença máxima de risco (nos Erros Padrão)** Permite para especificar uma regra de poda mais liberal. A regra de erro padrão permite que o algoritmo selecione a árvore mais simples cuja estimativa de risco esteja próxima (e possivelmente maior) que a da subárvore com o menor risco. O valor indica o tamanho da diferença permitida na estimativa de risco entre a árvore podada e a árvore com o menor risco em termos de estimativa de risco. Por exemplo, se você especificar 2, uma árvore cuja estimativa de risco seja (2 × erro padrão) maior que a da árvore integral pode ser selecionada.

Máximo de substitutos. Substitutos é um método para lidar com valores omissos. Para cada divisão na árvore, o algoritmo identifica os campos de entrada que forem mais semelhantes ao campo de divisão selecionado. Esses campos são os *substitutos* para essa divisão. Quando um registro tiver que ser classificado, mas tiver um valor omisso para um campo de divisão, seu valor em um campo substituto poderá ser utilizado para fazer a divisão. Aumentar essa configuração permitirá maior flexibilidade para lidar com valores omissos, mas também pode aumentar o uso de memória e os tempos de treinamento.

Nós de Árvore de Decisão - Regras de Parada

Estas opções controlam como a árvore é construída. As regras de parada determinam quando parar a divisão de ramificações específicas da árvore. Configure os tamanhos mínimos de ramificação para evitar que as divisões criem subgrupos muito pequenos. **Mínimo de registros na ramificação pai** evitará uma divisão se o número de registros no nó a serem divididos (o *pai*) for menor que o valor especificado. **Mínimo de registros na ramificação filha** evitará uma divisão se o número de registros em qualquer ramificação criada pela divisão (a *filha*) for menor que o valor especificado.

- **Usar porcentagem** Especifique tamanhos em termos de porcentagem de dados de treinamento gerais.
- **Usar valor absoluto** Especifique tamanhos conforme o número absoluto de registros.

Nós de Árvore de Decisão - Combinações

Essas configurações determinam o comportamento da combinação que ocorre quando efetuar boosting, bagging ou quando conjuntos de dados muito grandes forem solicitados nos Objetivos. As opções que não se aplicarem ao objetivo selecionado são ignoradas.

Bagging e Conjuntos de Dados Muito Grandes. Ao escorar uma combinação, essa é a regra utilizada para combinar os valores preditos a partir dos modelos base para calcular o valor de escore de combinação.

- **Regra de combinação padrão para variáveis resposta categórica**. Os valores preditos de combinação para variável resposta categórica podem ser combinados utilizando votação, probabilidade mais alta ou probabilidade média mais alta. **Votação** seleciona a categoria que tem a probabilidade mais alta e mais frequente nos modelos base. **Probabilidade mais alta** seleciona a categoria que atinge a única probabilidade mais alta em todos os modelos base. **Probabilidade média mais alta** Seleciona a categoria com o valor mais alto quando a média das probabilidades da categoria é calculada entre os modelos base.
- **Regra de combinação padrão para variáveis resposta contínuas.** Os valores preditos de combinação para variáveis resposta contínuas podem ser combinados utilizando a média ou mediana dos valores preditos a partir dos modelos base.

Observe que quando o objetivo é melhorar a precisão do modelo, as seleções da regra de combinação são ignoradas. O boosting sempre utiliza uma votação por maioria ponderada para escorar variáveis resposta categóricas e uma média ponderada para escorar variáveis resposta contínuas.

Boosting e Bagging. Especifique o número de modelos base para construção quando o objetivo for melhorar a precisão ou a estabilidade do modelo; para bagging, este é o número de amostras de bootstrap. Ele deve ser um número inteiro positivo.

Árvore C&R e Nós QUEST - Custos e Informações a priori

Custos de classificação errada

Em alguns contextos, determinados tipos de erros são mais caros que outros. Por exemplo, pode ser mais caro classificar um solicitante de crédito de alto risco como baixo risco (um tipo de erro) do que classificar um solicitante de baixo risco como alto risco (um tipo diferente de erro). Os custos de classificação errada permitem especificar a importância relativa de diferentes tipos de erros de predição.

Os custos de classificação errada são basicamente ponderações aplicadas a resultados específicos. Essas ponderações são fatoradas no modelo e podem, na realidade, alterar a predição (como uma forma de proteger contra erros caros).

Com exceção dos modelos do C5.0, os custos de classificação errada não serão aplicados ao escorar um modelo e não são levados em conta quando classificar ou comparar modelos usando um nó Classificador Automático, gráfico de avaliação, ou nó Análise. Um modelo que inclui custos poderá não produzir menos erros do que aquele que não inclui e poderá não ter uma classificação mais alta em termos de precisão geral, mas provavelmente executará melhor em termos práticos por possuir um viés integrado a favor de erros *menos caros*.

A matriz de custo mostra o custo para cada combinação possível de categoria predita e categoria real. Por padrão, todos os custos de classificação errada são configurados como 1,0. Para inserir valores de custo customizado, selecione **Usar custos de classificação errada** e insira os valores customizados na matriz de custo.

Para alterar um custo de classificação errada, selecione a célula correspondente à combinação desejada de valores preditos e reais, exclua o conteúdo existente da célula e insira o custo desejado para a célula. Os custos não são simétricos automaticamente. Por exemplo, se você configurar o custo de classificação errada de *A* como *B* para 2,0, o custo da classificação errada de *B* como *A* ainda terá o valor padrão de 1,0, a menos que você também o altere explicitamente.

Anteriores

Essas opções permitem especificar as probabilidades anteriores para categorias ao prever um campo de destino categórico. As **Probabilidades anteriores** são estimativas da frequência relativa geral de cada categoria de destino na população a partir da qual os dados de treinamento são obtidos. Em outras palavras, elas são as estimativas de probabilidade que você faria para cada valor de destino possível *antes* de saber qualquer informação sobre os valores do preditor. Existem três métodos de configuração de informações a priori:

- **Baseado em dados de treinamento.** Este é o padrão. As probabilidades anteriores baseiam-se nas frequências relativas das categorias nos dados de treinamento.
- **Igual para todas as classes.** As probabilidades anteriores para todas as categorias são definidas como $1/k$, em que k é o número de categorias de destino.
- **Customizado.** É possível especificar suas próprias probabilidades anteriores. Os valores iniciais das probabilidades anteriores são configurados como iguais para todas as classes. É possível ajustar as probabilidades de categorias individuais para valores definidos pelo usuário. Para ajustar a probabilidade de uma categoria específica, selecione a célula de probabilidade na tabela correspondente à categoria desejada, exclua o conteúdo da célula e insira o valor desejado.

As probabilidades anteriores de todas as categorias devem somar 1,0 (a **restrição de probabilidade**). Se elas não somarem 1,0, um aviso será exibido, com uma opção para normalizar automaticamente os valores. Esse ajustamento automático preserva as proporções entre as categorias ao aplicar a restrição de probabilidade. É possível executar este ajustamento a qualquer momento clicando no botão **Normalizar**. Para reconfigurar a tabela para valores iguais para todas as categorias, clique no botão **Igualar**.

Ajustar informações a priori usando custos de classificação errada. Esta opção permite ajustar as informações a priori com base nos custos de classificação errada (especificado na guia Custos). Isso permite incorporar informações de custo diretamente no processo de crescimento de árvore para árvores que usam a medida de impureza de Twoing. (Quando esta opção não estiver selecionada, as informações de custo são utilizadas apenas na classificação de registros e no cálculo das estimativas de risco para árvores com base na medida de Twoing).

Nó CHAID - Custos

Em alguns contextos, determinados tipos de erros são mais caros que outros. Por exemplo, pode ser mais caro classificar um solicitante de crédito de alto risco como baixo risco (um tipo de erro) do que classificar um solicitante de baixo risco como alto risco (um tipo diferente de erro). Os custos de classificação errada permitem especificar a importância relativa de diferentes tipos de erros de predição.

Os custos de classificação errada são basicamente ponderações aplicadas a resultados específicos. Essas ponderações são fatoradas no modelo e podem, na realidade, alterar a predição (como uma forma de proteger contra erros caros).

Com exceção dos modelos do C5.0, os custos de classificação errada não serão aplicados ao escorar um modelo e não são levados em conta quando classificar ou comparar modelos usando um nó Classificador Automático, gráfico de avaliação, ou nó Análise. Um modelo que inclui custos poderá não produzir menos erros do que aquele que não inclui e poderá não ter uma classificação mais alta em termos de precisão geral, mas provavelmente executará melhor em termos práticos por possuir um viés integrado a favor de erros *menos caros*.

A matriz de custo mostra o custo para cada combinação possível de categoria predita e categoria real. Por padrão, todos os custos de classificação errada são configurados como 1,0. Para inserir valores de custo customizado, selecione **Usar custos de classificação errada** e insira os valores customizados na matriz de custo.

Para alterar um custo de classificação errada, selecione a célula correspondente à combinação desejada de valores preditos e reais, exclua o conteúdo existente da célula e insira o custo desejado para a célula. Os custos não são simétricos automaticamente. Por exemplo, se você configurar o custo de classificação errada de *A* como *B* para 2,0, o custo da classificação errada de *B* como *A* ainda terá o valor padrão de 1,0, a menos que você também o altere explicitamente.

Nó Árvore C&R – Avançado

As opções avançadas permitem fazer um ajuste preciso do processo de construção da árvore.

Mudança mínima na impureza. Especifique mudança mínima na impureza para criar uma nova divisão na árvore. **Impureza** refere-se à extensão em que subgrupos definidos pela árvore possuem uma ampla variedade de valores de campo de saída dentro de cada grupo. Para variáveis resposta categóricas, um nó é considerado "puro" se 100% dos casos no nó caírem em uma categoria específica do campo de destino. O objetivo da construção de árvore é criar subgrupos com valores de saída semelhantes, em outras palavras, para minimizar a impureza dentro de cada nó. Se a melhor divisão de uma ramificação reduzir a impureza em um nível menor que a quantia especificada, a divisão não será feita.

Medida de impureza para destinos categóricos. Para campos de destino categóricos, especifique o método utilizado para medir a impureza da árvore. (Para destinos contínuos, essa opção é ignorada e a medida de impureza de **desvio de quadrado mínimo** é sempre utilizada).

- **Gini** é uma medida de impureza geral baseada em probabilidades da associação de categoria para a ramificação.
- **Twoing** é uma medida de impureza que enfatiza a divisão binária e que mais pode levar a ramificações de tamanhos aproximadamente iguais em uma divisão.
- **Ordenadas** inclui a restrição adicional de que apenas as classes de destino contínuas podem ser agrupadas, já que isso se aplica apenas com destinos ordinais. Se esta opção for selecionada para um destino nominal, a medida de twoing padrão será utilizada por padrão.

Conjunto de prevenção ao super ajuste. O algoritmo separa internamente os registros em um conjunto de construção de modelo e em um conjunto de prevenção ao super ajuste, que é um conjunto independente de registros de dados utilizados para rastrear erros durante o treinamento a fim de evitar que o método modele a variação de chances nos dados. Especifique uma porcentagem de registros. O padrão é 30.

Replicar resultados. Configurar uma semente aleatória permite replicar análises. Especifique um número inteiro ou clique em **Gerar**, que criará um pseudonúmero inteiro aleatório entre 1 e 2147483647, inclusive.

Nó QUEST - Avançado

As opções avançadas permitem fazer um ajuste preciso do processo de construção da árvore.

Nível de significância para divisão. Especifica o nível de significância (alpha) para divisão de nós. O valor deve estar entre 0 e 1. Valores mais baixos tendem a produzir árvores com menos nós.

Conjunto de prevenção ao super ajuste. O algoritmo separa internamente os registros em um conjunto de construção de modelo e em um conjunto de prevenção ao super ajuste, que é um conjunto independente de registros de dados utilizados para rastrear erros durante o treinamento a fim de evitar que o método modele a variação de chances nos dados. Especifique uma porcentagem de registros. O padrão é 30.

Replicar resultados. Configurar uma semente aleatória permite replicar análises. Especifique um número inteiro ou clique em **Gerar**, que criará um pseudonúmero inteiro aleatório entre 1 e 2147483647, inclusive.

Nó CHAID - Avançado

As opções avançadas permitem fazer um ajuste preciso do processo de construção da árvore.

Nível de significância para divisão. Especifica o nível de significância (alpha) para divisão de nós. O valor deve estar entre 0 e 1. Valores mais baixos tendem a produzir árvores com menos nós.

Nível de significância para mesclagem. Especifica o nível de significância (alfa) para mesclar as categorias. O valor deve ser maior que 0 e menor ou igual a 1. Para evitar qualquer mesclagem de categorias, especifique um valor de 1. Para variáveis resposta contínuas, isso significa que o número de categorias para a variável na árvore final corresponde ao número especificado de intervalos. Essa opção não está disponível para o Exhaustive CHAID.

Ajustar valores de significância usando o método de Bonferroni. Ajusta os valores de significância ao testar as várias combinações de categoria de um preditor. Os valores são ajustados com base no número de testes, que está diretamente relacionado ao número de categorias e ao nível de medição de um preditor. Isso é geralmente desejável porque controla melhor a taxa de erros de falso positivo. Desativar essa opção aumentará o poder da sua análise para localizar diferenças reais, mas ao custo de uma taxa maior de falso positivo. Em particular, desativar essa opção pode ser recomendado para pequenas amostras.

Permitir redivisão de categorias mescladas dentro de um nó. O algoritmo CHAID tenta mesclar categorias para produzir a árvore mais simples que descreve o modelo. Se selecionada, esta opção permite que categorias mescladas sejam divididas novamente, se isso resultar em uma solução melhor.

Qui-quadrado para variáveis resposta categóricas. Para variáveis resposta categóricas, é possível especificar o método utilizado para calcular estatísticas qui-quadrado.

- **Pearson.** Este método fornece cálculos mais rápidos, mas deve ser utilizado com cuidado em amostras pequenas.
- **Razão de verossimilhança.** Esse método é mais robusto que o Pearson, mas leva mais tempo para calcular. Para pequenas amostras, este é o método preferencial. Para variáveis resposta contínuas, este método é sempre utilizado.

Mudança mínima nas frequências de célula esperadas. Ao estimar as frequências de célula (para o modelo nominal e para o modelo ordinal de efeitos de linha), um procedimento iterativo (epsilon) é utilizado para convergir na estimativa ideal utilizada no teste qui-quadrado para uma divisão específica. O epsilon determina quanta mudança deve ocorrer para que as iterações continuem; se a mudança na

última iteração for menor que o valor especificado, a iteração parará. Se você tiver tendo problemas com o algoritmo que não converge, será possível aumentar esse valor ou aumentar o número máximo de iterações até que a convergência ocorra.

Máximo de iterações por convergência. Especifica o número máximo de iterações antes de parar, independentemente se a convergência tiver ocorrido ou não.

Conjunto de prevenção ao super ajuste. (Esta opção está disponível apenas ao utilizar o construtor de árvore interativo). O algoritmo separa internamente os registros em um conjunto de construção de modelo e em um conjunto de prevenção ao super ajuste, que é um conjunto independente de registros de dados utilizados para rastrear erros durante o treinamento a fim de evitar que o método modele a variação de chances nos dados. Especifique uma porcentagem de registros. O padrão é 30.

Replicar resultados. Configurar uma semente aleatória permite replicar análises. Especifique um número inteiro ou clique em **Gerar**, que criará um pseudonúmero inteiro aleatório entre 1 e 2147483647, inclusive.

Opções do Modelo do Nó Árvore de Decisão

Na guia Opções do modelo, é possível escolher se deseja especificar um nome para o modelo ou gerar um nome automaticamente. Também é possível optar por obter informações de importância do preditor, bem como escores de propensão bruta e ajustada para respostas de flag.

Nome do modelo É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Avaliação de modelo

Calcular a importância do preditor. Para modelos que produzem uma medida apropriada de importância, é possível exibir um gráfico que indica a importância relativa de cada preditor na estimativa do modelo. Geralmente, você deseja concentrar seus esforços de modelagem nos preditores que forem de maior importância e considerar eliminar ou ignorar aqueles que forem de menor importância. Observe que a importância do preditor poderá levar mais tempo para calcular para alguns modelos, principalmente quando estiver trabalhando com grandes conjuntos de dados, e é desativada por padrão para alguns modelos como resultado. A importância do preditor não está disponível para modelos de lista de decisão. Consulte o tópico “Importância do preditor” na página 44 para obter mais informações.

Escores de Propensão

Os escores de propensão podem ser ativados no nó de modelagem e na guia Configurações no nugget do modelo. Esta funcionalidade estará disponível apenas quando o destino selecionado for um campo de flag. Consulte o tópico “Escores de Propensão” na página 36 para obter mais informações.

Calcular escores de propensão bruta. Os escores de propensão bruta são derivados do modelo com base apenas nos dados de treinamento. Se o modelo prever o valor *true* (responderá), então a propensão será a mesma que P, em que P é a probabilidade da predição. Se o modelo prever o valor false, então e propensão é calculada como (1-P).

- Se você escolher essa opção ao construir o modelo, os escores de propensão serão ativados no nugget do modelo por padrão. No entanto, sempre é possível optar por ativar os escores de propensão bruta no nugget do modelo independentemente se você selecioná-los no nó de modelagem ou não.
- Ao escorar o modelo, os escores de propensão bruta serão incluídos em um campo com as letras *RP* anexadas ao prefixo padrão. Por exemplo, se as predições estiverem em um campo denominado *$R-churn*, o nome do campo de escore de propensão será *$RRP-churn*.

Calcular escores de propensão ajustada. As propensões brutas baseiam-se puramente nas estimativas fornecidas pelo modelo, que podem ser super ajustadas e gerar estimativas de propensão super otimistas. As propensões ajustadas tentam compensar isso ao examinar como o modelo é executado nas partições de teste ou de validação e ajustar as propensões para fornecer uma estimativa melhor de acordo.

- Essa configuração requer que um campo de partição válido esteja presente no fluxo.
- Diferentemente dos escores de confiança bruta, os escores de propensão ajustada devem ser calculados ao construir o modelo; caso contrário, eles não estarão disponíveis quando escorar o nugget do modelo.
- Ao escorar o modelo, os escores de propensão ajustada serão incluídos em um campo com as letras *AP* anexadas ao prefixo padrão. Por exemplo, se as predições estiverem em um campo denominado *$R-churn*, o nome do campo de escore de propensão será *$RAP-churn*. Os escores de propensão ajustada não estão disponíveis para modelos de regressão logística.
- Ao calcular os escores de propensão ajustada, a partição de teste ou de validação utilizada para o cálculo não deverá ter sido balanceada. Para evitar isso, assegure-se de que a opção **Balancear somente dados de treinamento** esteja selecionada em qualquer nó Balanceamento de envio de dados. Além disso, se uma amostra complexa tiver sido obtida anteriormente, isto invalidará os escores de propensão ajustada.
- Os escores de propensão ajustada não estão disponíveis para modelos de árvore e de conjunto de regras "impulsionados". Consulte o tópico “Modelos do C5.0 Impulsionados” na página 118 para obter mais informações.

Baseado em. Para que os escores de propensão ajustada sejam calculados, um campo de partição deverá estar presente no fluxo. É possível especificar se deseja utilizar a partição de teste ou de validação para este cálculo. Para obter melhores resultados, a partição de teste ou de validação deverá incluir pelo menos tantos registros quanto a partição usou para treinar o modelo original.

Nó C5.0

Nota: esse recurso está disponível no SPSS Modeler Professional e no SPSS Modeler Premium.

Esse nó utiliza o algoritmo C5.0 para construir uma **árvore de decisão** ou um **conjunto de regras**. Um modelo C5.0 funciona dividindo a amostra com base no campo que fornece o máximo de **ganho de informações**. Cada subamostra definida pela primeira divisão é, então, divida novamente, geralmente com base em um campo diferente, e o processo é repetido até que as subamostras não possam ser divididas ainda mais. Por último, as divisões de nível inferior são reexaminadas e aquelas que não contribuírem significativamente com o valor do modelo são removidas ou **podadas**.

Nota: o nó C5.0 pode prever apenas uma variável resposta categórica. Ao analisar dados com campos categóricos (nominal ou ordinal), o nó mais provavelmente agrupará categorias juntas do que as versões do C5.0 antes da liberação 11.0.

C5.0 pode produzir dois tipos de modelos. Uma **árvore de decisão** é uma descrição clara das divisões localizadas pelo algoritmo. Cada nó terminal (ou "folha") descreve um subconjunto específico dos dados de treinamento, e cada caso nos dados de treinamento pertence exatamente a um nó terminal na árvore. Em outras palavras, exatamente uma predição é possível para qualquer registro de dados específico apresentado para uma árvore de decisão.

Em contraste, um **conjunto de regras** é um conjunto de regras que tenta fazer predições para registros individuais. Os conjuntos de regras são derivados das árvores de decisão e, de certa forma, representam uma versão simplificada ou destilada das informações localizadas na árvore de decisão. Os conjuntos de regras geralmente podem reter a maioria das informações importantes de uma árvore de decisão integral, mas com um modelo menos complexo. Devido à maneira com que os conjuntos de regras funcionam, eles não possuem as mesmas propriedades que as árvores de decisão. A diferença mais importante é que, com um conjunto de regras, mais de uma regra poderá ser aplicada a qualquer registro específico ou nenhuma regra poderá ser aplicada. Se diversas regras se aplicarem, cada regra receberá um "voto" ponderado com

base na confiança associada a essa regra, e a predição final é decidida combinando os votos ponderados de todas as regras que se aplicarem ao registro em questão. Se nenhuma regra se aplicar, uma predição padrão será designada ao registro.

Exemplo. Um pesquisador médico coletou dados de um conjunto de pacientes que tiveram a mesma doença. Durante o curso do tratamento, cada paciente respondeu a um de cinco medicamentos. É possível utilizar um modelo C5.0, em conjunto com outros nós, para ajudar a descobrir qual droga poderá ser apropriada para um paciente futuro com a mesma doença.

Requisitos. Para treinar um modelo C5.0, deverá haver um campo *Destino* categórico (ou seja, nominal ou ordinal) e um ou mais campos de *Entrada* de qualquer tipo. Campos configurados para *Ambos* ou *Nenhum* são ignorados. Os campos utilizados no modelo devem ter seus tipos totalmente instanciados. Um campo de ponderação também pode ser especificado.

Intensidades. Os modelos C5.0 são bastante robustos na presença de problemas como dados omissos e grandes números de campos de entrada. Eles geralmente não requerem longos tempos de treinamento para estimar. Além disso, os modelos C5.0 tendem a ser mais fáceis de entender do que alguns outros tipos de modelo, já que as regras derivadas do modelo possuem uma interpretação muito clara. O C5.0 também oferece o método de **boosting** poderoso para aumentar a precisão do método da classificação.

Nota: a velocidade da construção de modelo do C5.0 poderá ser beneficiada com a ativação do processamento paralelo.

Opções de Modelo do Nó C5.0

Nome do modelo. Especifique o nome do modelo a ser produzido.

- **Automático.** Com essa opção selecionada, o nome do modelo será gerado automaticamente, com base no nome ou nomes do campo de destino. Esse é o padrão.
- **Customizado.** Selecione esta opção para especificar seu próprio nome para o nugget do modelo que será criado por esse nó.

Utilizar dados particionados. Se um campo de partição for definido, essa opção assegurará que apenas os dados da partição de treinamento sejam utilizados para construir o modelo.

Criar modelos de divisão. Constrói um modelo separado para cada valor possível de campos de entrada que são especificados como campos de divisão. Consulte o tópico "Construindo Modelos de Divisão" na página 28 para obter mais informações.

Tipo de saída. Especifique se deseja que o nugget do modelo resultante seja uma **Árvore de Decisão** ou um **Conjunto de regras**.

Agrupar simbólicos. Se essa opção for selecionada, o C5.0 tentará combinar valores simbólicos que tiverem padrões similares com relação ao campo de saída. Se essa opção não for selecionada, o C5.0 criará um nó filho para cada valor do campo simbólico utilizado para dividir o nó pai. Por exemplo, se o C5.0 dividir em um campo *COLOR* (com valores *RED*, *GREEN* e *BLUE*), ele criará uma divisão de três vias por padrão. No entanto, se essa opção for selecionada, e os registros em que *COLOR = RED* forem muito semelhantes aos registros em que *COLOR = BLUE*, ele criará uma divisão de duas vias, com os *GREEN*s em um grupo e os *BLUE*s e *RED*s juntos no outro grupo.

Usar boosting. O algoritmo C5.0 possui um método especial para melhorar a sua taxa de precisão, denominada **boosting**. Ele funciona ao construir diversos modelos em uma sequência. O primeiro modelo é construído da maneira usual. Em seguida, um segundo modelo é construído para focar nos registros que forem classificados incorretamente pelo primeiro modelo. Em seguida, um terceiro modelo é construído para focar nos erros do segundo modelo, e assim por diante. Por último, os casos são classificados ao aplicar o conjunto inteiro de modelos neles, utilizando um procedimento de votação ponderada para combinar as predições separadas em uma predição geral. O boosting pode melhorar

significativamente a precisão de um modelo C5.0, como também requer um treinamento mais longo. A opção **Número de avaliações** permite controlar quantos modelos são utilizados para o modelo impulsionado.Esta variável baseia-se na investigação da Freund e Schapire, com algumas melhorias proprietárias para manipular melhor os dados ruidosos.

Validação cruzada. Se essa opção for selecionada, o C5.0 utilizará um conjunto de modelos construídos em subconjuntos dos dados de treinamento para estimar a precisão de um modelo construído no conjunto de dados integral. Isso será útil se o seu conjunto de dados for muito pequeno para dividir em conjuntos de treinamento e de teste tradicionais. Os modelos de validação cruzada são descartados após a estimativa de precisão ser calculada. É possível especificar o **número de dobras** ou o número de modelos utilizados para validação cruzada. Observe que em versões anteriores do IBM SPSS Modeler, construir o modelo e executar uma validação cruzada dele eram duas operações separadas. Na versão atual, nenhum passo de construção de modelo separado é necessário. A construção de modelo e a validação cruzada são executadas ao mesmo tempo.

Modo. Para treinamento **Simples**, a maioria dos parâmetros do C5.0 é configurada automaticamente. O treinamento **Especialista** permite um controle mais direto sobre os parâmetros de treinamento.

Opções do Modo Simples

Favor. Por padrão, C5.0 tentará produzir a árvore mais precisa possível. Em algumas instâncias, isso pode levar a super ajuste, podendo resultar em um fraco desempenho quando o modelo for aplicado aos novos dados. Selecione **Generalidade** para utilizar as configurações de algoritmo que são susceptíveis a esse problema.

Nota: não é garantido que modelos construídos com a opção **Generalidade** selecionada sejam mais bem generalizados do que outros modelos. Quando generalidade for uma questão crítica, sempre valide seu modelo com relação a uma amostra de teste validada.

Ruído esperado (%). Especifique a proporção esperada de dados ruidosos ou errôneos no conjunto de treinamento.

Opções do Modo Especialista

Severidade de Poda. Determina a extensão até a qual a árvore de decisão ou conjunto de regras será podado. Aumente esse valor para obter uma árvore menor e mais concisa. Diminua-o para obter uma árvore mais precisa. Essa configuração afeta apenas a poda local (consulte "Usar poda global" abaixo).

Registros mínimos por ramificação filha. O tamanho de subgrupos pode ser utilizado para limitar o número de divisões em qualquer ramificação da árvore. Uma ramificação da árvore será dividida apenas se duas ou mais sub-ramificações resultantes contiverem pelo menos esta quantidade de registros do conjunto de treinamento. O valor padrão é 2. Aumente este valor para ajudar a evitar **super treinamento** com dados ruidosos.

Usar poda global. As árvores são podadas em dois estágios: primeiro, um estágio de poda local que examina as subárvores e reduz as ramificações para aumentar a precisão do modelo. Segundo, um estágio de poda global considera a árvore como um todo e subárvores fracas podem ser reduzidas. A poda global é executada por padrão. Para omitir o estágio de poda global, desmarque essa opção.

Atributos de Winnow. Se essa opção for selecionada, o C5.0 examinará a utilidade dos preditores antes de começar a construir o modelo. Os preditores que forem considerados irrelevantes serão, então, excluídos do processo de construção de modelo. Esta opção pode ser útil para modelos com muitos campos preditores e pode ajudar a evitar super ajuste.

Nota: a velocidade da construção de modelo do C5.0 poderá ser beneficiada com a ativação do processamento paralelo.

Nó Árvore do AS

O nó Árvore do AS pode ser utilizado com dados em um ambiente distribuído e requer conexão com o IBM SPSS Analytic Server. Neste nó, é possível optar por construir árvores de decisão utilizando um modelo de CHAID ou Exhaustive CHAID.

O CHAID, ou Chi-squared Automatic Interaction Detection, é um método de classificação para construir árvores de decisão usando estatísticas qui-quadrado para identificar divisões ideais.

O CHAID primeiro examina a tabulações cruzadas entre cada um dos campos de entrada e o resultado, e testa a significância usando um teste de independência qui-quadrado. Se mais de uma dessas relações forem estatisticamente significativas, o CHAID selecionará o campo de entrada que for mais significativo (menor valor de *p*). Se uma entrada tiver mais de duas categorias, elas serão comparadas e as categorias que não mostrarem diferenças no resultado serão reduzidas juntas. Isso é feito ao unir sucessivamente o par de categorias que mostrarem a diferença menos significativa. Esse processo de mesclagem de categoria para quando todas as categorias restantes diferirem no nível de teste especificado. Para campos de entrada nominal, quaisquer categorias podem ser mescladas e, para um conjunto ordinal, apenas categorias contínuas podem ser mescladas.

O Exhaustive CHAID é uma modificação do CHAID que executa uma tarefa mais completa de examinar todas as divisões possíveis para cada preditor, mas leva mais tempo para calcular.

Requisitos. Os campos de destino e de entrada podem ser contínuos ou categóricos e os nós podem ser divididos em dois ou mais subgrupos em cada nível. Todos os campos ordinais utilizados no modelo devem ter armazenamento numérico (não sequência de caracteres). Se necessário, use o nó Reclassificar para convertê-los.

Intensidades. O CHAID pode gerar árvores não binárias, significando que algumas divisões possuem mais de duas ramificações. Isso, portanto, tende a criar uma árvore mais larga do que os métodos de crescimento binários. O CHAID funciona para todos os tipos de entradas e aceita ambas ponderações de caso e variáveis de frequência.

Opções de campos do nó Árvore do AS

Na guia Campos, você escolhe se deseja utilizar as configurações de papel do campo já definidas em nós de envio de dados ou fazer as designações de campo manualmente.

Usar papéis predefinidos Esta opção utiliza as configurações de papel (destinos, preditores e assim por diante) a partir de um nó Tipo de envio de dados (ou na guia Tipos de um nó de origem de envio de dados).

Usar designações de campo customizadas Para designar manualmente os destinos, preditores e outros papéis, selecione esta opção.

Campos Use os botões de seta para designar itens manualmente dessa lista para os vários campos de papel à direita da tela. Os ícones indicam os níveis de medição válidos para cada campo de papel.

Para selecionar todos os campos na lista, clique no botão **Tudo**, ou clique em um botão de nível de medição individual para selecionar todos os campos com esse nível de medição.

Destino Selecione um campo como o destino para a predição.

Preditores Selecione um ou mais campos como entradas para a predição.

Ponderação de Análise Para usar um campo como uma ponderação de caso, especifique esse campo aqui. As ponderações de caso são utilizadas para considerar as diferenças na variância nos níveis do campo de saída. Para obter informações adicionais, consulte "Usando Campos de Frequência e de Ponderação" na página 33.

Opções de criação do nó Árvore do AS

A guia Opções de Criação é onde você configura todas as opções para construir o modelo. Obviamente, é possível clicar somente no botão **Executar** para construir um modelo com todas as opções padrão, no entanto, você normalmente deseja customizar a construção para seus próprios propósitos.

A guia contém várias áreas de janela diferentes na quais você configura as customizações que são específicas para seu modelo.

Nó Árvore do AS - Básicos

Especifique as opções básicas sobre como a árvore de decisão deve ser construída.

Algoritmo de crescimento de árvore Selecione o tipo de algoritmo **CHAID** que você deseja utilizar. O **Exhaustive CHAID** é uma modificação do CHAID que executa uma tarefa mais completa de examinar todas as divisões possíveis para cada preditor, mas leva mais tempo para calcular.

Profundidade máxima da árvore Especifique o número máximo de níveis abaixo do nó raiz (o número de vezes em que a amostra será dividida recursivamente); o padrão é 5. O número máximo de níveis (também referidos como *nós*) é 50.000.

Categorização Se você utilizar dados contínuos, deve-se categorizar as entradas. Isso pode ser feito em um nó precedente, no entanto, o nó Árvore do AS categoriza automaticamente quaisquer entradas contínuas. Se utilizar o nó Árvore do AS para categorizar automaticamente os dados, selecione o **Número de categorias** no qual as entradas devem ser divididas. Os dados são divididos em categorias com frequência igual e as opções disponíveis são 2, 4, 5, 10, 20, 25, 50 ou 100.

Nó Árvore do AS - Crescimento

Utilize as opções de crescimento para fazer um ajuste preciso do processo de construção da árvore.

Limite de registro para alternar de valores de p para tamanhos de efeito Especifique o número de registros no qual o modelo alternará do uso de **Configurações de valores de p** para o uso de **Configurações de tamanho de efeito** ao construir a árvore. O padrão é 1.000.000.

Nível de significância para divisão Especifique o nível de significância (alfa) para divisão de nós. O valor deve estar entre 0,05 e 0,95. Valores mais baixos tendem a produzir árvores com menos nós.

Nível de significância para mesclagem Especifique o nível de significância (alfa) para mesclagem de categorias. O valor deve estar entre 0,05 e 0,95. Essa opção não está disponível para o Exhaustive CHAID.

Ajustar valores de significância usando o método de Bonferroni Ajuste os valores de significância quando estiver testando as várias combinações de categoria de um preditor. Os valores são ajustados com base no número de testes, que está diretamente relacionado ao número de categorias e ao nível de medição de um preditor. Isso é geralmente desejável porque controla melhor a taxa de erros de falso positivo. Desativar essa opção aumenta o poder da sua análise para localizar diferenças reais, mas ao custo de uma taxa maior de falso positivo. Em particular, desativar essa opção pode ser recomendado para pequenas amostras.

Limite de tamanho do efeito (apenas variáveis resposta contínuas) Configure o limite de tamanho do efeito a ser utilizado quando dividir nós e mesclar categorias, quando utilizar uma variável resposta contínua. O valor deve estar entre 0,01 e 0,99.

Limite de tamanho do efeito (apenas variáveis resposta categóricas) Configure o limite de tamanho do efeito a ser utilizado quando dividir nós e mesclar categorias, quando utilizar uma variável resposta categórica. O valor deve estar entre 0,01 e 0,99.

Permitir redivisão de categorias mescladas dentro de um nó O algoritmo CHAID tenta mesclar categorias para produzir a árvore mais simples que descreve o modelo. Se selecionada, esta opção permite que categorias mescladas sejam divididas novamente, se isso resultar em uma solução melhor.

Nível de significância para agrupar nós folha Especifique o nível de significância que determina como os grupos de nós folha são formados ou como nós folha incomuns são identificados.

Qui-Quadrado para variáveis resposta categórica Para variáveis resposta categóricas, é possível especificar o método utilizado para calcular estatísticas qui-quadrado.

- **Pearson** Esse método fornece cálculos mais rápidos, mas deve ser utilizado com cuidado em pequenas amostras.
- **Razão de verossimilhança** Este método é mais robusto que o Pearson, mas leva mais tempo para calcular. Para pequenas amostras, este é o método preferencial. Para variáveis resposta contínuas, este método é sempre utilizado.

Nó Árvore do AS - Regras de parada

Estas opções controlam como a árvore é construída. As regras de parada determinam quando parar a divisão de ramificações específicas da árvore. Configure os tamanhos mínimos de ramificação para evitar que as divisões criem subgrupos muito pequenos. **Mínimo de registros na ramificação pai** evitará uma divisão se o número de registros no nó a serem divididos (o *pai*) for menor que o valor especificado. **Mínimo de registros na ramificação filha** evitará uma divisão se o número de registros em qualquer ramificação criada pela divisão (a *filha*) for menor que o valor especificado.

- **Usar porcentagem** Especifique tamanhos em termos de porcentagem de dados de treinamento gerais.
- **Usar valor absoluto** Especifique tamanhos conforme o número absoluto de registros.

Mínimo de mudança nas frequências de célula esperadas Ao estimar as frequências de célula (para o modelo nominal e para o modelo ordinal de efeitos de linha), um procedimento iterativo (epsilon) é utilizado para convergir na estimativa ideal utilizada no teste qui-quadrado para uma divisão específica. O epsilon determina quanta mudança deve ocorrer para que as iterações continuem; se a mudança na última iteração for menor que o valor especificado, a iteração parará. Se você tiver tendo problemas com o algoritmo que não converge, será possível aumentar esse valor ou aumentar o número máximo de iterações até que a convergência ocorra.

Máximo de iterações para convergência Especifica o número máximo de iterações antes de parar, independentemente se a convergência tiver ocorrido ou não.

Nó Árvore do AS - Custos

Em alguns contextos, determinados tipos de erros são mais caros que outros. Por exemplo, pode ser mais caro classificar um solicitante de crédito de alto risco como baixo risco (um tipo de erro) do que classificar um solicitante de baixo risco como alto risco (um tipo diferente de erro). Os custos de classificação errada permitem especificar a importância relativa de diferentes tipos de erros de predição.

Os custos de classificação errada são basicamente ponderações aplicadas a resultados específicos. Essas ponderações são fatoradas no modelo e podem, na realidade, alterar a predição (como uma forma de proteger contra erros caros).

Um modelo que inclui custos poderá não produzir menos erros do que aquele que não inclui e poderá não ter uma classificação mais alta em termos de precisão geral, mas provavelmente executará melhor em termos práticos por possuir um viés integrado a favor de erros menos caros.

A matriz de custo mostra o custo para cada combinação possível de categoria predita e categoria real. Por padrão, todos os custos de classificação errada são configurados como 1,0. Para inserir valores de custo customizado, selecione **Usar custos de classificação errada** e insira os valores customizados na matriz de custo.

Para alterar um custo de classificação errada, selecione a célula correspondente à combinação desejada de valores preditos e reais, exclua o conteúdo existente da célula e insira o custo desejado para a célula. Os custos não são simétricos automaticamente. Por exemplo, se você configurar o custo de classificação errada de *A* como *B* para 2,0, o custo da classificação errada de *B* como *A* ainda terá o valor padrão de 1,0, a menos que você também o altere explicitamente.

Apenas para respostas ordinais, é possível selecionar o **Aumento de custo padrão para resposta ordinal** e configurar os valores padrão na matriz de custos. As opções disponíveis são descritas na lista a seguir.

- **Nenhum aumento** – Um valor padrão de 1,0 para cada predição correta.
- **Linear** - Cada predição incorreta sucessiva aumenta o custo em 1.
- **Quadrado** – Cada predição incorreta sucessiva é o quadrado do valor linear. Nesse caso, os valores podem ser: 1, 4, 9, e assim por diante.
- **Customizado** – Se você editar manualmente quaisquer valores na tabela, a opção suspensa alterará automaticamente para **Customizado**. Se você alterar a seleção suspensa para qualquer uma das outras opções, seus valores editados serão substituídos pelos valores da opção selecionada.

Opções de modelo do nó Árvore do AS

Na guia Opções do modelo, é possível escolher se deseja especificar um nome para o modelo ou gerar um nome automaticamente. Também é possível escolher calcular valores de confiança e incluir um ID de identificação durante a escoragem do modelo.

Nome do modelo É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Calcular confianças Marque essa caixa de seleção para incluir um campo de confiança quando o modelo é escorado.

Identificador de Regra Marque essa caixa de seleção para incluir um campo quando o modelo que contém o ID do nó folha ao qual um registro foi designado é escorado.

Nugget do Modelo Árvore do AS

Saída do nugget do modelo Árvore do AS

Depois de criar um modelo de Árvore do AS, as informações a seguir estão disponíveis no visualizador de saída.

Tabela de informações de modelo

A tabela Informações do Modelo fornece informações chave sobre o modelo. A tabela identifica algumas configurações de modelo de alto nível, como:

- O tipo de algoritmo utilizado, CHAID ou Exhaustive CHAID.
- O nome do campo de destino selecionado na guia Campos do nó Tipo ou do nó Árvore do AS.
- Os nomes dos campos selecionados como preditores na guia Campos do nó Tipo ou do nó Árvore do AS.
- O número de registros nos dados.
- O número de *nós folha* na árvore gerada.

- O número de níveis na árvore, ou seja, a profundidade da árvore.

Importância do preditor

O gráfico Importância do Preditor mostra a importância das 10 principais entradas (preditores) no modelo como um gráfico de barras.

Se houver mais de 10 campos no gráfico, será possível alterar a seleção dos preditores incluídos no gráfico utilizando a régua de controle abaixo do gráfico. As marcas do indicador na régua de controle são uma largura fixa, e cada marca na régua de controle representa 10 campos. É possível mover as marcas do indicador ao longo da régua de controle para exibir os próximos ou os últimos 10 campos, ordenados por importância do preditor.

É possível clicar duas vezes no gráfico para abrir uma caixa de diálogo separada na qual poderá editar as configurações do gráfico. Por exemplo, é possível corrigir itens, como o tamanho do gráfico e o tamanho e a cor das fontes utilizadas. Quando fechar esta caixa de diálogo de edição separada, as mudanças são aplicadas no gráfico que é exibido na guia Saída.

Tabela de Principais Regras de Decisão

Por padrão, esta tabela interativa exibe as estatísticas das regras para os cinco principais nós folha na saída, com base na porcentagem do total de registros que estiverem contidos no nó folha.

É possível clicar duas vezes na tabela para abrir uma caixa de diálogo separada na qual poderá editar as informações de regra que são mostradas na tabela. As informações que são exibidas e as opções que estiverem disponíveis na caixa de diálogo dependem do tipo de dados da resposta, por exemplo, categóricos ou contínuos.

As informações de regra a seguir são mostradas na tabela:
- ID da regra
- Os detalhes de como a regra é aplicada e composta
- Contagem de registros para cada regra
- Porcentagem de registros para cada regra

Além disso, para uma variável resposta contínua, uma coluna extra na tabela mostra o valor da **Média** para cada regra.

É possível alterar o layout da tabela de regra utilizando as opções de **Conteúdo da tabela** a seguir:
- **Principais regras de decisão** As cinco principais regras de decisão são classificadas pela porcentagem do total de registros contidos nos nós folha.
- **Todas as regras** A tabela contém todos os nós folha produzidos pelo modelo, mas mostra apenas 20 regras por página. Ao selecionar este layout, é possível procurar uma regra utilizando as opções adicionais de **Localizar regra por ID** e **Página**.

Além disso, para uma variável resposta categórica, é possível alterar o layout da tabela de regras utilizando a opção **Principais regras por categoria**. As cinco principais regras de decisão são classificadas pela porcentagem do total de registros para uma **Categoria de destino** que você selecionar.

Se você alterar o layout da tabela de regras, será possível copiar a tabela de regras modificada de volta para o visualizador de Saída clicando no botão Copiar para o Visualizador no canto superior esquerdo da caixa de diálogo.

Configurações do nugget do modelo Árvore do AS

Na guia Configurações de um nugget do modelo Árvore do AS, você especifica opções para confiança e para a geração de SQL durante a escoragem de modelo. Esta guia estará disponível somente após o nugget do modelo ser incluído em um fluxo.

Calcular confianças Marque essa caixa de seleção para incluir confianças nas operações de escoragem. Ao escorar modelos no banco de dados, excluir confianças significa que é possível gerar SQL mais eficiente. Para árvores de regressão, as confianças não são designadas.

Identificador de Regra Marque essa caixa de seleção para incluir um campo na saída de escoragem que indica o ID do nó terminal para o qual cada registro é designado.

Gerar SQL para este modelo Ao usar dados de um banco de dados, código SQL pode ser enviado por push de volta para o banco de dados para execução, fornecendo desempenho superior para muitas operações.

Selecione uma das opções a seguir para especificar como o SQL é gerado:

- **Padrão: Escorar usando o Server Scoring Adapter (se instalado) no processo** Se conectado a um banco de dados com um adaptador de escoragem instalado, gera a SQL utilizando o adaptador de escoragem e funções definidas pelo usuário (UDF) associadas e escora seu modelo no banco de dados. Quando nenhum adaptador de escoragem estiver disponível, essa opção busca seus dados novamente a partir do banco de dados e os escora no SPSS Modeler.
- **Escorar fora do Banco de dados** Se selecionada, esta opção busca seus dados novamente a partir do banco de dados e os escora no SPSS Modeler.

Nuggets do modelo de modelo de árvore Árvore C&R, CHAID, QUEST e C5.0

Os nuggets do modelo de árvore de decisão representam as estruturas de árvore para prever um campo de saída específico descoberto por um dos nós de modelagem de árvore de decisão (Árvore C&R, CHAID, QUEST ou C5.0). Os modelos de árvore podem ser gerados diretamente a partir do nó de construção de árvore, ou indiretamente a partir do construtor de árvore interativo. Consulte o tópico “O Construtor de Árvore Interativo” na página 83 para obter mais informações.

Escorando Modelos de Árvore

Ao executar um fluxo contendo um nugget do modelo de árvore, o resultado específico depende do tipo da árvore.

- Para árvores de classificação (variável resposta categórica), dois novos campos contendo o valor predito e a confiança para cada registro são incluídos nos dados. A predição baseia-se na categoria mais frequentes para o nó terminal para o qual o registro é designado; se a maioria dos respondentes de um determinado nó for *sim*, a predição para todos os registros designados a esse nó será sim.
- Para árvores de regressão, apenas os valores preditos são gerados e as confianças não são designadas.
- Opcionalmente, para modelos CHAID, QUEST e Árvores C&R, um campo adicional pode ser incluído que indica o ID do nó ao qual cada registro é designado.

Os novos nomes de campo são derivados do nome do modelo ao incluir prefixos. Para Árvore C&R, CHAID e QUEST, os prefixos são *$R-* para o campo de predição, *$RC-* para o campo de confiança e *$RI-* para o campo identificador de nó. Por árvores C5.0, os prefixos são *$C-* para o campo de predição e *$CC-* para o campo confiança. Se diversos nós de modelo de árvore estiverem presentes, os novos nomes de campo incluirão os números no *prefixo* para distingui-los, se necessário, por exemplo, *$R1-* e *$RC1-* e *$R2-*.

Trabalhando com Nuggets do Modelo de Árvore

É possível salvar ou exportar informações relacionadas ao modelo de diversas maneiras.

Nota: Muitas dessas opções também estão disponíveis a partir da janela do construtor de árvore.

A partir do construtor de árvore ou de um nugget do modelo de árvore, é possível:

- Gerar um filtro ou selecionar um nó com base na árvore atual. Consulte o tópico "Gerando Nós Filtro e Seleção" na página 93 para obter mais informações.
- Gerar um nugget do Conjunto de Regras que representa a estrutura em árvore como um conjunto de regras que definem as ramificações de terminal da árvore. Consulte o tópico "Gerando um Conjunto de Regras a partir de uma Árvore de Decisão" na página 93 para obter mais informações.
- Além disso, é possível exportar o modelo no formato PMML apenas para nuggets do modelo de árvore. Consulte o tópico "A Paleta de Modelos" na página 41 para obter mais informações. Se o modelo incluir quaisquer divisões customizadas, essas informações não serão preservadas no PMML exportado. (A divisão é preservada, mas o fato de que ela é customizada ao invés de escolhida pelo algoritmo não é).
- Gerar um gráfico com base em uma parte selecionada da árvore atual. *Nota*: isto funcionará apenas para um nugget quando ele estiver anexado a outros nós em um fluxo. Consulte o tópico "Gerando Gráficos" na página 118 para obter mais informações.
- Apenas para modelos C5.0 impulsionados, é possível escolher **Árvore de Decisão Única (Tela)** ou **Árvore de Decisão Única (Paleta GM)** para criar um novo conjunto de regras único derivado da regra atualmente selecionada. Consulte o tópico "Modelos do C5.0 Impulsionados" na página 118 para obter mais informações.

Nota: Embora o nó Regra de Construção tenha sido substituído pelo nó Árvore C&R, os nós de árvore de decisão em fluxos existentes que foram criados originalmente utilizando um nó Regra de Construção ainda continuarão a funcionar corretamente.

Nuggets do Modelo de Árvore Única

Se você selecionar **Construir uma única árvore** como o objetivo principal no nó de modelagem, o nugget do modelo resultante conterá as guias a seguir.

Tabela 7. Guias em um nugget de árvore única

Tabulação	Descrição	Informações Adicionais
Modelo	Exibe as regras que definem o modelo.	Consulte o tópico "Regras de Modelos de Árvore de Decisão" na página 115 para obter mais informações.
Visualizador	Exibe a visualização em árvore do modelo.	Consulte o tópico "Visualizador de Modelos de Árvore de Decisão" na página 116 para obter mais informações.
Sumarização	Exibe informações sobre os campos, as configurações de construção e o processo de estimação do modelo.	Consulte o tópico "Sumarização / Informações do Nugget do Modelo" na página 43 para obter mais informações.
Configurações	Permite especificar opções para confiança e para a geração de SQL durante a escoragem do modelo.	Consulte o tópico "Configurações de Nugget do Modelo Árvore de Decisão/Conjunto de Regras" na página 117 para obter mais informações.
Anotação	Permite incluir anotações descritivas, especificar um nome customizado, incluir texto da dica de ferramenta e especificar palavras-chave de procura para o modelo.	

Regras de Modelos de Árvore de Decisão

A guia Modelo de um nugget da árvore de decisão exibe as regras que definem o modelo. Opcionalmente, um gráfico de importância do preditor e um terceiro painel com informações sobre o histórico, frequências e substitutos também podem ser exibidos.

Nota: Se selecionar a opção **Criar um modelo para conjuntos de dados muito grandes** na guia Opções de Criação do nó CHAID (painel Objetivo), a guia Modelo exibirá apenas os detalhes da regra da árvore.

Regras de Árvore

A área de janela esquerda exibe uma lista de condições que definem o particionamento de dados descobertos pelo algoritmo – essencialmente uma série de regras que podem ser utilizadas para designar registros individuais para os nós filhos com base nos valores de preditores diferentes.

As árvores de decisão funcionam ao particionar recursivamente os dados com base nos valores de campo de entrada. As partições de dados são chamadas de *ramificações*. A ramificação inicial (às vezes chamada de *raiz*) inclui todos os registros de dados. A raiz é dividida em subconjuntos, ou *ramificações filas*, com base no valor de um determinado campo de entrada. Cada ramificação filha pode ser dividida ainda mais em sub-ramificações, que podem, por sua vez, ser divididas novamente, e assim por diante. No nível mais baixo da árvore estão as ramificações que não possuem mais divisões. Essas ramificações são conhecidas como *ramificações terminais* (ou *folhas*).

Detalhes de regra de árvore

O navegador de regras mostra os valores de entrada que definem cada partição ou ramificação e uma sumarização dos valores de campo de saída para os registros nessa divisão. Para obter informações gerais sobre como utilizar o navegador do modelo, consulte "Procurando Nuggets do Modelo" na página 42.

Para divisões com base em campos numéricos, a ramificação é mostrada por uma linha no formato:

```
fieldname relation value [summary]
```

em que *relation* é uma relação numérica. Por exemplo, uma ramificação definida por valores maiores que 100 para o campo *revenue* seria mostrada como:

```
revenue > 100 [summary]
```

Para divisões com base em campos simbólicos, a ramificação é mostrada por uma linha no formato:

```
fieldname = value [summary] or fieldname in [values] [summary]
```

em que *values* representa os valores de campo que definem a ramificação. Por exemplo, uma ramificação que inclui registros em que o valor de *region* pode ser *North*, *West* ou *South* seria representada como:

```
region in ["North" "West" "South"] [summary]
```

Para ramificações terminais, uma predição também é fornecida, incluindo uma seta e o valor predito para o término da condição da regra. Por exemplo, uma folha definida por *revenue* > 100 que prediz um valor de *high* para o campo de saída seria exibida como:

```
revenue > 100 [Mode: high] → high
```

A *sumarização* para a ramificação é definida de forma diferente para campos de saída simbólicos e numéricos. Para árvores com campos de saída numéricos, a sumarização é o valor da *média* da ramificação, e o *efeito* da ramificação é a diferença entre a média da ramificação e a média de sua ramificação pai. Para árvores com campos de saída simbólicos, a sumarização é o *modo*, ou o valor mais frequente, dos registros na ramificação.

Para descrever completamente uma ramificação, é necessário incluir a condição que define a ramificação, mais as condições que definem as divisões mais acima na árvore. Por exemplo, na árvore:

```
revenue > 100
  region = "North"
  region in ["South" "East" "West"]
    revenue <= 200
```

a ramificação representada pela segunda linha é definida pelas condições *revenue > 100* e *region = "North"*.

Se você clicar em **Mostrar Instâncias/Confiança** na barra de ferramentas, cada regra também mostrará informações sobre o número de registros aos quais a regra se aplica (*Instâncias*) e a proporção desses registros para os quais a regra é verdadeira (*Confiança*).

Importância do preditor

Opcionalmente, um gráfico que indica a importância relativa de cada preditor na estimativa do modelo também pode ser exibido na guia Modelo. Geralmente você desejará focar seus esforços de modelagem nos preditores que forem mais importantes e considerar descartar ou ignorar aqueles que forem menos importantes.

Nota: Este gráfico estará disponível apenas se **Calcular a importância do preditor** for selecionada na guia Análise antes de gerar o modelo. Consulte o tópico "Importância do preditor" na página 44 para obter mais informações.

Informações Adicionais do Modelo

Se você clicar em **Mostrar Painel de Informações Adicionais** na barra de ferramentas, um painel será aberto na parte inferior da janela exibindo informações detalhadas da regra selecionada. O painel de informações contém três guias.

Histórico. Esta guia rastreia as condições de divisão do nó raiz até o nó selecionado. Isso fornece uma lista de condições que determinam quando um registro é designado para o nó selecionado. Os registros para os quais todas as condições forem verdadeiras serão designados a este nó.

Frequências. Para modelos com campos de destino simbólicos, essa guia mostra, para cada valor de destino possível, o número de registros designados a esse nó (nos dados de treinamento) que possuírem esse valor de destino. A figura de frequência, expressa como uma porcentagem (mostrada com um máximo de três casas decimais) também é exibida. Para modelos com destinos numéricos, esta guia estará vazia.

Substitutos. Quando aplicável, quaisquer substitutos para o campo de divisão primário são mostrados para o nó selecionado. Os substitutos são campos alternativos utilizados se o valor do preditor primário estiver omisso para um determinado registro. O número máximo de substitutos permitidos para uma determinada divisão é especificado no nó de construção de árvore, mas o número real dependerá dos dados de treinamento. Em geral, quanto mais houver dados omissos, mais substitutos deverão ser utilizados. Para outros modelos de árvore de decisão, essa guia estará vazia.

Nota: Para serem incluídos no modelo, os substitutos devem ser identificados durante a fase de treinamento. Se a amostra de treinamento não possuir nenhum valor omisso, então nenhum substituto será identificado e quaisquer registros com valores omissos encontrados durante o teste ou escoragem serão incluídos automaticamente no nó filho com o maior número de registros. Se valores omissos forem esperados durante os testes ou escoragens, assegure-se de que os valores estejam omissos também na amostra de treinamento. Substitutos não estão disponíveis para as árvores de CHAID.

Visualizador de Modelos de Árvore de Decisão

A guia Visualizador de um nugget do modelo de árvore de decisão é semelhante à exibição no construtor de árvore. A diferença principal é que, quando procurar o nugget do modelo, não é possível crescer ou

modificar a árvore. Outras opções para visualizar e customizar a exibição são semelhantes entre os dois componentes. Consulte o tópico “Customizando a Visualização em Árvore” na página 85 para obter mais informações.

Nota: a guia Visualizador não é exibida para nuggets do modelo CHAID construídos se você selecionar a opção **Criar um modelo para conjuntos de dados muito grandes** no painel Objetivo da guia Opções de Criação.

Ao visualizar regras de divisão na guia Visualizador, os colchetes indicam que o valor adjacente é incluído no intervalo, ao passo que os parênteses indicam que o valor adjacente é excluído do intervalo. A expressão (23,37], portanto, significa de 23 exclusivos para 37 inclusivos, ou seja, exatamente acima de 23 para 37. Na guia Modelo, a mesma condição seria exibida como:

```
Age > 23 and Age <= 37
```

Configurações de Nugget do Modelo Árvore de Decisão/Conjunto de Regras

A guia Configurações de uma árvore de decisão ou de um nugget do modelo Conjunto de Regras permite especificar opções para confianças e para a geração de SQL durante a escoragem de modelo. Esta guia estará disponível somente após o nugget do modelo ter sido incluído em um fluxo.

Calcular confianças Selecione para incluir confianças nas operações de escoragem. Ao escorar modelos no banco de dados, excluir confianças permite gerar SQL mais eficiente. Para árvores de regressão, as confianças não são designadas.

Nota: Se selecionar a opção **Criar um modelo para conjuntos de dados muito grandes** na guia Opções de Criação – painel Método para modelos do CHAID - esta caixa de seleção estará disponível apenas nos nuggets do modelo com variáveis resposta categóricas de nominal ou de flag.

Calcular escores de propensão bruta Para modelos com uma resposta de flag (que retornam uma predição de sim ou não), é possível solicitar os escores de propensão que indicam a probabilidade do resultado real especificado para o campo de destino. Esses são um complemento dos outros valores de predição e de confiança que podem ser gerados durante a escoragem.

Nota: Se selecionar a opção **Criar um modelo para conjuntos de dados muito grandes** na guia Opções de Criação – painel Método para modelos do CHAID - esta caixa de seleção estará disponível apenas nos nuggets do modelo com uma variável resposta categórica de flag.

Calcular escores de propensão ajustada Os escores de propensão bruta baseiam-se apenas nos dados de treinamento e esses podem ser altamente otimistas devido à tendência de muitos modelos para super ajustar esses dados. As propensões ajustadas tentam compensar ao avaliar o desempenho do modelo com relação à partição de teste ou de validação. Essa opção requer que um campo de partição seja definido no fluxo e que os escores de propensão ajustada sejam ativados no nó de modelagem antes de gerar o modelo.

Nota: Os escores de propensão ajustada não estão disponíveis para modelos de árvore e de conjunto de regras impulsionados. Consulte o tópico “Modelos do C5.0 Impulsionados” na página 118 para obter mais informações.

Identificador de regra Para modelos CHAID, QUEST e Árvore C&R, essa opção inclui um campo na saída da escoragem que indica o ID do nó terminal ao qual cada registro é designado.

Nota: Quando essa opção é selecionada, a geração de SQL não está disponível.

Gerar SQL para este modelo Ao usar dados de um banco de dados, código SQL pode ser enviado por push de volta para o banco de dados para execução, fornecendo desempenho superior para muitas operações.

Selecione uma das opções a seguir para especificar como a geração de SQL é executada.

- **Padrão: Escorar usando o Server Scoring Adapter (se instalado) no processo** Se conectado a um banco de dados com um adaptador de escoragem instalado, gera a SQL utilizando o adaptador de escoragem e funções definidas pelo usuário (UDF) associadas e escora seu modelo no banco de dados. Quando nenhum adaptador de escoragem estiver disponível, essa opção busca seus dados novamente a partir do banco de dados e os escora no SPSS Modeler.
- **Escorar ao converter em SQL nativo sem suporte para valor omisso** Se selecionada, gera SQL nativo para escorar o modelo no banco de dados, sem a sobrecarga de manipular valores omissos. Esta opção simplesmente configura a predição para nulo (`$null$`) quando um valor omisso é encontrado ao escorar um caso.

  Nota: Essa opção não está disponível para modelos CHAID. Para outros tipos de modelo, ela está disponível apenas para árvores de decisão (não conjuntos de regras).
- **Escorar ao converter em SQL nativo com suporte para valor omisso** Para modelos CHAID, QUEST, e Árvore C&R, é possível gerar SQL nativo para escorar o modelo no banco de dados com suporte total para valor omisso. Isso significa que o SQL é gerado de modo que os valores omissos sejam manipulados conforme especificado no modelo. Por exemplo, Árvores C&R utilizam regras substitutas e o maior fallback de filho.

  Nota: Para modelos do C5.0, essa opção está disponível somente para conjuntos de regras (não para árvores de decisão).
- **Escorar fora do Banco de dados** Se selecionada, esta opção busca seus dados novamente a partir do banco de dados e os escora no SPSS Modeler.

Modelos do C5.0 Impulsionados

Nota: esse recurso está disponível no SPSS Modeler Professional e no SPSS Modeler Premium.

Ao criar um modelo do C5.0 impulsionado (um conjunto de regras ou uma árvore de decisão), você na realidade cria um conjunto de modelos relacionados. O navegador de regras de modelo para um modelo do C5.0 impulsionado mostra a lista de modelos no nível superior da hierarquia, junto com a precisão estimada de cada modelo e a precisão geral da combinação dos modelos impulsionados. Para examinar as regras ou divisões para um modelo específico, selecione esse modelo e expanda-o conforme você faria com uma regra ou ramificação em um modelo único.

Também é possível extrair um modelo específico a partir do conjunto de modelos impulsionados e criar um novo nugget do modelo de Conjunto de Regras contendo apenas esse modelo. Para criar um novo conjunto de regras a partir de um modelo do C5.0 impulsionado, selecione o conjunto de regras ou a árvore de interesse e escolha **Árvore de Decisão Única (Paleta GM)** ou **Árvore de Decisão Única (Tela)** a partir do menu Gerar.

Gerando Gráficos

Os nós Árvore fornecem muitas informações, no entanto, essas informações nem sempre podem estar em um formato facilmente acessível para usuários de negócios. Para fornecer os dados de modo que possam ser facilmente incorporados em relatórios de negócios, apresentações, e assim por diante, é possível produzir gráficos de dados selecionados. Por exemplo, nas guias Modelo ou Visualizador de um nugget do modelo, ou na guia Visualizador de uma árvore interativa, é possível gerar um gráfico de uma parte selecionada da árvore, criando, assim, um gráfico apenas para os casos na árvore ou no nó de ramificação selecionado.

Nota: é possível gerar um gráfico a partir de um nugget apenas quando ele estiver anexado a outros nós em um fluxo.

Gerar um gráfico

O primeiro passo é selecionar as informações a serem mostradas no gráfico:

- Na guia Modelo de um nugget, expanda a lista de condições e regras na área de janela à esquerda e selecione aquela na qual você está interessado.
- Na guia Visualizador de um nugget, expanda a lista de ramificações e selecione o nó no qual você está interessado.
- Na guia Visualizador de uma árvore interativa, expanda a lista de ramificações e selecione o nó no qual você está interessado.

  Nota: não é possível selecionar o nó superior em qualquer guia Visualizador.

A maneira de criar um gráfico é a mesma, independentemente de como você selecionar os dados a serem mostrados:

1. No menu Gerar, selecione **Gráfico (da seleção)**; como alternativa, na guia Visualizador, clique no botão **Gráfico (da seleção)** no canto inferior esquerdo. A guia Gráfico Básico é exibida.

   Nota: somente as guias Básico e Detalhado estão disponíveis quando exibir o Gráfico desta maneira.
2. Utilizando as configurações da guia Básico ou Detalhado, especifique os detalhes a serem exibidos no gráfico.
3. Clique em OK para gerar o gráfico.

O título do gráfico identifica os nós ou as regras que foram escolhidas para inclusão.

Nuggets do Modelo para Boosting, Bagging e Conjuntos de Dados Muito Grandes

Se você selecionar **Melhorar a precisão do modelo (boosting)**, **Melhorar a estabilidade do modelo (bagging)** ou **Criar um modelo para conjuntos de dados muito grandes** como objetivo principal no nó de modelagem, o IBM SPSS Modeler construirá uma combinação de diversos modelos. Consulte o tópico "Modelos para Combinações" na página 45 para obter mais informações.

O nugget do modelo resultante contém as guias a seguir. A guia Modelo fornece um número de visualizações diferentes do modelo.

Tabela 8. Guias disponíveis no nugget do modelo

Tabulação	Visualizar	Descrição	Informações Adicionais
Modelo	Sumarização do modelo	Exibe uma sumarização da qualidade e (exceto para os modelos e variáveis resposta contínuas impulsionados) diversidade da combinação, uma medida do quanto as predições variam nos diferentes modelos.	Consulte o tópico "Sumarização do Modelo" na página 46 para obter mais informações.
	Importância do preditor	Exibe um gráfico que indica a importância relativa de cada preditor (campo de entrada) na estimativa do modelo.	Consulte o tópico "Importância do Preditor" na página 46 para obter mais informações.
	Frequência do Preditor	Exibe um gráfico que mostra a frequência relativa com a qual cada preditor é utilizado no conjunto de modelos.	Consulte o tópico "Frequência do Preditor" na página 47 para obter mais informações.
	Precisão de Modelo de Componente	Representa um gráfico da precisão preditiva de cada um dos modelos diferentes na combinação.	

Tabela 8. Guias disponíveis no nugget do modelo (continuação)

Tabulação	Visualizar	Descrição	Informações Adicionais
	Detalhes de Modelo de Componente	Exibe informações sobre cada um dos modelos diferentes na combinação.	Consulte o tópico "Detalhes de Modelo de Componente" na página 47 para obter mais informações.
	Informações	Exibe informações sobre os campos, as configurações de construção e o processo de estimação do modelo.	Consulte o tópico "Sumarização / Informações do Nugget do Modelo" na página 43 para obter mais informações.
Configurações		Permite incluir confianças nas operações de escoragem.	Consulte o tópico "Configurações de Nugget do Modelo Árvore de Decisão/Conjunto de Regras" na página 117 para obter mais informações.
Anotação		Permite incluir anotações descritivas, especificar um nome customizado, incluir texto da dica de ferramenta e especificar palavras-chave de procura para o modelo.	

Nuggets do modelo de conjunto de regras Árvore C&R, CHAID, QUEST, C5.0 e a priori

Um nugget do modelo Conjunto de Regras representa as regras para prever um campo de saída específico descoberto pelo nó de modelagem de regra de associação (a priori) ou por um dos nós de construção de árvore (Árvore C&R, CHAID, QUEST ou C5.0). Para regras de associação, o conjunto de regras deve ser gerado a partir de um nugget Regra não refinado. Para árvores, um conjunto de regras pode ser gerado a partir do construtor de árvore interativo, a partir de um nó de construção de modelo C5.0 ou de qualquer nugget do modelo de árvore. Ao contrário dos nuggets Regra não refinados, os nuggets Conjunto de Regras podem ser colocados em fluxos para gerar predições.

Ao executar um fluxo contendo um nugget Conjunto de Regras, dois novos campos são incluídos no fluxo que contém o valor predito e a confiança de cada registro com os dados. Os novos nomes de campo são derivados do nome do modelo ao incluir prefixos. Para conjuntos de regras de associação, os prefixos são *$A-* para o campo de predição e *$AC-* para o campo de confiança. Para conjuntos de regras C5.0, os prefixos são *$C-* para o campo de predição e *$CC-* para o campo de confiança. Para conjuntos de regras Árvore C&R, os prefixos são *$R-* para o campo de predição e *$RC-* para o campo de confiança. Em um fluxo com diversos nuggets Conjunto de Regras em uma série que prediz o mesmo campo ou campos de saída, os novos nomes de campo incluirão números no *prefixo* para diferenciá-los uns dos outros. O primeiro nugget Conjunto de Regras de associação no fluxo utilizará os nomes comuns, o segundo nó utilizará nomes que iniciam com *$A1-* e *$AC1-*, o terceiro nó utilizará nomes que iniciam com *$A2-* e *$AC2-*, e assim por diante.

Como as regras são aplicadas. Os Conjuntos de Regras gerados a partir de regras de associação são diferentes de outros nuggets do modelo porque mais de uma predição pode ser gerada para qualquer registro específico e nem todas essas predições podem estar de acordo. Há dois métodos para gerar predições de conjuntos de regras.

Nota: Os Conjuntos de Regras que forem gerados a partir de árvores de decisão retornam os mesmos resultados, independentemente do método utilizado, uma vez que as regras derivadas de uma árvore de decisão são mutuamente exclusivas.

- **Votação.** Este método tenta combinar as predições de todas as regras que se aplicam ao registro. Para cada registro, todas as regras são examinadas e cada regra que se aplicar ao registro será utilizada para gerar uma predição e uma confiança associada. A soma dos valores de confiança para cada valor de saída é calculada, e o valor com a maior confiança será escolhido como a predição final. A confiança para a predição final é a soma da confiança para esse valor dividido pelo número de regras que forem disparadas para esse registro.
- **Primeira ocorrência**. Este método simplesmente testa as regras em ordem e a primeira regra que se aplicar ao registro é aquela utilizada para gerar a predição.

O método utilizado pode ser controlado nas opções de fluxo.

Gerando nós. O menu Gerar permite criar novos nós com base no conjunto de regras.

- **Nó Filtro** Cria um novo nó Filtro para filtrar campos que não forem utilizados pelas regras no conjunto de regras.
- **Nó Seleção** Cria um novo nó Seleção para selecionar registros aos quais a regra selecionada se aplica. O nó gerado selecionará registros para os quais todos os antecedentes da regra forem verdadeiros. Esta opção requer que uma regra seja selecionada.
- **Nó Rastreio de Regra** Cria um novo SuperNode que calculará um campo que indica qual regra foi utilizada para criar a predição para cada registro. Quando um conjunto de regras é avaliado utilizando o método de primeira ocorrência, isso é apenas um símbolo indicando a primeira regra que seria disparada. Quando o conjunto de regras é avaliado utilizando o método de votação, esta será uma sequência de caracteres mais complexa mostrando a entrada para o mecanismo de votação.
- **Árvore de Decisão Única** (Tela) / **Árvore de Decisão Única** (Paleta GM). Cria um novo nugget Conjunto de Regras único derivado da regra atualmente selecionada. Disponível apenas para os modelos C5.0 **impulsionados**. Consulte o tópico “Modelos do C5.0 Impulsionados” na página 118 para obter mais informações.
- **Modelo para Paleta** Retorna o modelo para a paleta de modelos. Isso é útil em situações em que um colega pode ter enviado um fluxo contendo o modelo e não o próprio modelo.

Nota: As guias Configurações e Sumarização no nugget Conjunto de Regras são idênticas àquelas para os modelos de árvore de decisão.

Guia do Modelo do Conjunto de Regras

A guia Modelo para um nugget de Conjunto de Regras exibe uma lista de regras extraídas dos dados pelo algoritmo.

As regras são divididas por subsequente (categoria predita) e são apresentadas no formato a seguir:

```
if antecedent_1
and antecedent_2
...
and antecedent_n
then predicted value
```

em que `consequent` e `antecedent_1` até `antecedent_n` são todos condições. A regra é interpretada como "para os registros em que `antecedent_1` até `antecedent_n` são todos verdadeiros, o `consequent` também deverá ser verdadeiro". Se você clicar no botão **Mostrar Instâncias/Confiança** na barra de ferramentas, cada regra também mostrará informações sobre o número de registros aos quais a regra se aplica, ou seja, para os quais os antecedentes são verdadeiros (**Instâncias**), e a proporção desses registros para os quais a regra inteira é verdadeira (**Confiança**).

Observe que a confiança é calculada de um modo um pouco diferente para os conjuntos de regras do C5.0. O C5.0 utiliza a fórmula a seguir para calcular a confiança de uma regra:

```
(1 + number of records where rule is correct)
/
(2 + number of records for which the rule's antecedents are true)
```

Este cálculo da estimativa de confiança é adequado para o processo de generalizar as regras a partir de uma árvore de decisão (que é o que o C5.0 faz quando ele cria um conjunto de regras).

Importando Projetos do AnswerTree 3.0

O IBM SPSS Modeler pode importar projetos salvos no AnswerTree 3.0 ou 3.1 utilizando a caixa de diálogo Abrir > Arquivo padrão, conforme a seguir:

1. Nos menus do IBM SPSS Modeler, escolha :

   Arquivo > **Abrir Fluxo**
2. Na lista suspensa de Arquivos do Tipo, selecione **Arquivos do Projeto AT (*.atp, *.atp;)**.

   Cada projeto importado é convertido em um fluxo do IBM SPSS Modeler com os seguintes nós:
   - Um nó de origem que define a origem de dados utilizada (por exemplo, um arquivo de dados ou origem do banco de dados do IBM SPSS Statistics).
   - Para cada árvore no projeto (pode haver várias), um nó Tipo é criado que define propriedades para cada campo (variável), incluindo tipo, papel (campo de entrada ou preditor versus campo de saída ou predito), valores omissos, e outras opções.
   - Para cada árvore no projeto, um nó Partição é criado que particiona os dados para uma amostra de treinamento ou de teste e um nó de construção de árvore é criado que define parâmetros para gerar a árvore (um nó Árvore C&R, QUEST ou CHAID).
3. Para visualizar a árvore ou árvores geradas, execute o fluxo.

Comentários

- As árvores de decisão geradas no IBM SPSS Modeler não podem ser exportadas para o AnswerTree; a importação do AnswerTree para o IBM SPSS Modeler é um percurso unidirecional.
- Os lucros definidos no AnswerTree não são preservados quando o projeto é importado no IBM SPSS Modeler.

Capítulo 7. Modelos de Rede Bayesiana

Nó Rede Bayesiana

O nó **Rede Bayesiana** permite construir um modelo de probabilidade ao combinar evidências observadas e registradas com um conhecimento do mundo real de "senso comum" para estabelecer a probabilidade das ocorrências usando atributos aparentemente desvinculados. O nó foca nas redes Tree Augmented Naïve Bayes (TAN) e Markov Blanket que são utilizadas principalmente para classificação.

As redes bayesianas são utilizadas para fazer predições em muitas situações variadas; alguns exemplos são:

- Selecionar oportunidades de empréstimo com baixo risco padrão.
- Estimar quando um equipamento precisará de serviço, peças ou substituição, com base na entrada do sensor e nos registros existentes.
- Resolver problemas do cliente por meio de ferramentas de resolução de problemas online.
- Diagnosticar e resolver problemas de redes de telefonia celular em tempo real.
- Avaliar possíveis riscos e recompensas de projetos de pesquisa e desenvolvimento para focar os recursos nas melhores oportunidades.

Uma rede bayesiana é um modelo gráfico que exibe as variáveis (geralmente referidas como **nós**) em um conjunto de dados e independências probabilísticas ou condicionais entre elas. Os relacionamentos causais entre os nós podem ser representados por uma rede bayesiana, no entanto, as ligações na rede (também conhecidas como **arcos**) não representam necessariamente causa e efeito diretos. Por exemplo, uma rede bayesiana pode ser utilizada para calcular a probabilidade de um paciente ter uma doença específica, dada a presença ou ausência de determinados sintomas e outros dados relevantes, se as independências probabilísticas entre os sintomas e a doença forem confirmadas, conforme exibido no gráfico. As redes são muito robustas onde as informações estiverem omissas e fazem a melhor predição possível utilizando qualquer informação que estiver presente.

Um exemplo comum e básico de uma rede bayesiana foi criado por Lauritzen e Spiegelhalter (1988). Ele é geralmente referido como o modelo "Ásia" e é uma versão simplificada de uma rede que pode ser utilizada para diagnosticar novos pacientes de um médico, com a direção das ligações correspondendo quase que à causalidade. Cada nó representa uma máscara que pode se relacionar à condição do paciente, por exemplo, "Smoking" confirma que ele é um fumante, e "VisitAsia" mostra se ele visitou a Ásia recentemente. Os relacionamentos de probabilidade são mostrados pelas ligações entre quaisquer nós, por exemplo, fumar aumenta as chances de o paciente desenvolver bronquite e câncer de pulmão, ao passo que idade parece estar associada apenas à possibilidade de desenvolver câncer de pulmão. Da mesma forma, anormalidades encontradas em uma radiografia dos pulmões podem ser causadas por tuberculose ou câncer de pulmão, ao passo que as chances de um paciente sofrer com falta de ar (dispneia) serão maiores se eles também sofrerem de bronquite ou câncer de pulmão.

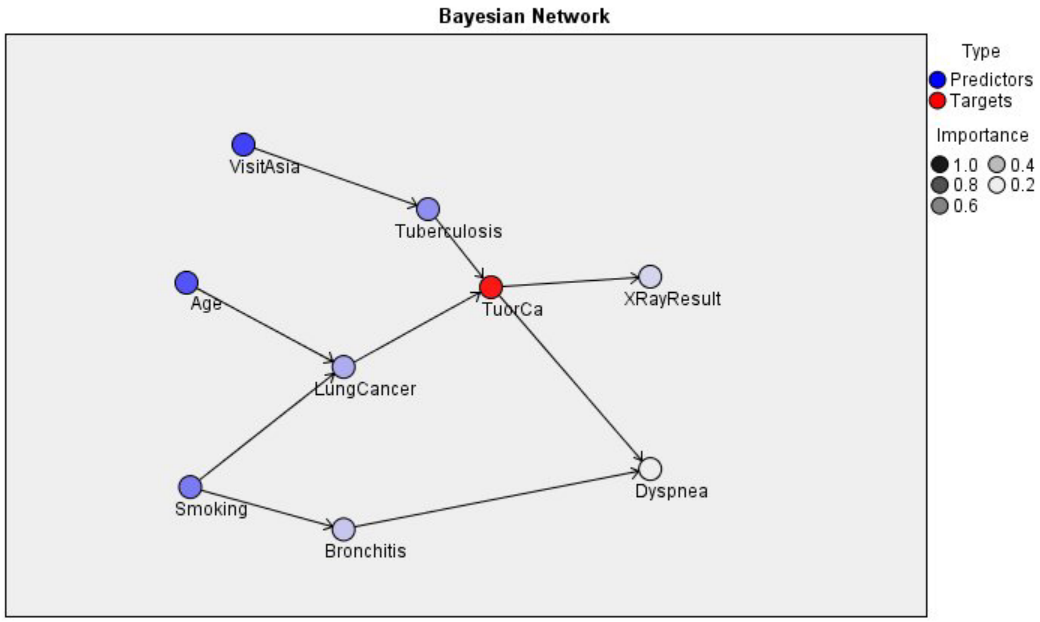


Figura 29. Exemplo da rede Asia de Lauritzen e Spegelhalter

Existem várias razões pelas quais você pode decidir utilizar uma rede bayesiana:

- Ela ajuda a aprender sobre relacionamentos causais. A partir disso, ela permite entender uma área do problema e prever as consequências de qualquer intervenção.
- A rede fornece uma abordagem eficaz para evitar super ajuste de dados.
- Uma visualização clara dos relacionamentos envolvidos é facilmente observada.

Requisitos. Os campos de destino devem ser categóricos e podem ter um nível de medição de *Nominal*, *Ordinal* ou *Flag*. As entradas podem ser campos de qualquer tipo. Os campos de entrada contínuos (intervalo numérico) são categorizados automaticamente, no entanto, se a distribuição estiver defasada, será possível obter melhores resultados ao categorizar manualmente os campos utilizando um nó Categorização antes do nó Rede Bayesiana. Por exemplo, use Categorização Ideal, em que o **Campo Supervisor** é igual ao campo **Destino** do nó Rede Bayesiana.

Exemplo. Um analista de um banco deseja prever clientes ou possíveis clientes que forem susceptíveis a ficarem inadimplentes na amortização do empréstimo. É possível utilizar um modelo de rede bayesiana para identificar as características de clientes mais propensos à inadimplência e construir vários tipos diferentes de modelo para determinar qual é o melhor na predição de possíveis inadimplentes.

Exemplo. Um operador de telecomunicações deseja reduzir o número de clientes que deixam a empresa (conhecido como "migração para o concorrente"), e atualizar o modelo mensalmente utilizando dados de cada mês precedente. É possível utilizar um modelo de rede bayesiana para identificar as características de clientes mais propensos a migrarem para o concorrente e continuar treinando o modelo todos os meses com novos dados.

Opções de Modelo do Nó Rede Bayesiana

Nome do modelo É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Utilizar dados particionados. Se um campo de partição for definido, essa opção assegurará que apenas os dados da partição de treinamento sejam utilizados para construir o modelo.

Construir modelo para cada divisão. Constrói um modelo separado para cada valor possível de campos de entrada que são especificados como campos de divisão. Consulte o tópico "Construindo Modelos de Divisão" na página 28 para obter mais informações.

Partição. Este campo permite especificar um campo utilizado para particionar os dados em amostras separadas para os estágios de treinamento, de teste e de validação de construção de modelo. Ao utilizar uma amostra para gerar o modelo e uma amostra diferente para testá-lo, é possível obter uma boa indicação do quão bem o modelo será generalizado para conjuntos de dados maiores que forem semelhantes aos dados atuais. Se diversos campos de partição tiverem sido definidos usando os nós Tipo ou Partição, um campo de partição único deverá ser selecionado na guia Campos em cada nó de modelagem que utiliza particionamento. (Se apenas uma partição estiver presente, ela será utilizada automaticamente sempre que o particionamento estiver ativado). Além disso, observe que para aplicar a partição selecionada à sua análise, o particionamento também deverá ser ativado na guia Opções de Modelo para o nó. (Desmarcar esta opção permite desativar o particionamento sem alterar as configurações do campo).

Divisões. Para os modelos de divisão, selecione o campo ou campos de divisão. Isso é semelhante a configurar o papel do campo para *Divisão* em um nó Tipo. É possível designar apenas os campos com um nível de medição de **Flag**, **Nominal**, **Ordinal** ou **Contínuo** como campos de divisão. Os campos escolhidos como campos de divisão não podem ser utilizados como campos de destino, de entrada, de partição, de frequência ou de ponderação. Consulte o tópico "Construindo Modelos de Divisão" na página 28 para obter mais informações.

Continuar treinando o modelo existente. Se essa opção for selecionada, os resultados mostrados na guia Modelo do nugget do modelo serão gerados novamente e atualizados toda vez que o modelo for executado. Por exemplo, isso poderá ser feito quando tiver incluído uma origem de dados nova ou atualizada em um modelo existente.

Nota: isso poderá apenas atualizar a rede existente; não será possível incluir ou remover nós ou conexões. Toda vez que você treinar novamente o modelo, a rede terá o mesmo formato, e apenas as probabilidades condicionais e a importância do preditor serão alteradas. Se os novos dados forem muito semelhantes aos seus dados antigos, isso não importa porque você espera que os mesmos itens sejam importantes; no entanto, se você desejar verificar ou atualizar *o que* é importante (ao contrário do quão importante ele é), será necessário construir um novo modelo, ou seja, construir uma nova rede

Tipo de estrutura. Selecione a estrutura a ser utilizada ao construir a rede bayesiana:

- **TAN.** O modelo Tree Augmented Naïve Bayes (TAN) cria um modelo de rede bayesiana simples que é uma melhoria do modelo padrão Naïve Bayes. Isso ocorre porque ele permite que cada preditor dependa de outro preditor além da variável de resposta, aumentando, assim, a precisão da classificação.
- **Markov Blanket.** Isso seleciona o conjunto de nós no conjunto de dados que contêm os pais da variável de resposta, seus filhos e os pais de seus filhos. Essencialmente, um Markov Blanket identifica todas as variáveis na rede que forem necessárias para prever a variável de resposta. Este método de construção de uma rede é considerado mais preciso, no entanto, com conjuntos de dados grandes, o tempo de processamento poderá ser penalizado devido ao alto número de variáveis envolvidas. Para reduzir a quantia de processamento, é possível utilizar as opções de **Seleção de Variável** na guia Especialista para selecionar as variáveis que estiverem significativamente relacionadas à variável de resposta.

Incluir passo de pré-processamento de seleção de variável. Selecionar essa caixa permite utilizar as opções de **Seleção de Variável** na guia Especialista.

Método de aprendizado de parâmetro. Os parâmetros da rede bayesiana referenciam as probabilidades condicionais de cada nó dados os valores de seus pais. Há duas seleções possíveis que podem ser utilizadas para controlar a tarefa de estimativa das tabelas de probabilidade condicional entre os nós nos quais os valores dos pais são conhecidos:

- **Máxima verossimilhança.** Selecione esta caixa quando utilizar um conjunto de dados grande. Esta é a seleção padrão.
- **Ajustamento do Bayes para pequenas contagens de célula.** Para conjuntos de dados menores, há um risco de super ajuste do modelo, além da possibilidade de um alto número de contagens zero. Selecione esta opção para minimizar esses problemas ao aplicar suavização para reduzir o efeito de qualquer contagem zero e quaisquer efeitos de estimativa não confiável.

Opções Avançadas do Nó Rede Bayesiana

As opções avançadas do nó permitem fazer um ajuste preciso do processo de construção de modelo. Para acessar as opções avançadas, configure o Modo para **Especialista** na guia Especialista.

Valores omissos. Por padrão, o IBM SPSS Modeler utiliza apenas registros que tiverem valores válidos para todos os campos utilizados no modelo. (Às vezes isso é chamado de **exclusão de lista** de valores omissos). Se houver muitos dados omissos, você poderá achar que essa abordagem elimina muitos registros, deixando-o sem dados suficientes para gerar um bom modelo. Nesses casos, é possível desmarcar a opção **Usar somente registros completos**. O IBM SPSS Modeler, em seguida, tenta usar o máximo de informações possível para estimar o modelo, incluindo registros nos quais alguns dos campos possuem valores omissos. (Às vezes isso é chamado de **exclusão dos pares** de valores omissos). No entanto, em algumas situações, usar registros incompletos dessa maneira pode levar a problemas computacionais durante a estimativa do modelo.

Incluir todas as probabilidades. Especifica se as probabilidades de cada categoria do campo de saída são incluídas em cada registro processado pelo nó. Se essa opção não estiver selecionada, apenas a probabilidade da categoria predita será incluída.

Teste de independência. Um teste de independência avalia se as observações emparelhadas em duas variáveis são independentes umas das outras. Selecione o tipo de teste a ser utilizado; as opções disponíveis são:

- **Razão de verossimilhança.** Testes de independência de preditor de resposta ao calcular uma razão entre a probabilidade máxima de um resultado sob duas hipóteses diferentes.
- **Qui-quadrado de Pearson.** Testes de independência de preditor de resposta utilizando uma hipótese nula de que as frequências relativas de ocorrência de eventos observados seguem uma distribuição de frequência especificada.

Os modelos de rede bayesiana realizam testes condicionais de independência em que variáveis adicionais são usadas além dos pares testados. Além disso, os modelos exploram não apenas as relações entre a resposta e os preditores, mas também as relações entre os próprios preditores.

Nota: as opções de teste de independência estarão disponíveis apenas se você selecionar **Incluir passo de pré-processamento de seleção de variável** ou um **Tipo de Estrutura** de Markov Blanket na guia Modelo.

Nível de significância. Utilizado em conjunto com as configurações de teste de Independência, permite configurar um valor de corte a ser utilizado durante a realização dos testes. Quanto menor o valor, menos ligações permanecerão na rede; o nível padrão é 0,01.

Nota: essa opção estará disponível apenas se você selecionar **Incluir passo de pré-processamento de seleção de variável** ou um **Tipo de Estrutura** de Markov Blanket na guia Modelo.

Tamanho máximo do conjunto de condicionamento. O algoritmo para criar uma estrutura de Markov Blanket utiliza conjuntos de condicionamento de tamanho crescente para realizar o teste de

independência e remover ligações desnecessárias da rede. Como os testes que envolvem um alto número de variáveis de condicionamento requerem mais tempo e memória para processamento, será possível limitar o número de variáveis a serem incluídas. Isso pode ser útil principalmente ao processar dados com dependências fortes entre muitas variáveis. No entanto, observe que a rede resultante pode conter algumas ligações supérfluas.

Especifique o número máximo de variáveis de condicionamento a serem utilizadas para teste de independência. A configuração padrão é 5.

Nota: essa opção estará disponível apenas se você selecionar **Incluir passo de pré-processamento de seleção de variável** ou um **Tipo de Estrutura** de Markov Blanket na guia Modelo.

Seleção de variável. Essas opções permitem restringir o número de entradas utilizadas ao processar o modelo para acelerar o processo de construção do modelo. Isso é útil principalmente quando criar uma estrutura de Markov Blanket devido ao grande número de entradas possível; ela permite selecionar as entradas que estiverem significativamente relacionadas à variável de resposta.

Nota: as opções de seleção de variáveis estarão disponíveis apenas se você selecionar **Incluir passo de pré-processamento de seleção de variável** na guia Modelo.

- **Entradas sempre selecionadas** Utilizando o Seletor de Campo (botão à direita do campo de texto), selecione os campos do conjunto de dados que sempre serão utilizados ao construir o modelo de rede bayesiana. Observe que o campo de destino é sempre selecionado.
- **Número máximo de entradas.** Especifique o número total de entradas do conjunto de dados a serem utilizadas ao construir o modelo de rede bayesiana. O número mais alto que pode ser inserido é o número total de entradas no conjunto de dados.

  Nota: se o número de campos selecionados em **Entradas sempre selecionadas** exceder o valor de **Número máximo de entradas**, uma mensagem de erro será exibida.

Nuggets do Modelo de Rede Bayesiana

Nota: Se você selecionar **Continuar treinamento de parâmetros existentes** na guia Modelo do nó de modelagem, as informações que são mostradas na guia Modelo do nugget do modelo são atualizadas toda vez que você gerar novamente o modelo.

A guia Modelo do nugget do modelo é dividida em duas áreas de janela.

Área de Janela Esquerda

Básico Esta visualização contém um gráfico de rede de nós que exibe o relacionamento entre a resposta e seus preditores mais importantes, e o relacionamento entre os preditores. A importância de cada preditor é mostrada pela densidade de sua cor, ou seja, uma cor forte mostra um preditor importante, e vice-versa.

Os valores de categoria para nós que representam um intervalo são exibidos em uma dica de ferramenta ao passar o ponteiro do mouse sobre o nó.

É possível utilizar as ferramentas do gráfico no IBM SPSS Modeler para interagir, editar e salvar o gráfico. Por exemplo, para uso em outros aplicativos como do MS Word.

Dica: Se a rede contiver muitos nós, será possível clicar para selecionar um nó e arrastá-lo para tornar o gráfico mais legível.

Distribuição Esta visualização exibe as probabilidades condicionais para cada nó na rede como um minigráfico. Passe o ponteiro do mouse sobre um gráfico para exibir seus valores em dicas de ferramenta.

Área de Janela Direita

Importância do Preditor Exibe um gráfico que indica a importância relativa de cada preditor na estimativa do modelo. Para obter informações adicionais, consulte “Importância do preditor” na página 44.

Probabilidades Condicionais Ao selecionar um nó ou um minigráfico de distribuição na área de janela esquerda, a tabela de probabilidades condicionais associada é exibida na área de janela direita. Esta tabela contém o valor da probabilidade condicional para cada valor do nó e cada combinação de valores em seus nós pai. Além disso, ele inclui o número de registros que são observados para cada valor de registro e cada combinação de valores nos nós pai.

Configurações do Modelo de Rede Bayesiana

A guia Configurações de um nugget do modelo de Rede Bayesiana especifica opções para modificar o modelo construído. Por exemplo, é possível utilizar o nó Rede Bayesiana para construir vários modelos diferentes usando os mesmos dados e configurações e, em seguida, utilizar essa guia em cada modelo para modificar um pouco as configurações para ver como isso afeta os resultados.

Nota: Esta guia estará disponível somente após o nugget do modelo ter sido incluído em um fluxo.

Calcular escores de propensão bruta. Para modelos com uma resposta de flag (que retornam uma predição de sim ou não), é possível solicitar os escores de propensão que indicam a probabilidade do resultado real especificado para o campo de destino. Esses são um complemento dos outros valores de predição e de confiança que podem ser gerados durante a escoragem.

Calcular escores de propensão ajustada. Os escores de propensão bruta baseiam-se apenas nos dados de treinamento e esses podem ser altamente otimistas devido à tendência de muitos modelos a super ajustar desses dados. As propensões ajustadas tentam compensar ao avaliar o desempenho do modelo com relação à partição de teste ou de validação. Essa opção requer que um campo de partição seja definido no fluxo e que os escores de propensão ajustada sejam ativados no nó de modelagem antes de gerar o modelo.

Incluir todas as probabilidades Especifica se as probabilidades de cada categoria do campo de saída são incluídas em cada registro processado pelo nó. Se essa opção não estiver selecionada, apenas a probabilidade da categoria predita será incluída.

A configuração padrão dessa caixa de seleção é determinada pela caixa de seleção correspondente na guia Especialista do nó de modelagem. Consulte o tópico “Opções Avançadas do Nó Rede Bayesiana” na página 126 para obter mais informações.

Gerar SQL para este modelo Ao usar dados de um banco de dados, código SQL pode ser enviado por push de volta para o banco de dados para execução, fornecendo desempenho superior para muitas operações.

Selecione uma das opções a seguir para especificar como a geração de SQL é executada.

- **Padrão: Escorar usando o Server Scoring Adapter (se instalado) no processo** Se conectado a um banco de dados com um adaptador de escoragem instalado, gera a SQL utilizando o adaptador de escoragem e funções definidas pelo usuário (UDF) associadas e escora seu modelo no banco de dados. Quando nenhum adaptador de escoragem estiver disponível, essa opção busca seus dados novamente a partir do banco de dados e os escora no SPSS Modeler.
- **Escorar fora do Banco de dados** Se selecionada, esta opção busca seus dados novamente a partir do banco de dados e os escora no SPSS Modeler.

Sumarização do Modelo de Rede Bayesiana

A guia Sumarização de um nugget do modelo exibe informações sobre o modelo em si (*Análise*), sobre os campos usados no modelo (*Campos*), sobre as configurações utilizadas ao construir o modelo (*Configurações de Construção*) e sobre o treinamento do modelo (*Sumarização do Treinamento*).

Ao procurar o nó pela primeira vez, os resultados da guia Sumarização são reduzidos. Para ver os resultados de interesse, utilize o controle expansor à esquerda de um item para desdobrá-lo ou clique no botão **Expandir Tudo** para mostrar todos os resultados. Para ocultar os resultados após terminar de visualizá-los, use o controle expansor para reduzir os resultados específicos que deseja ocultar ou clique no botão **Reduzir Tudo** para reduzir todos os resultados.

Análise. Exibe informações sobre o modelo específico.

Campos. Lista os campos utilizados como o destino e as entradas na construção do modelo.

Configurações da Construção. Contém informações sobre as configurações utilizadas na construção do modelo.

Sumarização do Treinamento. Mostra o tipo de modelo, o fluxo utilizado para criá-lo, o usuário que o criou, quando ele foi construído e o tempo decorrido para construir o modelo.

Capítulo 8. Redes neurais

Uma **rede neural** pode aproximar uma ampla variedade de modelos preditivos às demandas mínimas na estrutura e suposição de modelo. O formato dos relacionamentos é determinado durante o processo de aprendizado. Se um relacionamento linear entre o destino e os preditores for apropriado, os resultados da rede neural deverão ficar muito próximos dos resultados de um modelo linear tradicional. Se um relacionamento não linear for mais apropriado, a rede neural se aproximará automaticamente da estrutura de modelo "correta".

O trade-off para essa flexibilidade é que a rede neural não é facilmente interpretável. Se estiver tentando explicar um processo subjacente que produz os relacionamentos entre o destino e o preditor, será melhor utilizar um modelo estatístico mais tradicional. No entanto, se a capacidade de interpretação do modelo não for importante, será possível obter boas predições utilizando uma rede neural.

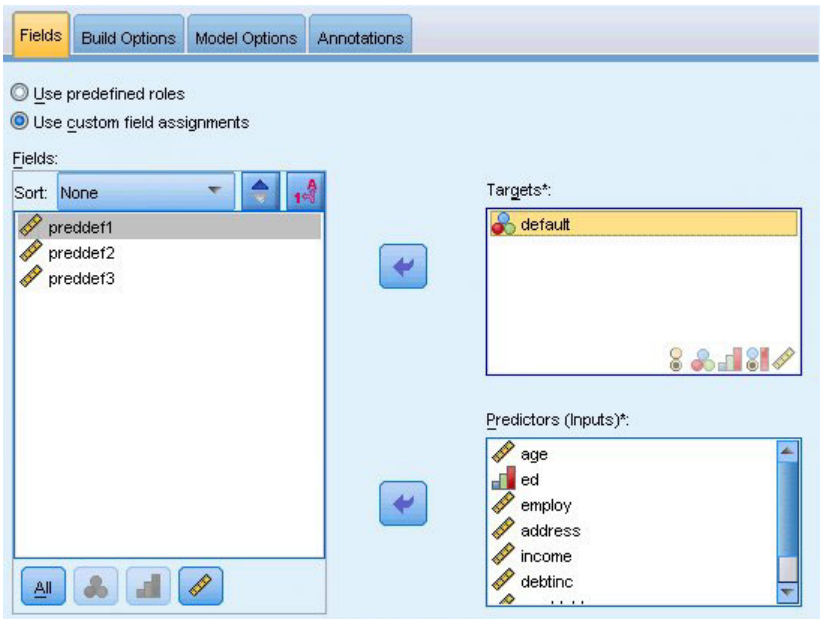


Figura 30. Guia Campos

Requisitos de campo. Deve haver pelo menos um destino e uma entrada. Campos configurados para Ambos ou Nenhum são ignorados. Não há restrições de nível de medição nos destinos ou preditores (entradas). Consulte o tópico “Opções de Campos do Nó de Modelagem” na página 31 para obter mais informações.

O Modelo de Redes Neurais

Redes Neurais são modelos simples do modo com que o sistema nervoso opera. As unidades básicas são **neurônios**, que normalmente são organizados em **camadas**, conforme mostrado na figura a seguir.

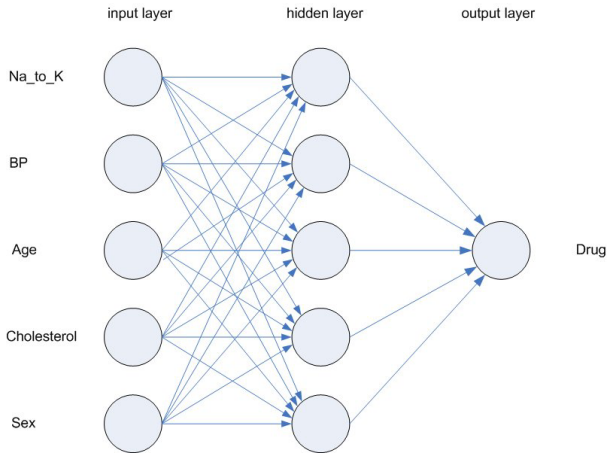


Figura 31. Estrutura de uma rede neural

Uma **rede neural** é um modelo simplificado da maneira com que o cérebro humano processa informações. Ele funciona ao simular um grande número de unidades de processamento interconectadas que lembram versões de neurônios abstratas.

As unidades de processamento são organizadas em camadas. Geralmente há três partes em uma rede neural: uma **camada de entrada**, com as unidades representando os campos de entrada, uma ou mais **camadas ocultas** e uma **camada de saída**, com uma unidade ou unidades representando um ou mais campos de destino. As unidades são conectadas com intensidades (ou **ponderações**) de conexão variadas. Os dados de entrada são apresentados na primeira camada e os valores são propagados de neurônio para neurônio na próxima camada. Por fim, um resultado é entregue a partir da camada de saída.

A rede aprende ao examinar os registros individuais, gerar uma predição para cada registro e fazer ajustamentos nas ponderações sempre que ela fizer uma predição incorreta. Esse processo é repetido muitas vezes e a rede continua a melhorar suas predições até que um ou mais critérios de parada tenham sido atendidos.

Inicialmente, todas as ponderações são aleatórias e as respostas vindas da rede provavelmente não têm sentido. A rede aprende por meio de **treinamento**. Os exemplos para os quais a saída é conhecida são repetidamente apresentados à rede e as respostas que ela fornece são comparadas com os resultados conhecidos. As informações a partir desta comparação são transmitidas de volta por meio da rede, alterando gradualmente as ponderações. Conforme o treinamento progride, a rede se torna cada vez mais precisa em replicar os resultados conhecidos. Depois de treinada, a rede pode ser aplicada em casos futuros nos quais o resultado é desconhecido.

Utilizando Redes Neurais com Fluxos Legados

A Versão 14 do IBM SPSS Modeler introduziu um novo nó Rede Neural, que suporta técnicas e otimização de boosting e de bagging para conjuntos de dados muito grandes. Os fluxos existentes que contêm o nó antigo ainda construirão e escorarão modelos nessa liberação. Entretanto, como esse suporte será removido em uma liberação futura, recomenda-se usar a nova versão a partir de agora.

A partir da versão 13, os campos com valores desconhecidos (ou seja, valores não presentes nos dados de treinamento) não são mais tratados automaticamente como valores omissos e são escorados com o valor `$null$`. Portanto, se desejar escorar campos com valores desconhecidos como não nulos utilizando um modelo de Rede Neural mais antigo (anterior à versão 13) nesta versão 13 ou posterior, deve-se marcar os valores desconhecidos como valores omissos (por exemplo, por meio do nó Tipo).

Observe que, para compatibilidade, quaisquer fluxos legados que ainda contiverem o nó antigo ainda poderão estar usando a opção *Limitar tamanho do conjunto* em **Ferramentas > Propriedades do Fluxo > Opções**; esta opção se aplica apenas a redes Kohonen e nós *K*.-Médias da versão 14 em diante.

Objetivos

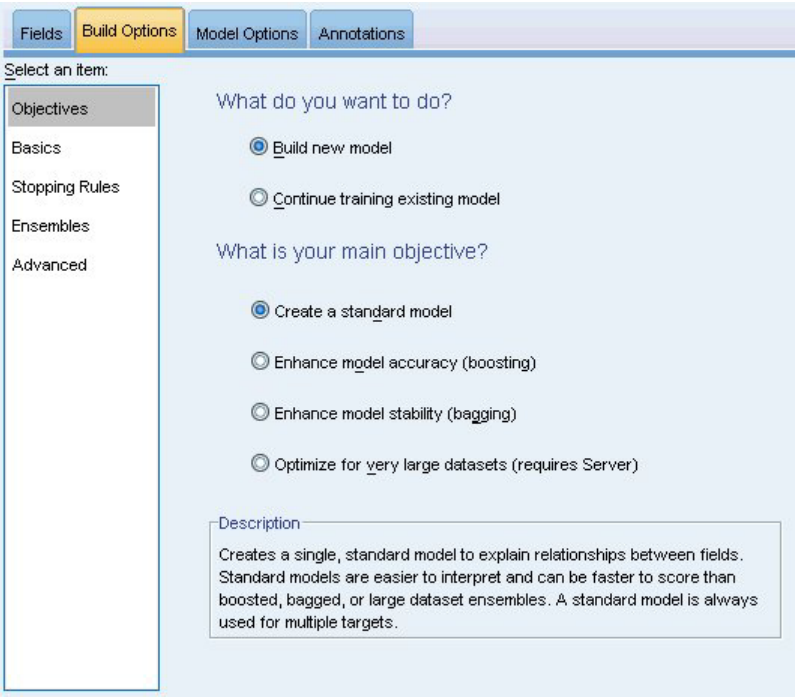


Figura 32. Configurações de objetivos

O que deseja fazer?

- **Construir um novo modelo.** Constrói um modelo completamente novo. Esta é a operação usual do nó.
- **Continuar treinamento de um modelo existente.** O treinamento continua com o último modelo produzido com sucesso pelo nó. Isso permite atualizar ou renovar um modelo existente sem precisar acessar os dados originais e poderá resultar em um desempenho significativamente mais rápido desde que apenas os registros novos ou atualizados sejam alimentados no fluxo. Detalhes do modelo anterior são armazenados com o nó de modelagem, o que permite utilizar essa opção mesmo se o nugget do modelo anterior não estiver mais disponível na paleta de fluxo ou de Modelos.

  Nota: quando essa opção é ativada, todos os outros controles nas guias Campos e Opções de Criação são desativados.

Qual é o seu principal objetivo? Selecione o objetivo apropriado.

- **Criar um modelo padrão.** O método constrói um único modelo para prever a resposta utilizando os preditores. Em geral, os modelos padrão são mais fáceis de interpretar e podem ser mais rápidos para escorar do que combinações de conjunto de dados impulsionados, empacotados ou grandes.

  Nota: Para modelos de divisão, para utilizar esta opção com **Continuar treinando um modelo existente**, você deverá estar conectado ao Analytic Server.
- **Aprimorar a precisão do modelo (boosting).** O método constrói um modelo de combinação usando boosting, que gera uma sequência de modelos para obter predições mais precisas. As combinações podem demorar mais tempo para construir e escorar do que um modelo padrão.

O boosting produz uma sucessão de "modelos de componentes", cada qual construído no conjunto de dados inteiro. Antes de construir cada modelo de componente sucessivo, os registros são ponderados com base nos resíduos do modelo do componente anteriores. Casos com resíduos grandes recebem ponderações de análise relativamente maiores para que o próximo modelo de componente foque na predição também desses registros. Juntos, esses modelos de componente formam um modelo de combinação. O modelo de combinação escora novos registros utilizando uma regra de combinação, e as regras disponíveis dependem do nível de medição da resposta.

- **Aprimorar a estabilidade do modelo (bagging).** O método constrói um modelo de combinação usando bagging (agregação de bootstrap), que gera diversos modelos para obter predições mais confiáveis. As combinações podem demorar mais tempo para construir e escorar do que um modelo padrão.

  A agregação de bootstrap (bagging) produz réplicas do conjunto de dados de treinamento por amostragem com substituição do conjunto de dados original. Isso cria amostras de bootstrap de tamanho igual ao do conjunto de dados original. Em seguida, um "modelo de componente" é construído em cada réplica. Juntos, esses modelos de componente formam um modelo de combinação. O modelo de combinação escora novos registros utilizando uma regra de combinação, e as regras disponíveis dependem do nível de medição da resposta.
- **Criar um modelo para conjuntos de dados muito grandes (requer o IBM SPSS Modeler Server).** O método constrói um modelo de combinação ao dividir o conjunto de dados em blocos de dados separados. Escolha essa opção quando o conjunto de dados for grande demais para construir qualquer um dos modelos acima, ou para construção de modelo incremental. Essa opção pode levar menos tempo para construir, mas pode levar mais tempo para escorar com relação a um modelo padrão. Esta opção requer conectividade com o IBM SPSS Modeler Server .

Quando houver diversos destinos, esse método criará apenas um modelo padrão, independentemente do objetivo selecionado.

Básicos

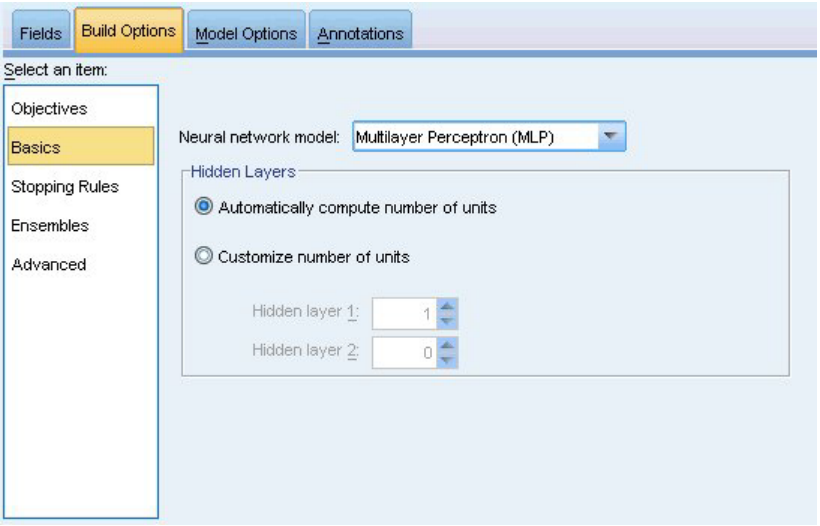


Figura 33. Configurações básicas

Modelo de rede neural. O tipo de modelo determina como a rede conecta os preditores às respostas por meio de uma ou mais camadas ocultas. O **perceptron multicamada (MLP)** permite relacionamentos mais complexos a um custo possivelmente maior de tempo de treinamento e escoragem. Já a **função de base radial (RBF)** pode demandar menos tempo de treinamento e escoragem, porém o poder preditivo pode ser menor em comparação com o MLP.

Camadas Ocultas. Uma ou mais camadas ocultas de uma rede neural contêm unidades não observáveis. O valor de cada unidade oculta é alguma função dos preditores, em que o formato exato da função depende em parte do tipo de rede. Um perceptron multicamadas pode ter uma ou duas camadas ocultas, ao passo que uma rede de função de base radial pode ter uma camada oculta.

- **Calcular automaticamente o número de unidades.** Esta opção constrói uma rede com uma camada oculta e calcula o "melhor" número de unidades na camada oculta.
- **Customizar número de unidades.** Esta opção permite especificar o número de unidades em cada camada oculta. A primeira camada oculta deve ter pelo menos uma unidade. Especificar 0 unidades para a segunda camada oculta constrói um perceptron multicamadas com uma camada oculta única.

  Nota: os valores devem ser escolhidos de modo que o número de nós não exceda o número de preditores contínuos mais o número total de categorias em todos os preditores categóricos (flag, nominal e ordinal).

Regras de Parada

Figura 34. Configurações de Regra de Parada

Estas são as regras que determinam quando parar o treinamento de redes perceptron multicamadas; essas configurações serão ignoradas quando o algoritmo de função de base radial for utilizado. O treinamento continua em pelo menos um ciclo (passagem de dados) e poderá ser interrompido de acordo com os critérios a seguir.

Usar o tempo máximo de treinamento (por modelo de componente). Escolha se deseja especificar o número máximo de minutos para o algoritmo executar. Especifique um número maior que 0. Quando um modelo de combinação for construído, este será o tempo de treinamento permitido para cada modelo de componente da combinação. Observe que treinamento poderá ir um pouco além do limite de tempo especificado para concluir o ciclo atual.

Customizar o número máximo de ciclos de treinamento. O número máximo de ciclos de treinamento permitidos. Se o número máximo de ciclos for excedido, então o treinamento é interrompido. Especifique um número inteiro maior que 0.

Usar precisão mínima. Com essa opção, o treinamento continuará até que a precisão especificada seja alcançada. Isso poderá nunca acontecer, mas é possível interromper o treinamento em qualquer ponto e salvar a rede com a melhor precisão alcançada até o momento.

O algoritmo de treinamento também parará se o erro no conjunto de prevenção ao super ajuste não diminuir após cada ciclo, se a mudança relativa no erro de treinamento for pequena ou se a razão do erro de treinamento atual for pequena em comparação com o erro inicial.

Combinações

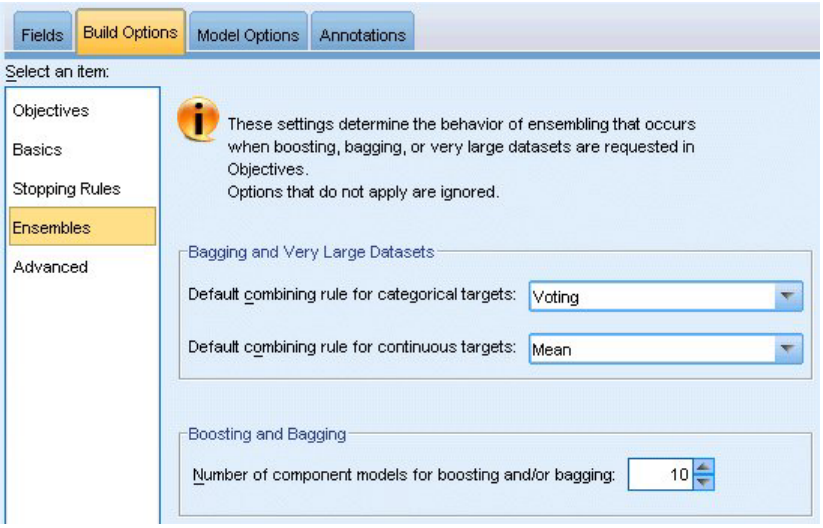


Figura 35. Configurações de combinações

Essas configurações determinam o comportamento da combinação que ocorre quando efetuar boosting, bagging ou quando conjuntos de dados muito grandes forem solicitados nos Objetivos. As opções que não se aplicarem ao objetivo selecionado são ignoradas.

Bagging e Conjuntos de Dados Muito Grandes. Ao escorar uma combinação, essa é a regra utilizada para combinar os valores preditos a partir dos modelos base para calcular o valor de escore de combinação.

- **Regra de combinação padrão para variáveis resposta categórica**. Os valores preditos de combinação para variável resposta categórica podem ser combinados utilizando votação, probabilidade mais alta ou probabilidade média mais alta. **Votação** seleciona a categoria que tem a probabilidade mais alta e mais frequente nos modelos base. **Probabilidade mais alta** seleciona a categoria que atinge a única probabilidade mais alta em todos os modelos base. **Probabilidade média mais alta** Seleciona a categoria com o valor mais alto quando a média das probabilidades da categoria é calculada entre os modelos base.
- **Regra de combinação padrão para variáveis resposta contínuas.** Os valores preditos de combinação para variáveis resposta contínuas podem ser combinados utilizando a média ou mediana dos valores preditos a partir dos modelos base.

Observe que quando o objetivo é melhorar a precisão do modelo, as seleções da regra de combinação são ignoradas. O boosting sempre utiliza uma votação por maioria ponderada para escorar variáveis resposta categóricas e uma média ponderada para escorar variáveis resposta contínuas.

Boosting e Bagging. Especifique o número de modelos base para construção quando o objetivo for melhorar a precisão ou a estabilidade do modelo; para bagging, este é o número de amostras de bootstrap. Ele deve ser um número inteiro positivo.

Avançado

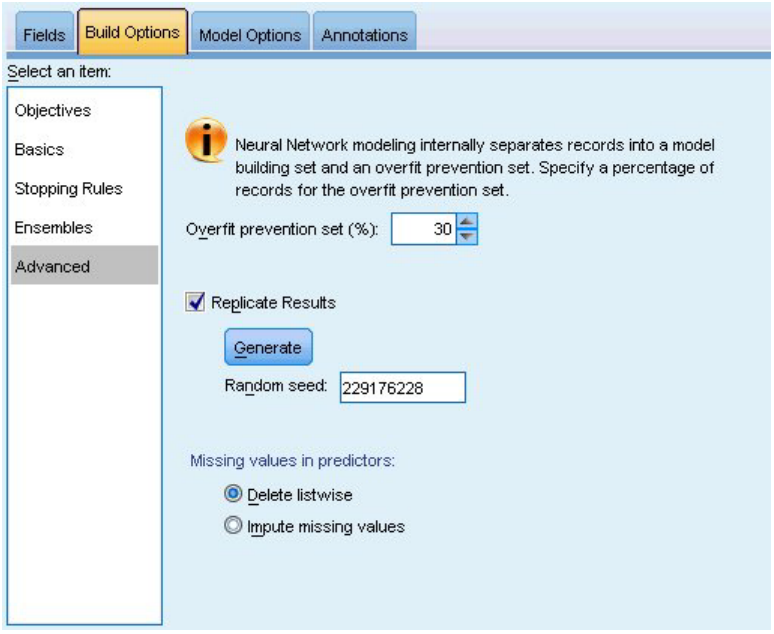


Figura 36. Configurações avançadas

As configurações avançadas fornecem controle sobre as opções que não se ajustarem aos outros grupos de configurações.

Conjunto de prevenção ao super ajuste. O método de rede neural separa internamente os registros em um conjunto de construção de modelo e em um conjunto de prevenção ao super ajuste, que é um conjunto independente de registros de dados utilizados para rastrear erros durante o treinamento a fim de evitar que o método modele a variação de chances nos dados. Especifique uma porcentagem de registros. O padrão é 30.

Replicar resultados. Configurar uma semente aleatória permite replicar análises. Especifique um número inteiro ou clique em **Gerar**, que criará um pseudonúmero inteiro aleatório entre 1 e 2147483647, inclusive. Por padrão, as análises são replicadas com a semente 229176228.

Valores omissos em preditores. Isso especifica como tratar os valores omissos. **Exclusão de lista** remove registros com valores omissos em preditores da construção de modelo. **Imputar valores omissos** substitui valores omissos nos preditores e usa esses registros na análise. Campos contínuos imputam a média dos valores mínimo e máximo observados e campos categóricos imputam a categoria que ocorre com mais frequência. Observe que os registros com valores omissos em qualquer outro campo especificado na guia Campos são sempre removidos da construção de modelo.

Opções de Modelo

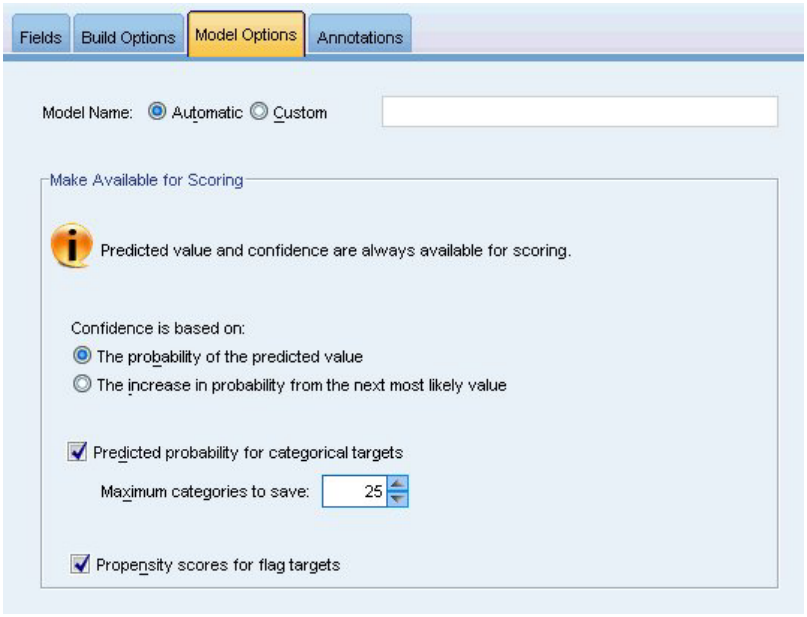


Figura 37. Guia Opções de Modelo

Nome do Modelo. É possível gerar o nome do modelo automaticamente com base nos campos de destino ou especificar um nome customizado. O nome gerado automaticamente é o nome do campo de destino. Se houver diversos destinos, então o nome do modelo será os nomes do campo na ordem, ligados pelo e comercial (símbolo &). Por exemplo, se *field1 field2 e field3* forem os destinos, então o nome do modelo será: *field1 & field2 & field3*.

Disponibilizar para Escoragem. Quando o modelo é escorado, os itens selecionados neste grupo devem ser produzidos. O valor predito (para todas as respostas) e a confiança (para variáveis resposta categóricas) são sempre calculados quando o modelo é escorado. A confiança calculada pode ser baseada na probabilidade do valor predito (a probabilidade predita mais alta) ou na diferença entre a probabilidade predita mais alta e a segunda probabilidade predita mais alta.

- **Probabilidade predita para variáveis resposta categóricas.** Isso produz as probabilidades preditas para variáveis resposta categóricas. Um campo é criado para cada categoria.
- **Escores de propensão para destino de sinalização.** Para modelos com uma resposta de flag (que retornam uma predição de sim ou não), é possível solicitar os escores de propensão que indicam a probabilidade do resultado real especificado para o campo de destino. O modelo produz escores de propensão bruta; se as partições estiverem em vigor, o modelo também produzirá escores de propensão ajustada com base na partição de teste.

Sumarização do Modelo

Model Summary

Target	Previously defaulted
Model	Multilayer Perceptron
Stopping Rule Used	Error cannot be further decreased
Hidden Layer 1 Neurons	4

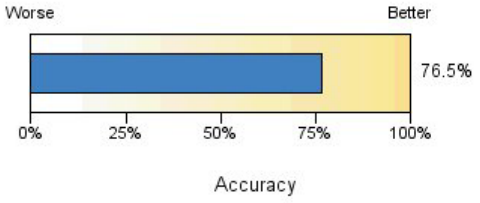


Figura 38. Visualização de Sumarização do Modelo de Redes Neurais

A visualização Sumarização do Modelo é uma captura instantânea ou uma sumarização rápida da precisão de preditiva ou de classificação da rede neural.

Sumarização do modelo. A tabela identifica a resposta, o tipo de rede neural treinado, a regra de parada que parou o treinamento (mostrado se uma rede perceptron multicamada foi treinada) e o número de neurônios em cada camada oculta da rede.

Qualidade da Rede Neural. O gráfico exibe a precisão do modelo final, apresentada em um formato maior e melhor. Para uma variável resposta categórica, a precisão é simplesmente a porcentagem de registros para a qual o valor predito corresponde ao valor observado. Para uma variável resposta contínua, a precisão é 1 menos a razão da média de erro absoluto na predição (a média dos valores absolutos dos valores preditos menos os valores observados) com o intervalo de valores preditos (o valor máximo predito menos o valor mínimo predito).

Diversas respostas. Se houver diversas respostas, então cada resposta será exibida na linha **Resposta** da tabela. A precisão exibida no gráfico é a média das precisões de respostas individuais.

Importância do Preditor

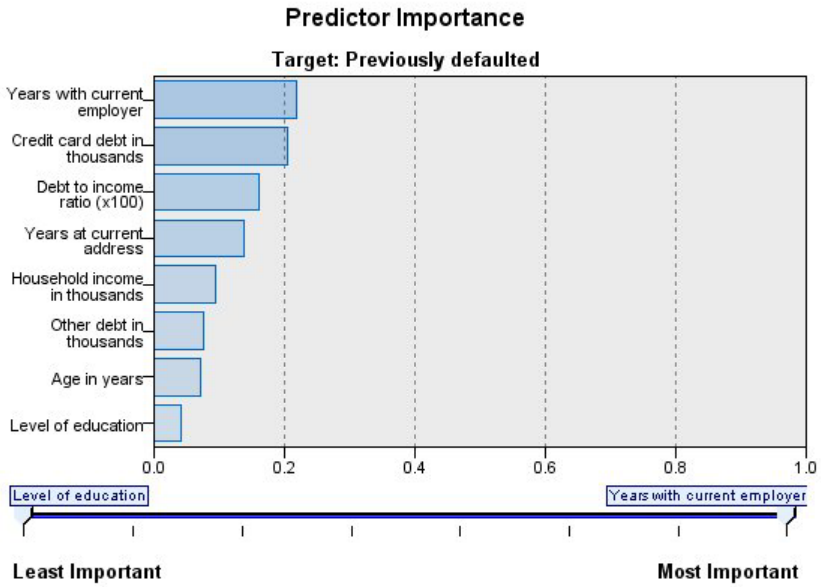


Figura 39. Visualização Importância do Preditor

Geralmente você desejará focar seus esforços de modelagem nos campos preditores que forem mais importantes e considerar descartar ou ignorar aqueles que forem menos importantes. O gráfico de importância do preditor ajuda a fazer isso ao indicar a importância relativa de cada preditor na estimativa do modelo. Como os valores são relativos, a soma dos valores para todos os preditores na tela é 1,0. A importância do preditor não tem relação com a precisão do modelo. Ela está relacionada apenas com a importância de cada preditor em fazer uma predição, e não se a predição é precisa ou não.

Diversas respostas. Se houver diversas respostas, então cada resposta será exibida em um gráfico separado e haverá uma lista suspensa **Resposta** que controla qual resposta exibir.

Predito Por Observado

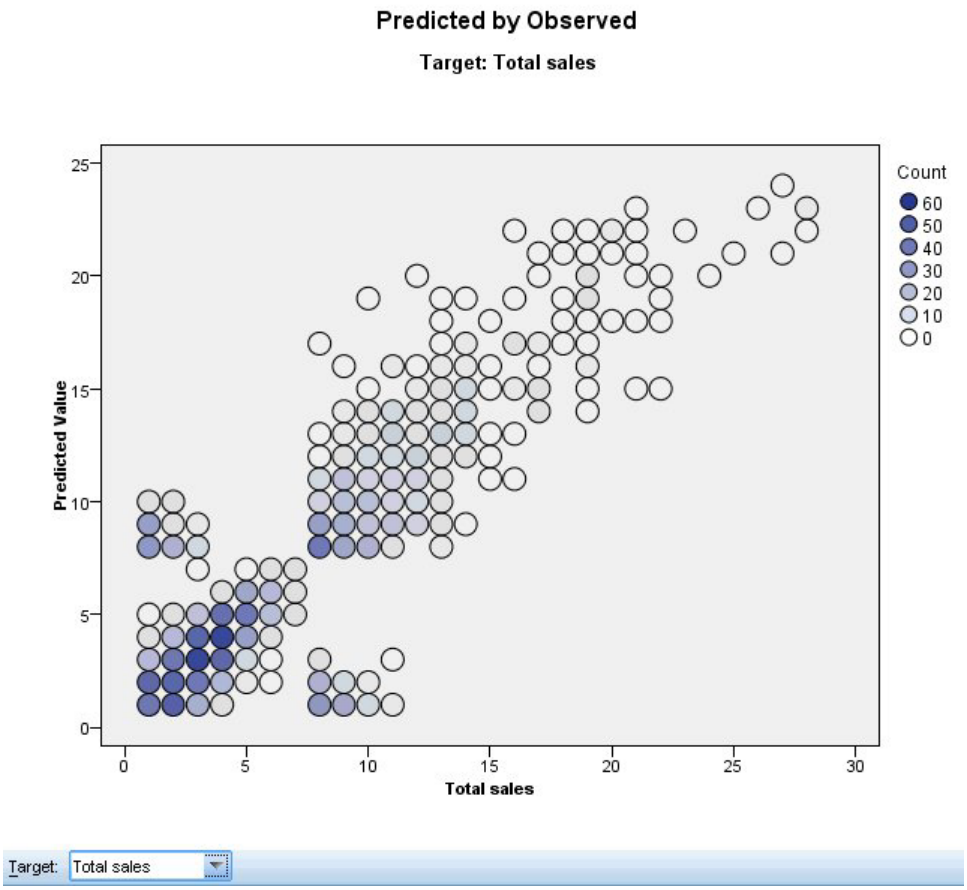


Figura 40. Visualização Predito por Observado

Para variáveis resposta contínuas, isso exibe um gráfico de dispersão categorizado dos valores preditos no eixo vertical em função dos valores observados no eixo horizontal.

Diversas respostas. Se houver diversas variáveis resposta contínuas, então cada resposta será exibida em um gráfico separado e haverá uma lista suspensa **Resposta** que controla qual resposta exibir.

Classificação

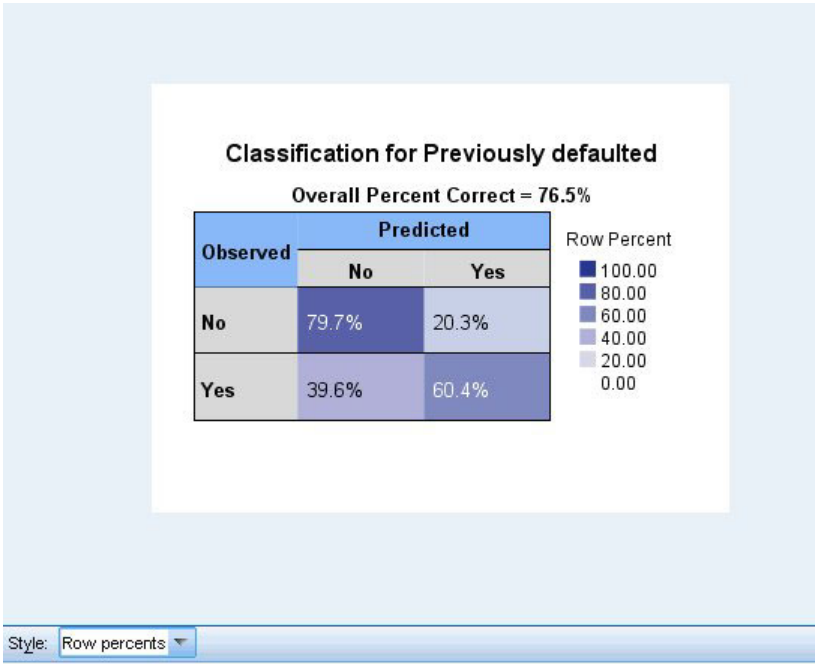


Figura 41. Visualização Classificação, estilos de percentuais de linha

Para variáveis resposta categóricas, isto exibe a classificação cruzada de valores observados versus valores preditos em um heat map, mais o percentual geral correto.

Estilos da tabela. Há vários estilos diferentes de exibição que podem ser acessados a partir da lista suspensa **Estilo**.

- **Percentuais de linhas.** Isso exibe as porcentagens de linhas (as contagens de células expressas como uma porcentagem dos totais de linhas) nas células. Este é o padrão.
- **Contagens de células.** Isso exibe as contagens de célula nas células. O sombreamento para o heat map ainda baseia-se nas porcentagens de linhas.
- **Heat map.** Isso não exibe nenhum valor nas células, apenas o sombreamento.
- **Compactado.** Isso não exibe nenhuma linha ou título da coluna, nem valores nas células. Este pode ser útil quando a resposta tiver um lote de categorias.

Omisso. Se quaisquer registros tiverem valores omissos na resposta, eles serão exibidos em uma linha **(Omisso)** abaixo de todas as linhas válidas. Os registros com valores omissos não contribuem com o percentual geral correto.

Diversas respostas. Se houver diversas variáveis resposta categóricas, então cada resposta será exibida em uma tabela separada e haverá uma lista suspensa **Resposta** que controla qual resposta exibir.

Tabelas Grandes. Se a resposta exibida possuir mais de 100 categorias, nenhuma tabela será exibida.

Rede

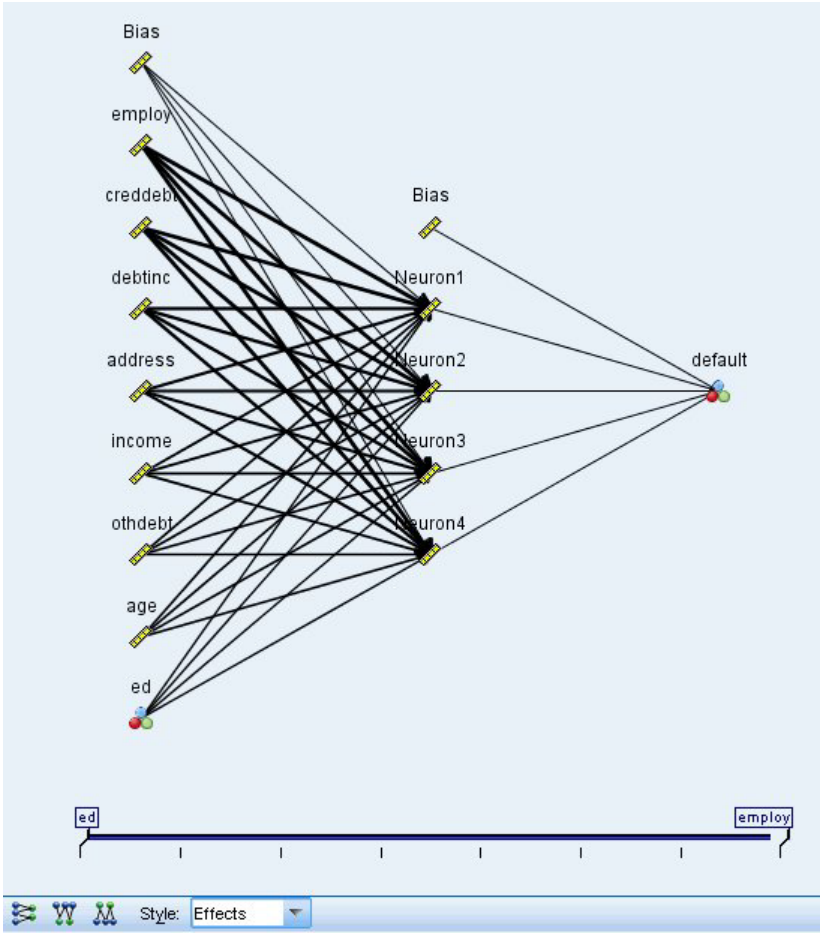


Figura 42. Visualização de rede, entradas à esquerda e estilo de efeitos

Isso exibe uma representação gráfica da rede neural.

Estilos de gráfico. Há dois estilos diferentes de exibição que podem ser acessados a partir da lista suspensa **Estilo**.

- **Efeitos.** Exibe cada preditor e resposta como um nó no diagrama, independentemente se a escala de medição for contínua ou categórica. Este é o padrão.
- **Coeficientes.** Exibe diversos nós indicadores para preditores e respostas categóricos. As linhas de conexão no diagrama de estilo de coeficientes são coloridas com base no valor estimado da ponderação sináptica.

Orientação do diagrama. Por padrão, o diagrama de rede é organizado com as entradas à esquerda e com as respostas à direita. Utilizando os controles da barra de ferramentas, é possível alterar a orientação de modo que as entradas estejam na parte superior e as respostas na parte inferior, ou vice-versa.

Importância do preditor. As linhas de conexão no diagrama são ponderadas com base na importância do preditor, com a largura da linha maior correspondendo à maior importância. Há uma régua de controle de Importância do Preditor na barra de ferramentas que controla quais preditores são mostrados no diagrama de rede. Isso não altera o modelo, apenas permite focar nos preditores mais importantes.

Diversas respostas. Se houver diversas respostas, todas as respostas serão exibidas no gráfico.

Configurações

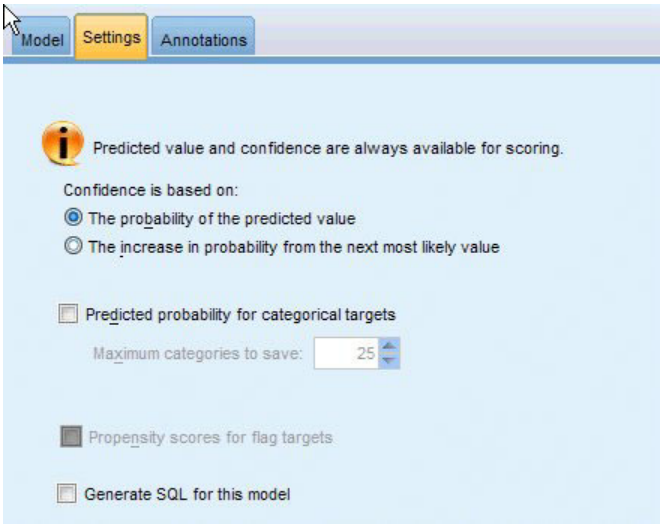


Figura 43. Guia Configurações

Quando o modelo é escorado, os itens selecionados nesta guia devem ser produzidos. O valor predito (para todas as respostas) e a confiança (para variáveis resposta categóricas) são sempre calculados quando o modelo é escorado. A confiança calculada pode ser baseada na probabilidade do valor predito (a probabilidade predita mais alta) ou na diferença entre a probabilidade predita mais alta e a segunda probabilidade predita mais alta.

- **Probabilidade predita para variáveis resposta categóricas.** Isso produz as probabilidades preditas para variáveis resposta categóricas. Um campo é criado para cada categoria.
- **Escores de propensão para destino de sinalização.** Para modelos com uma resposta de flag (que retornam uma predição de sim ou não), é possível solicitar os escores de propensão que indicam a probabilidade do resultado real especificado para o campo de destino. O modelo produz escores de propensão bruta; se as partições estiverem em vigor, o modelo também produzirá escores de propensão ajustada com base na partição de teste.

Gerar SQL para este modelo Ao usar dados de um banco de dados, código SQL pode ser enviado por push de volta para o banco de dados para execução, fornecendo desempenho superior para muitas operações.

Padrão: Escorar usando o Server Scoring Adapter (se instalado) no processo Se conectado a um banco de dados com um adaptador de escoragem instalado, gera a SQL utilizando o adaptador de escoragem e funções definidas pelo usuário (UDF) associadas e escora seu modelo no banco de dados. Quando nenhum adaptador de escoragem estiver disponível, essa opção busca seus dados novamente a partir do banco de dados e os escora no SPSS Modeler.

Escorar ao converter para SQL nativo Se selecionada, gera SQL nativo para escorar o modelo no banco de dados.

Nota: Embora essa opção possa fornecer resultados mais rápidos, o tamanho e a complexidade do SQL nativo aumentam conforme a complexidade do modelo aumenta.

Escorar fora do Banco de dados Se selecionada, esta opção busca seus dados novamente a partir de seu banco de dados e os escora no SPSS Modeler.

Capítulo 9. Lista de Decisão

Os modelos do Lista de Decisão identificam subgrupos ou **segmentos** que mostram uma probabilidade maior ou menor de um resultado binário (sim ou não) com relação à amostra global. Por exemplo, é possível procurar por clientes que forem menos propensos a migrarem para o concorrente ou mais propensos a responderem favoravelmente a uma oferta ou campanha específica. A Decision List Viewer fornece controle completo sobre o modelo, permitindo editar segmentos, incluir suas próprias regras de negócios, especificar como cada segmento é escorado e customizar o modelo de várias outras maneiras para otimizar a proporção de ocorrências em todos os segmentos. Dessa forma, ela é bem apropriada principalmente para gerar listas de distribuição ou, de outra forma, identificar quais registros destinar para uma campanha específica. Também é possível utilizar diversas **tarefas de mineração** para combinar abordagens de modelagem, por exemplo, ao identificar segmentos de alto e baixo desempenho dentro do mesmo modelo e incluir ou excluir cada um deles no estágio de escoragem conforme apropriado.

Segmentos, Regras e Condições

Um modelo consiste em uma lista de segmentos, cada qual definido por uma regra que seleciona registros correspondentes. Uma determinada regra pode ter diversas condições, por exemplo:

```
RFM_SCORE > 10 and
MONTHS_CURRENT <= 9
```

As regras são aplicadas na ordem listada, com a primeira regra de correspondência determinando o resultado de um determinado registro. Obtidas independentemente, as regras ou condições podem se sobrepor, porém a ordem das regras resolve a ambiguidade. Se nenhuma regra corresponder, o registro será designado para a regra restante.

Controle Completo sobre Escoragem

O Decision List Viewer permite visualizar, modificar e reorganizar segmentos e escolher quais deles incluir ou excluir para propósitos de escoragem. Por exemplo, é possível optar por excluir um grupo de clientes e incluir outros para ofertas futuras e ver imediatamente como isso afeta a taxa de ocorrência geral. Os modelos do Lista de Decisão retornam um escore de *Sim* para segmentos incluídos e *$null$* para todos os demais. Esse controle direto sobre a escoragem torna os modelos do Lista de Decisão ideais para gerar listas de distribuição, que são amplamente utilizadas no gerenciamento de relacionamento com o cliente, incluindo central de atendimento ou aplicativos de marketing.

Tarefas de Mineração, Medidas e Seleções

O processo de modelagem é orientado por **tarefas de mineração**. Cada tarefa de mineração inicia efetivamente uma nova execução de modelagem e retorna um novo conjunto de modelos alternativos para escolher. A tarefa padrão baseia-se em suas especificações iniciais no nó do Lista de Decisão, mas é possível definir qualquer número de tarefas customizadas. Também é possível aplicar tarefas iterativamente - por exemplo, é possível executar uma procura de alta probabilidade no conjunto de treinamento inteiro e, em seguida, executar uma procura de baixa probabilidade no restante para eliminar segmentos de baixo desempenho.

Seleções de Dados

É possível definir seleções de dados e medidas de modelo customizadas para construção e avaliação de modelo. Por exemplo, é possível especificar uma seleção de dados em uma tarefa de mineração para customizar o modelo para uma região específica e criar uma medida customizada para avaliar o grau de desempenho desse modelo no país inteiro. Ao contrário de tarefas de mineração, as medidas não alteram o modelo subjacente, mas fornecem outras perspectivas para avaliar o seu grau de desempenho.

Incluindo Seu Conhecimento de Negócios

Ao fazer um ajuste preciso ou estender o segmentos identificados pelo algoritmo, a Decision List Viewer permite incorporar seu conhecimento de negócios diretamente no modelo. É possível editar os segmentos gerados pelo modelo ou incluir segmentos adicionais com base nas regras que você especificar. Em seguida, é possível aplicar as mudanças e visualizar os resultados.

Para obter um insight adicional, uma ligação dinâmica com o Excel permite exportar seus dados para o Excel, no qual ele poderá ser utilizado para criar gráficos de apresentação e calcular medidas customizadas, como lucro complexo e o ROI, que podem ser visualizados na Decision List Viewer enquanto estiver construindo o modelo.

Exemplo. O departamento de marketing de uma instituição financeira deseja obter resultados mais rentáveis em campanhas futuras ao corresponder a oferta certa para cada cliente. É possível utilizar um modelo de Lista de Decisão para identificar as características de clientes com maior probabilidade de responderem favoravelmente com base em promoções anteriores e gerar uma lista de distribuição com base nos resultados.

Requisitos. Um campo de destino categórico único com um nível de medição do tipo *Flag* ou *Nominal* que indica o resultado binário que deseja prever (sim/não), e pelo menos um campo de entrada. Quando o tipo de campo de destino for *Nominal*, deve-se escolher manualmente um valor único a ser tratado como uma **ocorrência** ou **resposta**, e todos os outros valores são agrupados como **não ocorrência**. Um campo de frequência opcional também pode ser especificado. Os campos de data/hora contínuos são ignorados. As entradas do intervalo numérico contínuo são categorizadas automaticamente pelo algoritmo, conforme especificado na guia Especialista no nó de modelagem. Para um controle mais fino sobre a categorização, inclua um nó de categorização de envio de dados e utilize o campo categorizado como entrada com um nível de medição de *Ordinal*.

Opções do Modelo da Lista de Decisão

Nome do modelo É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Utilizar dados particionados. Se um campo de partição for definido, essa opção assegurará que apenas os dados da partição de treinamento sejam utilizados para construir o modelo.

Criar modelos de divisão. Constrói um modelo separado para cada valor possível de campos de entrada que são especificados como campos de divisão. Consulte o tópico “Construindo Modelos de Divisão” na página 28 para obter mais informações.

Modo. Especifica o método utilizado para construir o modelo.

- **Gerar modelo.** Gera automaticamente um modelo na paleta de modelos quando o nó é executado. O modelo resultante pode ser incluído em fluxos para propósitos de escoragem, mas não poderá ser editado posteriormente.
- **Ativar sessão interativa.** Abre a janela de modelagem interativa (saída) do Decision List Viewer, permitindo escolher entre diversas alternativas e aplicar repetidamente o algoritmo com configurações diferentes para aumentar ou modificar progressivamente o modelo. Consulte o tópico “Decision List Viewer” na página 149 para obter mais informações.
- **Usar informações da sessão interativa salva.** Ativa uma sessão interativa utilizando as configurações salvas anteriormente. As configurações interativas podem ser salvas a partir da Decision List Viewer usando o menu Gerar (para criar um modelo ou nó de modelagem) ou o menu Arquivo (para atualizar o nó a partir do qual a sessão foi ativada).

Valor do destino. Especifica o valor do campo de destino que indica o resultado que você deseja modelar. Por exemplo, se o campo de destino de migração para o concorrente for codificado como `0 = no` e `1 = yes`, especifique 1 para identificar as regras que indicam quais registros poderão perder clientes.

Localizar segmentos com. Indica se a procura para a variável de destino deve procurar por uma ocorrência de **Alta probabilidade** ou **Baixa probabilidade**. Localizá-las e excluí-las podem ser uma maneira útil de melhorar seu modelo e podem ser particularmente úteis quando o restante tiver uma probabilidade baixa.

Número máximo de segmentos. Especifica o número máximo de segmentos para retornar. Os *N* principais segmentos são criados, em que o melhor segmento é aquele com a probabilidade mais alta ou, se mais de um modelo possuir a mesma probabilidade, a cobertura mais alta. A configuração mínima permitida é 1; não há configuração máxima.

Tamanho mínimo do segmento. As duas configurações a seguir determinam o tamanho mínimo do segmento. O maior dos dois valores terá precedência. Por exemplo, se o valor da porcentagem for igual a um número maior que o valor absoluto, a configuração da porcentagem terá precedência.

- **Como porcentagem do segmento anterior (%).** Especifica o tamanho mínimo do grupo como uma porcentagem de registros. A configuração mínima permitida é 0, e a configuração máxima permitida é 99,9.
- **Como valor absoluto (N).** Especifica o tamanho mínimo do grupo como um número absoluto de registros. A configuração mínima permitida é 1; não há configuração máxima.

Regras de segmento.

Número máximo de atributos. Especifica o número máximo de condições por regra de segmento. A configuração mínima permitida é 1; não há configuração máxima.

- **Permitir reutilização do atributo.** Quando ativada, cada ciclo pode considerar todos os atributos, mesmo aqueles que foram usados em ciclos anteriores. As condições para um segmento são construídas em ciclos, e cada ciclo inclui uma nova condição. O número de ciclos é definido usando a configuração **Número máximo de atributos**.

Intervalo de confiança para novas condições (%). Especifica o nível de confiança para testar a significância do segmento. Esta configuração desempenha um papel importante no número de segmentos (se houver) que são retornados, bem como a regra de número de condições por segmento. Quanto maior o valor, menor será o conjunto de resultados retornado. A configuração mínima permitida é 50 e a configuração máxima permitida é 99,9.

Opções Avançadas do Nó de Lista de Decisão

As opções avançadas permitem fazer um ajuste preciso do processo de construção de modelo.

Método de categorização. O método utilizado para categorizar campos contínuos (contagem igual ou largura igual).

Número de categorias. O número de categorias para criar campos contínuos. A configuração mínima permitida é 2; não há configuração máxima.

Largura de procura de modelo. O número máximo de resultados de modelo por ciclo que podem ser utilizados para o próximo ciclo. A configuração mínima permitida é 1; não há configuração máxima.

Largura de procura de regra. O número máximo de resultados de regra por ciclo que podem ser utilizados para o próximo ciclo. A configuração mínima permitida é 1; não há configuração máxima.

Fator de mesclagem de categoria. A quantia mínima pela qual um segmento deve crescer quando mesclado com seu vizinho. A configuração mínima permitida é 1,01; não há configuração máxima.

- **Permitir valores omissos nas condições.** `True` para permitir o teste `IS MISSING` nas regras.
- **Descartar resultados intermediários.** Quando `True`, apenas os resultados finais do processo de procura são retornados. Um resultado final é um resultado que não é refinado ainda mais no processo de procura. Quando `False`, os resultados intermediários também são retornados.

  Número máximo de alternativas. Especifica o número máximo de alternativas que podem ser retornadas ao executar a tarefa de mineração. A configuração mínima permitida é 1; não há configuração máxima.

  Observe que a tarefa de mineração retornará apenas o número real de alternativas, até o máximo especificado. Por exemplo, se o máximo for configurado como 100 e apenas 3 alternativas forem localizadas, apenas os 3 serão mostrados.

Nugget do Modelo da Lista de Decisão

Um modelo consiste em uma lista de **segmentos**, cada qual definido por uma **regra** que seleciona registros correspondentes. É possível visualizar ou modificar facilmente os segmentos antes de gerar o modelo e escolher quais deles incluir ou excluir. Quando utilizados na escoragem, os modelos de Lista de Decisão retornam *Sim* para segmentos incluídos e *$null$* para todo o restante. Esse controle direto sobre a escoragem torna os modelos de Lista de Decisão ideais para gerar listas de distribuição, que são amplamente utilizadas no gerenciamento de relacionamento com o cliente, incluindo central de atendimento ou aplicativos de marketing.

Ao executar um fluxo contendo um modelo de Lista de Decisão, o nó inclui três novos campos contendo o escore, *1* (significando *Sim*) para campos incluídos ou *$null$* para os campos excluídos, a probabilidade (taxa de ocorrência) para o segmento no qual o registro cai e o número de ID do segmento. Os nomes dos novos campos são derivados do nome do campo de saída que está sendo predito, prefixado com *$D-* para o escore, *$DP-* para a probabilidade e *$DI-* para o ID do segmento.

O modelo é escorado com base no valor de destino especificado quando o modelo foi construído. Os segmentos podem ser excluídos manualmente para que eles sejam escorados como *$null$*. Por exemplo, se você executar uma procura de baixa probabilidade para localizar segmentos com taxas de ocorrências menores que a média, esses segmentos "baixos" serão escorados como *Sim*, a menos que eles sejam excluídos manualmente. Se necessário, nulos podem ser recodificaos como *Não* utilizando um nó Derivar ou Preenchedor.

PMML

Um modelo de Lista de Decisão pode ser armazenado como um RuleSetModel PMML com o critério de seleção "primeira ocorrência". No entanto, espera-se que todas as regras tenham a mesma escoragem. Para permitir mudanças no campo de destino ou no valor de destino, diversos modelos de conjunto de regras podem ser armazenados em um arquivo a ser aplicado na ordem, os casos que não forem correspondidos pelo primeiro modelo são transmitidos para o segundo, e assim por diante. O nome do algoritmo *DecisionList* é usado para indicar este comportamento não padrão, e apenas os modelos de conjunto de regras com este nome são reconhecidos como modelos de Lista de Decisão e escorados como tal.

Configurações do Nugget do Modelo da Lista de Decisão

A guia Configurações de um nugget do modelo Lista de Decisão permite obter os escores de propensão e ativar ou desativar a otimização de SQL. Esta guia estará disponível somente após incluir o nugget do modelo em um fluxo.

Calcular escores de propensão bruta. Para modelos com uma resposta de flag (que retornam uma predição de sim ou não), é possível solicitar os escores de propensão que indicam a probabilidade do

resultado real especificado para o campo de destino. Esses são um complemento dos outros valores de predição e de confiança que podem ser gerados durante a escoragem.

Calcular escores de propensão ajustada. Os escores de propensão bruta baseiam-se apenas nos dados de treinamento e esses podem ser altamente otimistas devido à tendência de muitos modelos a super ajustar desses dados. As propensões ajustadas tentam compensar ao avaliar o desempenho do modelo com relação à partição de teste ou de validação. Essa opção requer que um campo de partição seja definido no fluxo e que os escores de propensão ajustada sejam ativados no nó de modelagem antes de gerar o modelo.

Gerar SQL para este modelo Ao usar dados de um banco de dados, código SQL pode ser enviado por push de volta para o banco de dados para execução, fornecendo desempenho superior para muitas operações.

Selecione uma das opções a seguir para especificar como a geração de SQL é executada.

- **Padrão: Escorar usando o Server Scoring Adapter (se instalado) no processo** Se conectado a um banco de dados com um adaptador de escoragem instalado, gera a SQL utilizando o adaptador de escoragem e funções definidas pelo usuário (UDF) associadas e escora seu modelo no banco de dados. Quando nenhum adaptador de escoragem estiver disponível, essa opção busca seus dados novamente a partir do banco de dados e os escora no SPSS Modeler.
- **Escorar ao converter para SQL nativo** Se selecionada, gera SQL nativo para escorar o modelo no banco de dados.

  Nota: Embora essa opção possa fornecer resultados mais rápidos, o tamanho e a complexidade do SQL nativo aumentam conforme a complexidade do modelo aumenta.
- **Escorar fora do Banco de dados** Se selecionada, esta opção busca seus dados novamente a partir de seu banco de dados e os escora no SPSS Modeler.

Decision List Viewer

A interface gráfica da Decision List Viewer baseada em tarefa e fácil de usar elimina a complexidade do processo de construção do modelo, libertando você dos detalhes de nível inferior das técnicas de mineração de dados e permitindo dedicar toda a sua atenção para as partes da análise que requerem intervenção do usuário, como configurar objetivos, escolher grupos de destino, analisar os resultados e selecionar o modelo ideal.

Área de Janela do Modelo de Trabalho

A área de janela de modelo de trabalho exibe o modelo atual, incluindo tarefas de mineração e outras ações que se aplicam ao modelo de trabalho.

ID. Identifica a ordem do segmento sequencial. Os segmentos modelo são calculados, em sequência, de acordo com seu número de ID.

Regras de Segmento. Fornece o nome do segmento e as condições do segmento definidas. Por padrão, o nome do segmento é o nome do campo ou nomes de campo concatenados utilizados nas condições, com uma vírgula como um separador.

Escore. Representa o campo que você deseja prever, cujo valor supõe-se que esteja relacionado aos valores de outros campos (os preditores).

Nota: as opções a seguir podem ser alternadas para exibir por meio do diálogo do "Organizando Medidas de Modelo" na página 159.

Cobertura. O gráfico de pizza identifica visualmente a cobertura que cada segmento tem com relação à cobertura inteira.

Cobertura (n). Lista a cobertura para cada segmento com relação à cobertura inteira.

Frequência. Lista o número de ocorrências recebidas com relação à cobertura. Por exemplo, quando a cobertura for 79 e a frequência for 50, isso significa que 50 dos 79 responderam para o segmento selecionado.

Probabilidade. Indica a probabilidade do segmento. Por exemplo, quando a cobertura for 79 e a frequência for 50, isso significa que a probabilidade para o segmento é de 63,29% (50 dividido por 79).

Erro. Indica o erro de segmento.

As informações na parte inferior da área de janela indicam a cobertura, a frequência e a probabilidade de todo o modelo.

Barra de Ferramentas do Modelo de Trabalho

A área de janela de modelo de trabalho fornece as funções a seguir por meio de uma barra de ferramentas.

Nota: algumas funções também estão disponíveis clicando com o botão direito em um segmento do modelo.

Tabela 9. Botões da barra de ferramentas do modelo de trabalho.

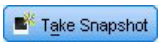
Botão da Barra de Ferramentas	Descrição
	Ativa o diálogo Gerar Novo Modelo, que fornece opções para criar um novo nugget do modelo.
	Salva o estado atual da sessão interativa. O nó de modelagem de Lista de Decisão é atualizado com as configurações atuais, incluindo tarefas de mineração, capturas instantâneas do modelo, seleções de dados e medidas customizadas. Para restaurar uma sessão para esse estado, selecione a caixa **Utilizar as informações da sessão salvas** na guia Modelo do nó de modelagem e clique em **Executar**.
	Exibe o diálogo Organizar Medidas de Modelo. Consulte o tópico "Organizando Medidas de Modelo" na página 159 para obter mais informações.
	Exibe o diálogo Organizar Seleções de Dados. Consulte o tópico "Organizando Seleções de Dados" na página 155 para obter mais informações.
	Exibe a guia Capturas Instantâneas. Consulte o tópico "Guia Capturas Instantâneas" na página 151 para obter mais informações.
	Exibe a guia Alternativos. Consulte o tópico "Guia Alternativos" na página 151 para obter mais informações.
	Faz uma captura instantânea da estrutura do modelo atual. As capturas instantâneas exibem a guia Capturas Instantâneas e são normalmente utilizadas para propósitos de comparação de modelo.
	Ativa o diálogo Inserir Segmentos, que fornece opções para criar novos segmentos modelo.
	Ativa o diálogo Editar Regras Segmento , que fornece opções para incluir condições em segmentos modelo ou alterar condições do segmento modelo definidas anteriormente.
	Move o segmento selecionado para cima na hierarquia do modelo.

Tabela 9. Botões da barra de ferramentas do modelo de trabalho (continuação).

	Move o segmento selecionado para baixo na hierarquia do modelo.
	Exclui o segmento selecionado.
	Alterna se o segmento selecionado é incluído no modelo. Quando removido, os resultados do segmento são incluídos no restante. Isso é diferente de excluir um segmento devido à opção para reativar o segmento.

Guia Alternativos

Gerada quando clicar em **Localizar Segmentos**, a guia Alternativos lista todos os resultados de mineração de alternativas para o modelo ou segmento selecionado na área de janela do modelo de trabalho.

Para promover uma alternativa para ser o modelo de trabalho, destaque a alternativa necessária e clique em **Carregar**; o modelo de alternativa é exibido na área de janela do modelo de trabalho.

Nota: a guia Alternativos será exibida apenas se você tiver configurado o **Número máximo de alternativas** na guia Especialista do no nó de modelagem Lista de Decisão para criar mais de uma alternativa.

Cada alternativa de modelo gerada exibe informações específicas do modelo:

Nome. Cada alternativa é numerada sequencialmente. A primeira alternativa geralmente contém os melhores resultados.

Resposta. Indica o valor da resposta. Por exemplo: 1, que é igual a "true".

Nº de Segmentos. O número de regras de segmento utilizadas no modelo alternativo.

Cobertura. A cobertura do modelo alternativo.

Freq. O número de ocorrências com relação à cobertura.

Prob. Indica a porcentagem de probabilidade do modelo alternativo.

Nota: os resultados alternativos não são salvos com o modelo e são válidos apenas durante a sessão ativa.

Guia Capturas Instantâneas

Uma captura instantânea é uma visualização de um modelo em um momento específico. Por exemplo, é possível fazer uma captura instantânea do modelo quando desejar carregar um modelo alternativo diferente na área de janela de modelo de trabalho, mas não desejar perder o trabalho no modelo atual. A guia Capturas Instantâneas lista todas as capturas instantâneas do modelo feitas manualmente para qualquer número de estados do modelo de trabalho.

Nota: as Capturas Instantâneas são salvas com o modelo. Recomenda-se fazer uma captura instantânea quando carregar o primeiro modelo. Essa captura instantânea irá, então, preservar a estrutura do modelo original, assegurando que você sempre possa retornar para o estado do modelo original. O nome da captura instantânea gerada é exibido como um registro de data e hora, indicando quando ela foi gerada.

Criar uma Captura Instantânea do Modelo

1. Selecione um modelo/alternativa apropriado para exibir na área de janela do modelo de trabalho.
2. Faça todas as mudanças necessárias no modelo de trabalho.

3. Clique em **Fazer Captura Instantânea**. Uma nova captura instantânea é exibida na guia de Capturas Instantâneas.

   Nome. O nome da captura instantânea. É possível alterar um nome de captura instantânea clicando duas vezes no nome da captura instantânea.

   Resposta. Indica o valor da resposta. Por exemplo: 1, que é igual a "true".

   Nº de Segmentos. O número de regras de segmento utilizadas no modelo.

   Cobertura. A cobertura do modelo.

   Freq. O número de ocorrências com relação à cobertura.

   Prob. Indica a porcentagem de probabilidade do modelo.
4. Para promover uma captura instantânea para ser o modelo de trabalho, destaque a captura instantânea necessária e clique em **Carregar**; o modelo de captura instantânea é exibido na área de janela do modelo de trabalho.
5. É possível excluir uma captura instantânea clicando em **Excluir** ou clicando com o botão direito na captura instantânea e escolhendo **Excluir** no menu.

Trabalhando com o Decision List Viewer

Um modelo que melhor preverá a resposta e o comportamento do cliente é construído em vários estágios. Quando o Decision List Viewer é ativado, o modelo de trabalho é preenchido com os segmentos e medidas de modelo definidos, prontos para iniciar uma tarefa de mineração, modificar os segmentos/medidas conforme necessário e gerar um novo modelo ou nó de modelagem.

É possível incluir uma ou mais regras de segmento até que você tenha desenvolvido um modelo satisfatório. É possível incluir regras de segmento no modelo ao executar tarefas de mineração ou utilizar a função **Editar Regra de Segmento**.

No processo de construção de modelo, é possível avaliar o desempenho do modelo ao validar o modelo com relação aos dados da medida, visualizar o modelo em um gráfico ou gerar medidas do Excel customizadas.

Quando sentir-se confiante com relação à qualidade do modelo, será possível gerar um novo modelo e colocá-lo na tela do IBM SPSS Modeler ou na paleta Modelo.

Tarefas de mineração

Uma **tarefa de mineração** é uma coleção de parâmetros que determinam a maneira com que novas regras são geradas. Alguns desses parâmetros são selecionáveis para fornecer flexibilidade para adaptar modelos às novas situações. Uma tarefa consiste em um modelo de tarefa (tipo), um destino e uma seleção de construção (conjunto de dados de mineração).

As seções a seguir detalham as várias operações de tarefa de mineração:

- “Executando Tarefas de Mineração”
- “Criando e Editando uma Tarefa de Mineração” na página 153
- “Organizando Seleções de Dados” na página 155

Executando Tarefas de Mineração: O Decision List Viewer permite incluir manualmente as regras de segmento em um modelo ao executar tarefas de mineração ou copiar e colar as regras de segmento entre os modelos. Uma tarefa de mineração contém informações sobre como gerar novas regras de segmento (as configurações de parâmetro de mineração de dados, como a estratégia de procura, os atributos de origem, a largura da procura, o nível de confiança, e assim por diante), sobre o comportamento do cliente para prever e sobre os dados a investigar. O objetivo de uma tarefa de mineração é procurar as melhores regras de segmento possíveis.

Para gerar uma regra de segmento modelo executando uma tarefa de mineração:

1. Clique na linha **Restante**. Se já houver segmentos exibidos na área de janela de modelo de trabalho, também é possível selecionar um dos segmentos para localizar regras adicionais com base no segmento selecionado. Após selecionar o restante ou o segmento, utilize um dos métodos a seguir para gerar o modelo ou modelos alternativos:

- No menu Ferramentas, escolha **Localizar Segmentos**.
- Clique com o botão direito na linha/segmento **Restante** e escolha **Localizar Segmentos**.
- Clique no botão **Localizar Segmentos** na área de janela do modelo de trabalho.

Enquanto a tarefa estiver sendo processada, o progresso é exibido na parte inferior da área de trabalho e informará quando a tarefa for concluída. O tempo exato que uma tarefa leva para concluir depende da complexidade da tarefa de mineração e do tamanho do conjunto de dados. Se houver apenas um único modelo nos resultados, ele será exibido na área de janela de modelo de trabalho assim que a tarefa for concluída; no entanto, se os resultados contiverem mais de um modelo, eles serão exibidos na guia Alternativos.

Nota: um resultado da tarefa será concluída com modelos, concluída sem modelo ou falha.

O processo de localizar novas regras de segmento pode ser repetido até que nenhuma nova regra seja incluída no modelo. Isso significa que todos os grupos significativos dos clientes foram localizados.

É possível executar uma tarefa de mineração em qualquer segmento modelo existente. Se o resultado de uma tarefa não for o que você está procurando, poderá optar por iniciar outra tarefa de mineração no mesmo segmento. Isso fornecerá regras localizadas adicionais com base no segmento selecionado. Os segmentos que estiverem "abaixo" do segmento selecionado (ou seja, incluídos no modelo posterior ao segmento selecionado) são substituídos pelos novos segmentos porque cada segmento depende de seus predecessores.

Criando e Editando uma Tarefa de Mineração: Uma tarefa de mineração é o mecanismo que procura a coleção de regras que compõem um modelo de dados. Além dos critérios de procura definidos no modelo selecionado, uma tarefa também define a resposta (a questão real que motivou a análise, como quantos clientes provavelmente responderão a uma correspondência) e identifica os conjuntos de dados a serem utilizados. O objetivo de uma tarefa de mineração é procurar os melhores modelos possíveis.

Criar uma tarefa de mineração

Para criar uma tarefa de mineração:

1. Selecione o segmento a partir do qual deseja minerar condições do segmento adicionais.
2. Clique em **Configurações**. O diálogo Criar/Editar Tarefa de Mineração é aberto. Este diálogo fornece opções para definir a tarefa de mineração.
3. Faça todas as mudanças necessárias e clique em **OK** para retornar para a área de janela do modelo de trabalho. O Decision List Viewer utiliza as configurações como os padrões de execução para cada tarefa até que uma tarefa ou configuração alternativa seja selecionada.
4. Clique em **Localizar Segmentos** para iniciar a tarefa de mineração no segmento selecionado.

Editar uma tarefa de mineração

O diálogo Criar/Editar Tarefa de Mineração fornece opções para definir uma nova tarefa de mineração ou editar uma existente.

A maioria dos parâmetros disponíveis para tarefas de mineração é semelhante aos parâmetros oferecidos no nó Lista de Decisão. As exceções são mostradas abaixo. Consulte o tópico “Opções do Modelo da Lista de Decisão” na página 146 para obter mais informações.

Configurações de Carregamento: Quando tiver criado mais de uma tarefa de mineração, selecione a tarefa necessária.

Novo ... Clique para criar uma nova tarefa de mineração com base nas configurações da tarefa exibida atualmente.

Destino

Campo de Destino: Representa o campo que você deseja prever, cujo valor supõe-se que esteja relacionado aos valores de outros campos (os preditores).

Valor do destino. Especifica o valor do campo de destino que indica o resultado que você deseja modelar. Por exemplo, se o campo de destino de migração para o concorrente for codificado como `0 = no` e `1 = yes`, especifique 1 para identificar as regras que indicam quais registros poderão perder clientes.

Configurações simples

Número máximo de alternativas. Especifica o número de alternativas que serão exibidas na execução de tarefa mineração. A configuração mínima permitida é 1; não há configuração máxima.

Configurações avançadas

Editar... Abre o diálogo **Editar Parâmetros Avançados** que permite definir as configurações avançadas. Consulte o tópico "Editar Parâmetros Avançados" para obter mais informações.

Dados

Seleção de construção. Fornece opções para especificar a medida de avaliação que o Decision List Viewer deve analisar para localizar novas regras. As medidas de avaliação listadas são criadas/editadas no diálogo Organizar Seleções de Dados.

Campos disponíveis. Fornece opções para exibir todos os campos ou selecionar manualmente quais campos serão exibidos.

Editar... Se a opção **Customizado** for selecionada, isto abre o diálogo **Customizar Campos Disponíveis** que permite selecionar quais campos estão disponíveis como atributos de segmento localizados pela tarefa de mineração. Consulte o tópico "Customizar Campos Disponíveis" na página 155 para obter mais informações.

Editar Parâmetros Avançados: O diálogo Editar Parâmetros Avançados fornece as opções de configuração a seguir.

Método de categorização. O método utilizado para categorizar campos contínuos (contagem igual ou largura igual).

Número de categorias. O número de categorias para criar campos contínuos. A configuração mínima permitida é 2; não há configuração máxima.

Largura de procura de modelo. O número máximo de resultados de modelo por ciclo que podem ser utilizados para o próximo ciclo. A configuração mínima permitida é 1; não há configuração máxima.

Largura de procura de regra. O número máximo de resultados de regra por ciclo que podem ser utilizados para o próximo ciclo. A configuração mínima permitida é 1; não há configuração máxima.

Fator de mesclagem de categoria. A quantia mínima pela qual um segmento deve crescer quando mesclado com seu vizinho. A configuração mínima permitida é 1,01; não há configuração máxima.

- **Permitir valores omissos nas condições.** `True` para permitir o teste `IS MISSING` nas regras.
- **Descartar resultados intermediários.** Quando `True`, apenas os resultados finais do processo de procura são retornados. Um resultado final é um resultado que não é refinado ainda mais no processo de procura. Quando `False`, os resultados intermediários também são retornados.

Customizar Campos Disponíveis: O diálogo Customizar Campos Disponíveis permite selecionar quais campos estão disponíveis como atributos de segmento localizados pela tarefa de mineração.

Disponíveis. Lista os campos que estão atualmente disponíveis como atributos do segmento. Para remover campos da lista, selecione os campos apropriados e clique em **Remover >>**. Os campos selecionados movem da lista Disponível para a lista Não Disponível.

Não Disponível. Lista os campos que não estão disponíveis como atributos de segmento. Para incluir os campos na lista de disponível, selecione os campos apropriados e clique em **<< Incluir**. Os campos selecionados são movidos da lista Não Disponível para a lista Disponível.

Organizando Seleções de Dados: Ao organizar as seleções de dados (um conjunto de dados de mineração), é possível especificar quais medidas de avaliação a Decision List Viewer deverá analisar para localizar novas regras e escolher quais seleções de dados serão utilizadas como base nas medidas.

Para organizar as seleções de dados:

1. No menu Ferramentas, escolha **Organizar Seleções de Dados** ou clique com o botão direito em um segmento e escolha a opção. O diálogo Organizar Seleções de Dados é aberto.

   Nota: o diálogo Organizar Seleções de Dados também permite editar ou excluir as seleções de dados existentes.
2. Clique no botão **Incluir nova seleção de dados**. Uma nova entrada de seleção de dados é incluída na tabela existente.
3. Clique em **Nome** e insira um nome de seleção apropriado.
4. Clique em **Partição** e selecione um tipo de partição apropriado.
5. Clique em **Condição** e selecione uma opção de condição apropriada. Quando **Especificar** é selecionado, o diálogo Especificar Condição de Seleção é aberto, fornecendo opções para definir condições de campo específicas.
6. Defina a condição apropriada e clique em **OK**.

As seleções de dados estão disponíveis na lista suspensa Seleção de Construção no diálogo Criar/Editar Tarefa de Mineração. A lista permite selecionar qual medida de avaliação é utilizada para uma tarefa de mineração específica.

Regras de segmento

Localize as regras de segmento modelo executando uma tarefa de mineração com base em um modelo de tarefa. É possível incluir manualmente regras de segmento em um modelo utilizando as funções Inserir Segmento ou Editar Regra de Segmento.

Se optar por minerar novas regras de segmento, os resultados, se houver, serão exibidos na guia Visualizador do diálogo Lista Interativa. É possível refinar rapidamente seu modelo ao selecionar um dos resultados alternativos no diálogo Álbuns de Modelo e clicar em **Carregar**. Desta maneira, é possível experimentar resultados diferentes até que você esteja pronto para construir um modelo que descreva precisamente seu grupos de destinos ideal.

Inserindo Segmentos: É possível incluir manualmente regras de segmento em um modelo utilizando a função Inserir Segmento.

Para incluir uma condição da regra de segmento em um modelo:

1. No diálogo Lista interativa, selecione um local onde deseja incluir um novo segmento. O novo segmento será inserido diretamente acima do segmento selecionado.
2. No menu Editar, escolha **Inserir Segmento** ou acesse esta seleção clicando com o botão direito em um segmento.

   O diálogo Inserir Segmento é exibido, permitindo inserir novas condições da regra de segmento.
3. Clique em **Inserir**. O diálogo Inserir Condição é exibido, permitindo definir os atributos para a nova condição de regra.
4. Selecione um campo e um operador nas listas suspensas.

   Nota: se você selecionar o operador **Not in**, a condição selecionada funcionará como uma condição de exclusão e será exibida em vermelho no diálogo Inserir Regra. Por exemplo, quando a condição `region = 'TOWN'` for exibida em vermelho, isso significa que `TOWN` é excluído do conjunto de resultados.
5. Insira um ou mais valores ou clique no ícone **Inserir Valor** para exibir o diálogo Inserir Valor. O diálogo permite escolher um valor definido para o campo selecionado. Por exemplo, o campo **casado** fornecerá os valores **sim** e **não**.
6. Clique em **OK** para retornar ao diálogo Inserir Segmento. Clique em **OK** uma segunda vez para incluir o segmento criado no modelo.

O novo segmento será exibido no local de modelo especificado.

Editando Regras de Segmento: A funcionalidade Editar Regra de Segmento permite incluir, alterar ou excluir condições regra de segmento.

Para alterar condição de regra de segmento:

1. Selecione o segmento modelo que você deseja editar.
2. No menu Editar, escolha **Editar Regra de Segmento** ou clique com o botão direito na regra para acessar essa seleção.

   O diálogo Editar Regra de Segmento é aberta.
3. Selecione a condição adequada e clique em **Editar**.

   O diálogo Editar Condição é exibido, permitindo definir os atributos para a condição da regra selecionada.
4. Selecione um campo e um operador nas listas suspensas.

   Nota: se você selecionar o operador **Not in**, a condição selecionada funcionará como uma condição de exclusão e será exibida em vermelho no diálogo Editar Regra de Segmento. Por exemplo, quando a condição `region = 'TOWN'` for exibida em vermelho, isso significa que `TOWN` é excluído do conjunto de resultados.
5. Insira um ou mais valores ou clique no botão **Inserir Valor** para exibir o diálogo Inserir Valor. O diálogo permite escolher um valor definido para o campo selecionado. Por exemplo, o campo **casado** fornecerá os valores **sim** e **não**.
6. Clique em **OK** para retornar ao diálogo Editar Regra de Segmento. Clique em **OK** uma segunda vez para retornar ao modelo de trabalho.

O segmento selecionado será exibido com as condições de regra atualizadas.

Excluindo Condições de Regra de Segmento: **Para excluir uma condição de regra de segmento:**

1. Selecione o segmento modelo que contém as condições de regra que deseja excluir.
2. No menu Editar, escolha **Editar Regra de Segmento** ou clique com o botão direito no segmento para acessar esta seleção.

   O diálogo Editar Regra de Segmento é exibido, permitindo excluir uma ou mais condições de regra de segmento.
3. Selecione a condição de regra apropriada e clique em **Excluir**.

4. Clique em **OK**.

Excluir uma ou mais condições de regra de segmento faz com que a área de janela do modelo de trabalho atualize suas métricas de medida.

Copiando Segmentos: O Decision List Viewer fornece uma maneira conveniente de copiar segmentos do modelo. Quando quiser aplicar um segmento de um modelo a outro modelo, apenas copie (ou recorte) o segmento a partir de um modelo e cole-o em outro modelo. Também é possível copiar um segmento de um modelo exibido no painel Visualização Alternativa e colá-lo no modelo exibido na área de janela do modelo de trabalho. Essas funções de recortar, copiar e colar utilizam uma área de transferência do sistema para armazenar ou recuperar dados temporários. Isso significa que as condições e o destino são copiados para a área de transferência. O conteúdo da área de transferência não é reservado exclusivamente para uso na Decision List Viewer, mas também pode ser colado em outros aplicativos. Por exemplo, quando o conteúdo da área de transferência for colado em um editor de texto, as condições e o destino são colados em formato XML.

Para copiar ou recortar segmentos modelo:

1. Selecione o segmento modelo que deseja utilizar em outro modelo.
2. No menu Editar, escolha **Copiar** (ou **Recortar**) ou clique com o botão direito no segmento modelo e selecione **Copiar** ou **Recortar**.
3. Abra o modelo apropriado (no qual o segmento modelo será colado).
4. Selecione um dos segmentos modelo e clique em **Colar**.

Nota: ao invés dos comandos **Recortar**, **Copiar** e **Colar**, também é possível utilizar as combinações de teclas: **Ctrl+X**, **Ctrl+C** e **Ctrl+V**.

Um segmento copiado (ou recortado) é inserido acima do segmento modelo selecionado anteriormente. As medidas do segmento colado e dos segmentos abaixo são recalculadas.

Nota: ambos os modelos neste procedimento devem ser baseados no mesmo modelo padrão subjacente e conter o mesmo destino, caso contrário, uma mensagem de erro será exibida.

Modelos Alternativos: Onde houver mais de um resultado, a guia Alternativos exibe os resultados de cada tarefa de mineração. Cada resultado consiste em condições nos dados selecionados que melhor correspondem ao destino, bem como em quaisquer alternativas "boas o suficiente". O número total de alternativas mostradas depende dos critérios de procura usados no processo de análise.

Para visualizar modelos alternativos:

1. Clique em um modelo alternativo na guia Alternativos. Os segmentos modelo alternativo exibem ou substituem os segmentos modelo atuais, no painel Visualização Alternativa.
2. Para trabalhar com um modelo alternativo na área de janela de modelo ativo, selecione o modelo e clique em **Carregar** no painel Visualização Alternativa ou clique com o botão direito em um nome alternativo na guia Alternativos e escolha **Carregar**.

Nota: os modelos alternativos não são salvos ao gerar um novo modelo.

Customizando um Modelo

Os dados não são estáticos. Os clientes se mudam, se casam e mudam de emprego. Os produtos perdem o foco do mercado e se tornam obsoletos.

A Decision List Viewer oferece aos usuários de negócios a flexibilidade de adaptar os modelos às novas situações de modo fácil e rápido. É possível alterar um modelo ao editar, priorizar, excluir ou desativar segmentos modelo específicos.

Priorizando Segmentos: É possível classificar as regras do modelo em qualquer ordem que escolher. Por padrão, os segmentos modelo são exibidos em ordem de prioridade, em que o primeiro segmento tem a prioridade mais alta. Ao designar uma prioridade diferente para um ou mais dos segmentos, o modelo é alterado de acordo. É possível alterar o modelo conforme necessário ao mover segmentos para uma posição de prioridade superior ou inferior.

Para priorizar segmentos modelo:

1. Selecione o segmento modelo ao qual você deseja designar uma prioridade diferente.
2. Clique em um dos dois botões de seta na barra de ferramentas da área de janela do modelo de trabalho para mover o segmento modelo selecionado para cima ou para baixo na lista.

Após a priorização, todos os resultados da avaliação anterior serão recalculados e os novos valores serão exibidos.

Excluindo Segmentos: Para excluir um ou mais segmentos:

1. Selecione um segmento modelo.
2. No menu Editar, escolha **Excluir Segmento**, ou clique no botão excluir na barra de ferramentas da área de janela do modelo de trabalho.

As medidas são recalculadas para o modelo modificado, e o modelo é alterado de acordo.

Excluindo Segmentos: Ao procurar grupos específicos, você provavelmente se baseará em ações de negócios em uma seleção dos segmentos do modelo. Ao implementar um modelo, é possível optar por excluir segmentos em um modelo. Os segmentos excluídos são escorados como valores nulos. Excluir um segmento não significa que o segmento não é utilizado; significa que todos os registros correspondentes a esta regra são excluídos da lista de distribuição. A regra ainda é aplicada, mas de forma diferente.

Para excluir segmentos específicos do modelo:

1. Selecione um segmento na área de janela do modelo de trabalho.
2. Clique no botão **Alternar Exclusão de Segmento** na barra de ferramentas da área de janela do modelo de trabalho. **Excluídos** é agora exibido na coluna Destino selecionada do segmento selecionado.

Nota: ao contrário dos segmentos excluídos, os segmentos removidos permanecem disponíveis para reutilização no modelo final. Os segmentos excluídos afetam os resultados do gráfico.

Alterar Valor de Resposta: O diálogo Alterar Valor de Resposta permite alterar o valor de resposta para o campo de destino atual.

Capturas instantâneas e resultados de sessão com um valor de resposta diferente do Modelo de Trabalho são identificados pela mudança do plano de fundo da tabela dessa linha para amarelo. Isso indica que uma captura instantânea/resultado da sessão está desatualizado.

O diálogo **Criar/Editar Tarefa de Mineração** exibe o valor de resposta para o modelo de trabalho atual. O valor de resposta não é salvo com a tarefa de mineração. Ao invés disso, ele é obtido do valor do Modelo de Trabalho.

Ao promover um modelo salvo para o Modelo de Trabalho que possui um valor de resposta diferente do modelo de trabalho atual (por exemplo, editando um resultado alternativo ou editando uma cópia de uma captura instantânea), o valor de resposta do modelo salvo é alterado para que seja o mesmo do modelo de trabalho (o valor de resposta mostrado na área de janela Modelo de Trabalho não é alterado). As métricas de modelo são reavaliadas com a nova resposta.

Gerar Novo Modelo

O diálogo Gerar Novo Modelo fornece opções para nomear o modelo e selecionar onde o novo nó é criado.

Nome do modelo. Selecione **Customizado** para ajustar o nome gerado automaticamente ou para criar um nome exclusivo para o nó, conforme exibido na tela de fluxo.

Criar nó em. Selecionar **Tela** coloca o novo modelo na tela de trabalho, selecionar **Paleta GM** coloca o novo modelo na paleta Modelos, e selecionar **Ambos** coloca o novo modelo na tela de trabalho e também na paleta Modelos.

Incluir estado de sessão interativa. Quando ativada, o estado da sessão interativa é preservado no modelo gerado. Quando gerar posteriormente um nó de modelagem a partir do modelo, o estado é conduzido e utilizado para inicializar a sessão interativa. Independentemente se a opção for selecionada, o próprio modelo escora os novos dados de forma idêntica. Quando a opção não estiver selecionada, o modelo ainda é capaz de criar um nó de construção, no entanto, ele será um nó de construção mais genérico que inicia uma nova sessão interativa ao invés de obter o local onde a sessão antiga parou. Se você alterar as configurações do nó, mas executar com um estado salvo, as configurações alteradas serão ignoradas a favor das configurações do estado salvo.

Nota: as métricas padrão são as únicas métricas que permanecem com o modelo. As métricas adicionais são preservadas com o estado interativo. O modelo gerado não representa o estado da tarefa de mineração interativa salvo. Ao ativar a Decision List Viewer, ela exibe as configurações originalmente feitas por meio do Visualizador.

Consulte o tópico “Gerando novamente um Nó de Modelagem” na página 49 para obter mais informações.

Avaliação de modelo

Uma modelagem bem-sucedida requer uma avaliação cautelosa do modelo antes que a implementação no ambiente de produção ocorra. O Decision List Viewer fornece diversas medidas de estatísticas e de negócios que podem ser utilizadas para avaliar o impacto de um modelo no mundo real. Estas incluem gráficos de ganhos e uma interoperabilidade total com o Excel, permitindo, assim, que cenários de custo/benefício sejam simulados para avaliar o impacto da implementação.

É possível avaliar seu modelo das seguintes formas:

- Utilizando as medidas de estatísticas e de modelo de negócios predefinidas disponíveis no Decision List Viewer (probabilidade, frequência).
- Avaliando medidas importadas no Microsoft Excel.
- Visualizando o modelo utilizando um gráfico de ganhos.

Organizando Medidas de Modelo: A Decision List Viewer fornece opções para definir as medidas que são calculadas e exibidas como colunas. Cada segmento pode incluir as medidas de cobertura, de frequência, de probabilidade e de erro padrão representadas como colunas. Também é possível criar novas medidas que serão exibidas como colunas.

Definindo Medidas do Modelo

Para incluir uma medida em seu modelo ou para definir uma medida existente:

1. No menu Ferramentas, escolha **Organizar Medidas de Modelo** ou clique com o botão direito no modelo para fazer esta seleção. O diálogo Organizar Medidas de Modelo é aberto.
2. Clique no botão **Incluir nova medida de modelo** (à direita da coluna Mostrar). Uma nova medida é exibida na tabela.
3. Forneça um nome para a medida e selecione um tipo, uma opção de exibição e uma seleção apropriados. A coluna Mostrar indica se a medida será exibida para o modelo de trabalho. Ao definir uma medida existente, selecione uma métrica e uma seleção apropriadas e indique se a medida será exibida para o modelo de trabalho.

4. Clique em **OK** para retornar para a área de trabalho da Decision List Viewer. Se a coluna Mostrar para a nova medida estiver selecionada, a nova medida será exibida para o modelo de trabalho.

Métricas Customizadas no Excel

Consulte o tópico "Avaliação no Excel" para obter mais informações.

Atualizando Medidas: Em determinados casos, pode ser necessário recalcular as medidas de modelo, por exemplo, quando aplicar um modelo existente a um novo conjunto de clientes.

Para recalcular (atualizar) as medidas de modelo:

No menu Editar, escolha **Atualizar Todas as Medidas**.

ou

Pressione F5.

Todas as medidas são recalculadas e os novos valores são mostrados para o modelo de trabalho.

Avaliação no Excel: O Decision List Viewer pode ser integrado ao Microsoft Excel, permitindo utilizar seus próprios cálculos de valores e fórmulas de lucro diretamente no processo de construção de modelo para simular cenários de custo/benefício. A ligação com o Excel permite exportar dados no Excel, que poderá ser utilizado para criar gráficos de apresentação, calcular medidas customizadas, como medidas complexas de lucro e de ROI, e visualizá-las na Decision List Viewer durante a construção de modelo.

Nota: para poder trabalhar com uma planilha do Excel, o especialista de CRM analítico deverá definir informações de configuração para a sincronização da Decision List Viewer com o Microsoft Excel. A configuração está contida em um arquivo de planilha do Excel e indica quais informações são transferidas da Decision List Viewer para o Excel, e vice-versa.

Os passos a seguir são válidos apenas quando o MS Excel estiver instalado. Se o Excel não estiver instalado, as opções para sincronizar os modelos com o Excel não serão exibidas.

Para sincronizar os modelos com o MS Excel:

1. Abra o modelo, execute uma sessão interativa e escolha **Organizar Medidas de Modelo** no menu Ferramentas.
2. Selecione **Sim** para a opção **Calcular medidas customizadas no Excel**. O campo **Planilha** é ativado, permitindo selecionar um modelo de planilha do Excel pré-configurado.
3. Clique no botão **Conectar-se ao Excel**. O diálogo Abrir é aberto, permitindo navegar para o local de modelo pré-configurado em seu sistema de arquivos local ou de rede.
4. Selecione o modelo Excel apropriado e clique em **Abrir**. O modelo do Excel selecionado é ativado; utilize a barra de tarefas do Windows (ou pressione Alt-Tab) para navegar de volta para o diálogo Escolher Entradas para Medidas Customizadas.
5. Selecione os mapeamentos apropriados entre os nomes de métricas definidos no modelo do Excel e os nomes de métrica do modelo e clique em **OK**.

Quando a ligação for estabelecida, o Excel começará com o modelo do Excel pré-configurado que exibe as regras de modelo na planilha. Os resultados calculados no Excel são exibidos como novas colunas na Decision List Viewer.

Nota: as métricas do Excel não permanecem quando o modelo é salvo, elas são válidas apenas durante a sessão ativa. No entanto, é possível criar capturas instantâneas que incluam métricas do Excel. As métricas do Excel salvas nas visualizações de captura instantânea são válidas apenas para fins de comparação histórica e não são atualizadas quando reabertas. Consulte o tópico "Guia Capturas

Instantâneas" na página 151 para obter mais informações. As métricas do Excel não serão exibidas nas capturas instantâneas até que você restabeleça uma conexão com o modelo do Excel.

Instalação da Integração do MS Excel: A integração entre a Decision List Viewer e o Microsoft Excel é realizada por meio do uso de um modelo de planilha do Excel pré-configurado. O modelo consiste em três planilhas:

Medidas de Modelo. Exibe as medidas da Decision List Viewer importadas, as medidas do Excel customizadas e os totais calculados (definidos na planilha Definições).

Definições. Fornece as variáveis para gerar cálculos com base nas medidas da Decision List Viewer importadas e nas medidas do Excel customizadas.

Configuração. Fornece opções para especificar quais medidas serão importadas da Decision List Viewer e para definir as medidas do Excel customizadas.

AVISO: A estrutura da planilha de Configuração é rigidamente definida. **NÃO** edite nenhuma célula na área sombreada em verde.

- **Métricas Do Modelo.** Indica quais métricas do Decision List Viewer são utilizadas nos cálculos.
- **Métricas Para Modelo.** Indica qual métrica ou métricas geradas pelo Excel serão retornadas para a Decision List Viewer. As métricas geradas pelo Excel são exibidas como novas colunas de medida no Decision List Viewer.

Nota: as métricas do Excel não permanecem com o modelo ao gerar um novo modelo, elas são válidas apenas durante a sessão ativa.

Alterando as Medidas do Modelo: Os exemplos a seguir explicam como alterar as Medidas do Modelo de várias maneiras:

- Alterar uma medida existente.
- Importar uma medida padrão adicional a partir do modelo.
- Exportar uma medida customizada adicional no modelo.

Alterar uma medida existente

1. Abra o modelo e selecione a planilha Configuração.
2. Edite qualquer **Nome** ou **Descrição** destacando-os e digitando sobre eles.

   Observe que se desejar alterar uma medida -- por exemplo, para solicitar ao usuário a Probabilidade ao invés de Frequência -- basta alterar o nome e a descrição em **Métricas do Modelo** que, em seguida, serão exibidos no modelo e o usuário poderá escolher a medida apropriada para mapear.

Importar uma medida padrão adicional a partir do modelo

1. Abra o modelo e selecione a planilha Configuração.
2. Nos menus, escolha:

   Ferramentas > **Proteção** > **Desproteger Planilha**
3. Selecione a célula A5 sombreada em amarelo que contém a palavra **Término**.
4. Nos menus, escolha:

   Inserir > **Linhas**
5. Digite o **Nome** e a **Descrição** da nova medida. Por exemplo, **Erro** e **Erro associado ao segmento**.
6. Na célula C5, insira a fórmula **=COLUMN('Model Measures'!N3)**.
7. Na célula D5, insira a fórmula **=ROW('Model Measures'!N3)+1**.

   Essas fórmulas fazem com que a nova medida seja exibida na coluna N da planilha Medidas de Modelo, que está atualmente vazia.

8. Nos menus, escolha:

   Ferramentas > **Proteção** > **Proteger Planilha**
9. Clique em **OK**.
10. Na planilha Medidas do Modelo, assegure-se de que a célula N3 tenha **Erro** como um título da nova coluna.
11. Selecione tudo na coluna N.
12. Nos menus, escolha:

    Formato > **Células**
13. Por padrão, todas as células possuem uma categoria de número **Geral**. Clique em **Porcentagem** para alterar como as figuras são exibidas. Isso ajuda a verificar suas figuras no Excel e também permite utilizar os dados de outras formas, por exemplo, como uma saída para um gráfico.
14. Clique em **OK**.
15. Salve a planilha como um modelo do Excel 2003, com um nome exclusivo e com a extensão de arquivo *.xlt*. Para facilitar a localização do novo modelo, recomendamos salvá-lo no local de modelo pré-configurado em seu sistema de arquivos local ou de rede.

Exportar uma medida customizada adicional no modelo.

1. Abra o modelo para o qual você incluiu a coluna Erro no exemplo anterior e selecione a planilha Configuração.
2. Nos menus, escolha:

   Ferramentas > **Proteção** > **Desproteger Planilha**
3. Selecione a célula A14 sombreada em amarelo que contém a palavra **Término**.
4. Nos menus, escolha:

   Inserir > **Linhas**
5. Digite o **Nome** e a **Descrição** da nova medida. Por exemplo, **Erro Escalado** e **Ajuste de escala aplicado ao erro a partir do Excel**.
6. Na célula C14, insira a fórmula **=COLUMN('Model Measures'!O3)**.
7. Na célula D14, insira a fórmula **=ROW('Model Measures'!O3)+1**.

   Essas fórmulas especificam que a coluna O fornecerá a nova medida para o modelo.
8. Selecione a planilha Configurações.
9. Na célula A17, insira a descrição **'- Erro Escalado**.
10. Na célula B17, insira o fator de ajuste de escala de **10**.
11. Na planilha Medidas de Modelo, insira a descrição **Erro Escalado** na célula O3 como um título para a nova coluna.
12. Na célula O4, insira a fórmula **=N4*Settings!B17**.
13. Selecione o canto da célula O4 e arraste-o para baixo até a célula O22 para copiar a fórmula em cada célula.
14. Nos menus, escolha:

    Ferramentas > **Proteção** > **Proteger Planilha**
15. Clique em **OK**.
16. Salve a planilha como um modelo do Excel 2003, com um nome exclusivo e com a extensão de arquivo *.xlt*. Para facilitar a localização do novo modelo, recomendamos salvá-lo no local de modelo pré-configurado em seu sistema de arquivos local ou de rede.

Ao conectar-se ao Excel utilizando esse modelo, o valor Erro estará disponível como uma nova medida customizada.

Visualizando Modelos

A melhor maneira de entender o impacto de um modelo é visualizá-lo. Ao usar um gráfico de ganhos, é possível obter insight diário de valor dos negócios e benefícios técnicos de seu modelo ao estudar o efeito de diversas alternativas em tempo real. A seção "Gráfico de Ganhos" mostra o benefício de um modelo na tomada de decisão aleatória e permite comparar diretamente diversos gráficos quando houver modelos alternativos.

Gráfico de Ganhos: O gráfico de ganhos representa os valores na coluna *% de Ganhos* da tabela. Os ganhos são definidos como a proporção de ocorrências em cada incremento com relação ao número total de ocorrências na árvore, utilizando a seguinte equação:

(ocorrência em incremento / número total de ocorrências) x 100%

Os gráficos de ganho ilustram eficientemente o quanto você ainda precisa efetuar cast da rede para capturar uma determinada percentagem de todas as ocorrências na árvore. A linha diagonal representa a resposta esperada para a amostra inteira se o modelo não for utilizado. Neste caso, a taxa de resposta seria constante, uma vez que a probabilidade de uma pessoa responder é a mesma que a de outra pessoa. Para dobrar seu lucro, você precisaria perguntar para o dobro de pessoas. A curva da linha indica o quanto é possível melhorar sua resposta, incluindo apenas aquelas que se classificam nos percentuais mais altos com base no ganho. Por exemplo, incluir os 50% principais pode render mais de 70% das respostas positivas. Quanto mais íngreme for a curva, maior será o ganho.

Para visualizar um gráfico de ganhos:

1. Abra um fluxo que contenha um nó Lista de Decisão e ative uma sessão interativa a partir do nó.
2. Clique na guia **Ganhos**. Dependendo de quais partições estiverem especificadas, é possível ver um ou dois gráficos (dois gráficos serão exibidos, por exemplo, quando ambas as partições de treinamento e de teste forem definidas para as medidas de modelo).

Por padrão, os gráficos são exibidos como segmentos. É possível alternar os gráficos para exibir como quantis ao selecionar **Quantis** e, em seguida, selecionar o método quantil apropriado no menu suspenso.

Opções de gráfico: O recurso Opções de Gráfico fornece opções para selecionar quais modelos e capturas instantâneas são representados em gráfico, quais partições também serão representadas e se rótulos de segmento devem ser exibidos ou não.

Modelos para Representação em Gráfico

Modelos Atuais. Permite selecionar quais modelos serão representados em gráfico. É possível selecionar o modelo de trabalho ou quaisquer modelos de captura instantânea criados.

Partições para Representação em Gráfico

Partições para o gráfico esquerdo. A lista suspensa fornece opções para exibir todas as partições definidas ou todos os dados.

Partições para o gráfico direito. A lista suspensa fornece opções para exibir todas as partições definidas, todos os dados ou apenas o gráfico esquerdo. Quando **Apenas gráfico esquerdo** é selecionada, apenas o gráfico esquerdo é exibido.

Exibir Rótulos do Segmento. Quando ativado, cada rótulo de segmento é exibido nos gráficos.

Capítulo 10. Modelos Estatísticos

Os modelos estatísticos utilizam equações matemáticas para codificar as informações extraídas dos dados. Em alguns casos, as técnicas de modelagem estatística podem fornecer modelos adequados muito rapidamente. Mesmo para problemas nos quais técnicas de aprendizado por máquina mais flexíveis (como redes neurais) podem definitivamente fornecer melhores resultados, é possível utilizar alguns modelos estatísticos como modelos preditivos de linha de base para avaliar o desempenho das técnicas mais avançadas.

Os nós de modelagem estatística a seguir estão disponíveis.

Os modelos de regressão lineares preveem uma variável resposta contínua com base em relacionamentos lineares entre o destino e um ou mais preditores.

A regressão logística é uma técnica estatística para classificar registros com base em valores de campos de entrada. Ela é semelhante a uma regressão linear, mas usa um campo de destino categórico ao invés de um intervalo numérico.

O nó PCA/Fator fornece técnicas poderosas de redução de dados para reduzir a complexidade de seus dados. A análise de componentes principais (PCA) localiza combinações lineares dos campos de entrada que executam a melhor tarefa de capturar a variância no conjunto inteiro de campos, em que os componentes são ortogonais (perpendiculares) entre si. A análise fatorial tenta identificar fatores subjacentes que explicam o padrão de correlações dentro de um conjunto de campos observados. Para ambas as abordagens, o objetivo é encontrar um número pequeno de campos derivados que efetivamente sumarizem as informações no conjunto de campos original.

A análise discriminante faz suposições mais rígidas do que a regressão logística, mas pode ser uma alternativa ou um complemento poderoso para uma análise de regressão logística quando essas suposições forem atendidas.

O modelo Linear Generalizado expande o modelo linear geral para que a variável dependente seja linearmente relacionada aos fatores e covariáveis por meio de uma função de ligação especificada. Além disso, o modelo permite à variável dependente ter uma distribuição não normal. Ele cobre a funcionalidade de um grande número de modelos estatísticos, incluindo regressão linear, regressão logística, modelos de log-linear para dados de contagem e modelos de sobrevivência censurados por intervalo.

Um modelo linear generalizado misto (GLMM) estende o modelo linear para que o destino possa ter uma distribuição não normal, esteja linearmente relacionado aos fatores e covariáveis por meio de uma função de ligação especificada e para que as observações possam ser correlacionadas. Os modelos lineares generalizados mistos abrangem uma ampla variedade de modelos, desde regressão linear simples até modelos multiníveis complexos para dados de longitude não normais.

O nó Regressão de Cox permite construir um modelo de sobrevivência para dados de sobrevivência na presença de registros censurados. O modelo produz uma função de sobrevivência que prediz a probabilidade de que o evento de interesse tenha ocorrido em um determinado momento (*t*) de acordo com os valores fornecidos para as variáveis de entrada.

Nó Lineares

A regressão linear é uma técnica estatística comum para classificar registros com base nos valores de campos de entrada numéricos. A regressão linear ajusta uma linha ou uma superfície reta que minimiza as discrepâncias entre os valores de saída preditos e reais.

Requisitos. Apenas campos numéricos podem ser utilizados em um modelo de regressão linear. Deve-se ter exatamente um campo de destino (com o papel configurado como **Destino**) e um ou mais preditores (com o papel configurado como **Entrada**). Os campos com um papel de **Ambos** ou **Nenhum** são ignorados, já que eles são campos não numéricos. (Se necessário, campos não numéricos podem ser recodificaos utilizando um nó Derivar).

Intensidades. Os modelos de regressão linear são relativamente simples e fornecem uma fórmula matemática facilmente interpretada para gerar predições. Como a regressão linear é um procedimento estatístico consagrado, as propriedades desses modelos são bem entendidas. Em geral, os modelos lineares também são muito rápidos para treinar. O nó Linear fornece métodos para seleção automática de campo para eliminar campos de entrada não significativos da equação.

Nota: Nos casos em que o campo de destino é categórico ao invés de um intervalo contínuo, como **yes/no** ou **churn/don't churn**, a regressão logística poderá ser utilizada como uma alternativa. A regressão logística também fornece suporte para entradas não numéricas, removendo a necessidade de recodificação destes campos. Consulte o tópico “Nó de Logística” na página 176 para obter mais informações.

Modelos lineares

Os modelos lineares preveem uma variável resposta contínua com base em relacionamentos lineares entre o destino e um ou mais preditores.

Os modelos lineares são relativamente simples e fornecem uma fórmula matemática facilmente interpretada para escoragem. As propriedades desses modelos são bem entendidas e geralmente podem ser construídas muito rapidamente em comparação com outros tipos de modelo (como redes neurais ou árvores de decisão) no mesmo conjunto de dados.

Exemplo. Uma empresa de seguros com recursos limitados para investigar reclamações de seguro residencial quer construir um modelo para estimar os custos da reclamação. Ao implementar esse modelo nos centros de serviço, os representantes poderão inserir informações da reclamação enquanto estiver no telefone com um cliente e obter imediatamente o custo "esperado" da reclamação com base nos dados passados.

Requisitos de campo. Deve haver um Destino e pelo menos uma Entrada. Por padrão, os campos com papéis predefinidos de Ambos ou Nenhum não são utilizados. O destino deve ser contínuo (escala). Não há restrições de nível de medição nos preditores (entradas); os campos categóricos (flag, nominal e ordinal) são utilizados como fatores no modelo e os campos contínuos são utilizados como covariáveis.

Objetivos

O que deseja fazer?

- **Construir um novo modelo.** Constrói um modelo completamente novo. Esta é a operação usual do nó.

- **Continuar treinamento de um modelo existente.** O treinamento continua com o último modelo produzido com sucesso pelo nó. Isso permite atualizar ou renovar um modelo existente sem precisar acessar os dados originais e poderá resultar em um desempenho significativamente mais rápido desde que apenas os registros novos ou atualizados sejam alimentados no fluxo. Detalhes do modelo anterior são armazenados com o nó de modelagem, o que permite utilizar essa opção mesmo se o nugget do modelo anterior não estiver mais disponível na paleta de fluxo ou de Modelos.

  Nota: quando essa opção é ativada, todos os outros controles nas guias Campos e Opções de Criação são desativados.

Qual é o seu principal objetivo? Selecione o objetivo apropriado.

- **Criar um modelo padrão.** O método constrói um único modelo para prever a resposta utilizando os preditores. Em geral, os modelos padrão são mais fáceis de interpretar e podem ser mais rápidos para escorar do que combinações de conjunto de dados impulsionados, empacotados ou grandes.

  Nota: Para modelos de divisão, para utilizar esta opção com **Continuar treinando um modelo existente**, você deverá estar conectado ao Analytic Server.
- **Aprimorar a precisão do modelo (boosting).** O método constrói um modelo de combinação usando boosting, que gera uma sequência de modelos para obter predições mais precisas. As combinações podem demorar mais tempo para construir e escorar do que um modelo padrão.

  O boosting produz uma sucessão de "modelos de componentes", cada qual construído no conjunto de dados inteiro. Antes de construir cada modelo de componente sucessivo, os registros são ponderados com base nos resíduos do modelo do componente anteriores. Casos com resíduos grandes recebem ponderações de análise relativamente maiores para que o próximo modelo de componente foque na predição também desses registros. Juntos, esses modelos de componente formam um modelo de combinação. O modelo de combinação escora novos registros utilizando uma regra de combinação, e as regras disponíveis dependem do nível de medição da resposta.
- **Aprimorar a estabilidade do modelo (bagging).** O método constrói um modelo de combinação usando bagging (agregação de bootstrap), que gera diversos modelos para obter predições mais confiáveis. As combinações podem demorar mais tempo para construir e escorar do que um modelo padrão.

  A agregação de bootstrap (bagging) produz réplicas do conjunto de dados de treinamento por amostragem com substituição do conjunto de dados original. Isso cria amostras de bootstrap de tamanho igual ao do conjunto de dados original. Em seguida, um "modelo de componente" é construído em cada réplica. Juntos, esses modelos de componente formam um modelo de combinação. O modelo de combinação escora novos registros utilizando uma regra de combinação, e as regras disponíveis dependem do nível de medição da resposta.
- **Criar um modelo para conjuntos de dados muito grandes (requer o IBM SPSS Modeler Server).** O método constrói um modelo de combinação ao dividir o conjunto de dados em blocos de dados separados. Escolha essa opção quando o conjunto de dados for grande demais para construir qualquer um dos modelos acima, ou para construção de modelo incremental. Essa opção pode levar menos tempo para construir, mas pode levar mais tempo para escorar com relação a um modelo padrão. Esta opção requer conectividade com o IBM SPSS Modeler Server .

Consulte “Combinações” na página 169 para obter configurações relacionadas ao boosting, bagging e conjuntos de dados muito grandes.

Básicos

Preparar dados automaticamente. Essa opção ativa o procedimento para transformar internamente a resposta e os preditores para maximizar o poder preditivo do modelo; todas as transformações são salvas com o modelo e aplicadas aos novos dados para escoragem. As versões originais dos campos transformados serão excluídas do modelo. Por padrão, a preparação automática de dados a seguir é executada.

- **Tratamento de Data e Hora.** Cada preditor de data é transformado em um novo preditor contínuo contendo o tempo decorrido desde uma data de referência (1970-01-01). Cada preditor de tempo é transformado em um novo preditor contínuo contendo o tempo decorrido desde o horário de referência (00:00:00).
- **Ajustar o nível de medição.** Preditores contínuos com menos de 5 valores distintos são reformulados como preditores ordinais. Os preditores ordinais com mais de 10 valores distintos são reformulados como preditores contínuos.
- **Tratamento de valor discrepante.** Valores de preditores contínuos que estiverem além de um valor de corte (3 desvios padrão da média) são configurados para o valor de corte.
- **Tratamento de valor omisso.** Valores omissos de preditores nominais são substituídos pelo modo da partição de treinamento. Valores omissos de preditores ordinais são substituídos pela mediana da partição de treinamento. Valores omissos de preditores contínuos são substituídos pela média da partição de treinamento.
- **Mesclagem supervisionada.** Isso torna um modelo mais simples ao reduzir o número de campos a serem processados em associação ao destino. Categorias similares são identificadas com base no relacionamento entre a entrada e o destino. As categorias que não forem significativamente diferentes (ou seja, que possuem um valor p maior que 0,1) são mescladas. Se todas as categorias forem mescladas em uma só, as versões original e derivada do campo serão excluídas do modelo porque elas não possuem nenhum valor como preditor.

Nível de confiança. Este é o nível de confiança utilizado para calcular estimativas de intervalo dos coeficientes do modelo na visualização Coeficientes. Especifique um valor maior que 0 e menor que 100. O padrão é 95.

Seleção de Modelo

Método de seleção de modelo. Escolha um dos métodos de seleção de modelo (detalhes abaixo) ou **Incluir todos os preditores**, que simplesmente insere todos os preditores disponíveis como termos modelo de efeitos principais. Por padrão, o **Forward stepwise** é utilizado.

Seleção de Forward Stepwise. Isso é iniciado sem efeitos no modelo e inclui e remove os efeitos um passo por vez até que mais nenhum possa ser incluído ou removido de acordo com os critérios de stepwise.

- **Critérios de entrada/remoção.** Essa é a estatística utilizada para determinar se um efeito deve ser incluído ou removido do modelo. O **Critério de Informações (AICC)** baseia-se na probabilidade do conjunto de treinamento dado o modelo, e é ajustado para penalizar modelos excessivamente complexos. As **Estatísticas de F** baseiam-se em um teste estatístico da melhoria no erro do modelo. O **R-quadrado ajustado** baseia-se no ajuste do conjunto de treinamento, e é ajustado para penalizar modelos excessivamente complexos. O **Critério de Prevenção ao Super Ajuste (ASE)** baseia-se no ajuste (erro quadrático médio, ou ASE) do conjunto de prevenção ao super ajuste. O conjunto de prevenção ao super ajuste é uma subamostra aleatória de aproximadamente 30% do conjunto de dados original que não é utilizado para treinar o modelo.

  Se qualquer critério diferente de **Estatísticas de F** for escolhido, então, em cada passo, o efeito que corresponder ao maior aumento positivo no critério será incluído no modelo. Quaisquer efeitos no modelo que corresponderem a uma diminuição no critério serão removidos.

  Se **Estatísticas de F** forem escolhidas como o critério, então, em cada passo, o efeito que tiver um valor de *p* menor que o limite especificado em **Incluir efeitos com valores de p menores que** será incluído no modelo. O padrão é 0,05. Quaisquer efeitos no modelo com um valor de *p* maior que o limite especificado em **Remover efeitos com valores de p maiores que** serão removidos. O padrão é 0,10.
- **Customizar o número máximo de efeitos no modelo final.** Por padrão, todos os efeitos disponíveis podem ser inseridos no modelo. Como alternativa, se o algoritmo stepwise terminar um passo com o número máximo de efeitos especificado, o algoritmo parará com o conjunto atual de efeitos.
- **Customizar número máximo de passos.** O algoritmo stepwise para após um determinado número de passos. Por padrão, isso é 3 vezes o número de efeitos disponíveis. Como alternativa, especifique um número inteiro máximo positivo de passos.

Seleção dos Melhores Subconjuntos. Isso verifica "todos os modelos possíveis", ou pelo menos um subconjunto maior de modelos possíveis do que o forward stepwise, para escolher o melhor modelo de acordo com o critério de melhores subconjuntos. O **Critério de Informações (AICC)** baseia-se na probabilidade do conjunto de treinamento dado o modelo, e é ajustado para penalizar modelos excessivamente complexos. O **R-quadrado ajustado** baseia-se no ajuste do conjunto de treinamento, e é ajustado para penalizar modelos excessivamente complexos. O **Critério de Prevenção ao Super Ajuste (ASE)** baseia-se no ajuste (erro quadrático médio, ou ASE) do conjunto de prevenção ao super ajuste. O conjunto de prevenção ao super ajuste é uma subamostra aleatória de aproximadamente 30% do conjunto de dados original que não é utilizado para treinar o modelo.

O modelo com o maior valor do critério é escolhido como o melhor modelo.

Nota: A seleção dos melhores subconjuntos requer cálculo computacional intensivo maior do que a seleção forward stepwise. Quando os melhores subconjuntos são executados em conjunto com boosting, bagging ou conjuntos de dados muito grandes, o tempo de construção será maior do que um modelo padrão construído utilizando a seleção forward stepwise.

Combinações

Essas configurações determinam o comportamento da combinação que ocorre quando efetuar boosting, bagging ou quando conjuntos de dados muito grandes forem solicitados nos Objetivos. As opções que não se aplicarem ao objetivo selecionado são ignoradas.

Bagging e Conjuntos de Dados Muito Grandes. Ao escorar uma combinação, essa é a regra utilizada para combinar os valores preditos a partir dos modelos base para calcular o valor de escore de combinação.

- **Regra de combinação padrão para variáveis resposta contínuas.** Os valores preditos de combinação para variáveis resposta contínuas podem ser combinados utilizando a média ou mediana dos valores preditos a partir dos modelos base.

Observe que quando o objetivo é melhorar a precisão do modelo, as seleções da regra de combinação são ignoradas. O boosting sempre utiliza uma votação por maioria ponderada para escorar variáveis resposta categóricas e uma média ponderada para escorar variáveis resposta contínuas.

Boosting e Bagging. Especifique o número de modelos base para construção quando o objetivo for melhorar a precisão ou a estabilidade do modelo; para bagging, este é o número de amostras de bootstrap. Ele deve ser um número inteiro positivo.

Avançado

Replicar resultados. Configurar uma semente aleatória permite replicar análises. O gerador de número aleatório é utilizado para escolher quais registros estão no conjunto de prevenção ao super ajuste. Especifique um número inteiro ou clique em **Gerar**, que criará um pseudonúmero inteiro aleatório entre 1 e 2147483647, inclusive. O padrão é 54752075.

Opções de Modelo

Nome do Modelo. É possível gerar o nome do modelo automaticamente com base nos campos de destino ou especificar um nome customizado. O nome gerado automaticamente é o nome do campo de destino.

Observe que o valor predito é calculado sempre que o modelo é escorado. O nome do novo campo é o nome do campo de destino, prefixado com *$L-*. Por exemplo, para um campo de destino chamado *sales*, o novo campo se chamará *$L-sales*.

Sumarização do Modelo

A visualização Sumarização do Modelo é uma captura instantânea ou uma sumarização rápida do modelo e de seu ajuste.

Tabela. A tabela identifica algumas configurações do modelo de alto nível, incluindo:

- O nome do destino especificado no guia Campos,
- Se a preparação automática de dados foi executada conforme especificado nas configurações de Básicos,
- O método e o critério de seleção de modelo especificados nas configurações de Seleção de Modelo. O valor do critério de seleção para o modelo final também é exibido, apresentado em um formato menor e melhor.

Gráfico. O gráfico exibe a precisão do modelo final, apresentada em um formato maior e melhor. O valor é 100 × o R^2 ajustado para o modelo final.

Preparação Automática de Dados

Essa visualização mostra informações sobre quais campos foram excluídos e como os campos transformados foram derivados no passo de preparação automática de dados (ADP). Para cada campo que foi transformado ou excluído, a tabela lista o nome do campo, o seu papel na análise e a ação executada pelo passo ADP. Os campos são classificados em ordem alfabética crescente de nomes de campo. Outras ações possíveis executadas para cada campo incluem:

- **Duração da derivação: meses** calcula o tempo decorrido em meses a partir dos valores em um campo que contém datas até a data atual do sistema.
- **Duração da derivação: horas** calcula o tempo decorrido em horas a partir dos valores em um campo que contém tempos até o tempo atual do sistema.
- **Alterar o nível de medição de contínuo para ordinal** reformula os campos contínuos com menos de 5 valores exclusivos como campos ordinais.
- **Alterar o nível de medição de ordinal para contínuo** reformula os campos ordinais com mais de 10 valores exclusivos como campos contínuos.
- **Aparar valores discrepantes** configura valores de preditores contínuos que estiverem além de um valor de corte (3 desvios padrão da média) com relação ao valor de corte.
- **Substituir valores omissos** substitui os valores omissos de campos nominais pelo modo, dos campos ordinais pela mediana e dos campos contínuos pela média.
- **Mesclar categorias para maximizar a associação com o destino** identifica categorias do preditor "semelhantes" com base no relacionamento entre a entrada e o destino. As categorias que não forem significativamente diferentes (ou seja, que possuem um valor de *p* maior que 0,05) são mescladas.
- **Excluir preditor constante / após manipulação de valor discrepante / após mesclagem de categorias** remove preditores que possuírem um valor único, possivelmente após outras ações ADP terem sido tomadas.

Importância do Preditor

Geralmente você desejará focar seus esforços de modelagem nos campos preditores que forem mais importantes e considerar descartar ou ignorar aqueles que forem menos importantes. O gráfico de importância do preditor ajuda a fazer isso ao indicar a importância relativa de cada preditor na estimativa do modelo. Como os valores são relativos, a soma dos valores para todos os preditores na tela é 1,0. A importância do preditor não tem relação com a precisão do modelo. Ela está relacionada apenas com a importância de cada preditor em fazer uma predição, e não se a predição é precisa ou não.

Predito Por Observado

Isso exibe um gráfico de dispersão categorizado dos valores preditos no eixo vertical em função dos valores observados no eixo horizontal. Idealmente, os pontos devem estar em uma linha de 45 graus, e essa visualização poderá informar se algum registro foi particularmente mal predito pelo modelo.

Resíduos

Isso exibe um gráfico de diagnóstico dos resíduos do modelo.

Estilos de gráfico. Há diferentes estilos de exibição que podem ser acessados a partir da lista suspensa **Estilo**.

- **Histograma.** Esse é um histograma categorizado dos resíduos estudentizados com uma sobreposição da distribuição normal. Os modelos lineares supõem que os resíduos possuem uma distribuição normal, de modo o histograma deverá se aproximar idealmente da linha plana.
- **Gráfico P-P.** Esse é um gráfico categorizado de probabilidade-probabilidade que compara os resíduos estudentizados com uma distribuição normal. Se a inclinação dos pontos representados for menos íngreme que a linha normal, os resíduos mostram uma variabilidade maior do que uma distribuição normal; se a inclinação for mais íngreme, os resíduos mostram uma variabilidade menor do que uma distribuição normal. Se os pontos representados tiverem uma curva em forma de S, então a distribuição dos resíduos está defasada.

Valores discrepantes

Esta tabela lista os registros que exercem influência indevida no modelo e exibe o ID do registro (se especificado na guia Campos), o valor de destino e a distância de Cook. A distância de Cook é uma medida do quanto os resíduos de todos os registros alterariam se um registro específico fosse excluído do cálculo dos coeficientes do modelo. Uma distância de Cook grande indica que excluir um registro altera os coeficientes substancialmente e, portanto, deve ser considerado influente.

Os registros influentes devem ser examinados cuidadosamente para determinar se é possível atribuir a eles uma ponderação menor na estimativa do modelo, truncar os valores distantes para algum limite aceitável ou remover os registros influentes completamente.

Efeitos

Esta visualização exibe o tamanho de cada efeito no modelo.

Estilos. Há diferentes estilos de exibição que podem ser acessados a partir da lista suspensa **Estilo**.

- **Diagrama.** Esse é um gráfico no qual os efeitos são ordenados de cima para baixo, por importância do preditor decrescente. As linhas de conexão no diagrama são ponderadas com base na significância do efeito, com a largura de linha maior correspondendo aos efeitos mais significativos (valores de *p* menores). Passar o mouse sobre a linha de conexão revela uma dica de ferramenta que mostra o valor de *p* e a importância do efeito. Este é o padrão.
- **Tabela.** Esta é uma tabela ANOVA para efeitos do modelo geral e de modelo individual. Os efeitos individuais são ordenados de cima para baixo, por importância do preditor decrescente. Observe que, por padrão, a tabela é reduzida para mostrar apenas os resultados para o modelo geral. Para ver os resultados dos efeitos do modelo individual, clique na célula **Modelo Corrigido** na tabela.

Importância do preditor. Há uma régua de controle de Importância do Preditor que controla quais preditores são mostrados na visualização. Isso não altera o modelo, apenas permite focar nos preditores mais importantes. Por padrão, os 10 principais efeitos são exibidos.

Significância. Há uma régua de controle Significância que controla ainda mais quais efeitos são mostrados na visualização, além daqueles mostrados com base na importância do preditor. Os efeitos com valores de significância maiores que o valor da régua de controle são ocultados. Isso não altera o modelo, apenas permite focar nos efeitos mais importantes. Por padrão, o valor é 1,00, de modo que nenhum efeito seja filtrado com base na significância.

Coeficientes

Essa visualização exibe o valor de cada coeficiente no modelo. Observe que os fatores (preditores categóricos) são codificados por indicador no modelo, de modo que os **efeitos** que contiverem fatores geralmente possuirão diversos **coeficientes** associados, um para cada categoria, exceto para a categoria correspondente ao parâmetro redundante (referência).

Estilos. Há diferentes estilos de exibição que podem ser acessados a partir da lista suspensa **Estilo**.

- **Diagrama.** Esse é um gráfico que exibe o intercepto primeiro e, em seguida, classifica os efeitos de cima para baixo, pela importância do preditor decrescente. Nos efeitos que contêm fatores, os coeficientes são classificados em ordem crescente de valores de dados. As linhas de conexão no

diagrama são coloridas com base no sinal do coeficiente (consulte a chave de diagrama) e ponderadas com base na significância do coeficiente, com a largura de linha maior correspondendo aos coeficientes mais significativos (valores de *p* menores). Passar o mouse sobre a linha de conexão revela uma dica de ferramenta que mostra o valor do coeficiente, seu valor de *p* e a importância do efeito com o qual o parâmetro está associado. Este é o estilo padrão.

- **Tabela.** Mostra os valores, os testes de significância e os intervalos de confiança para coeficientes de modelo individuais. Após o intercepto, os efeitos são ordenados de cima para baixo por importância do preditor decrescente. Nos efeitos que contêm fatores, os coeficientes são classificados em ordem crescente de valores de dados. Observe que, por padrão, a tabela é reduzida para mostrar apenas o coeficiente, a significância e a importância de cada parâmetro de modelo. Para ver o erro padrão, a estatística *t* e o intervalo de confiança, clique na célula **Coeficiente** na tabela. Passar o mouse sobre o nome de um parâmetro de modelo na tabela revela uma dica de ferramenta que mostra o nome do parâmetro, o efeito com o qual o parâmetro está associado e (para preditores categóricos), os rótulos de valor associados ao parâmetro de modelo. Isso pode ser particularmente útil para ver as novas categorias criadas quando a preparação automática de dados mescla categorias similares de um preditor categórico.

Importância do preditor. Há uma régua de controle de Importância do Preditor que controla quais preditores são mostrados na visualização. Isso não altera o modelo, apenas permite focar nos preditores mais importantes. Por padrão, os 10 principais efeitos são exibidos.

Significância. Há uma régua de controle Significância que controla ainda mais quais coeficientes são mostrados na visualização, além daqueles mostrados com base na importância do preditor. Os coeficientes com valores de significância maiores que o valor da régua de controle são ocultados. Isso não altera o modelo, apenas permite focar nos coeficientes mais importantes. Por padrão, o valor é 1,00, de modo que nenhum coeficiente seja filtrado com base na significância.

Médias Estimadas

Esses gráficos são exibidos para preditores significativos. O gráfico exibe o valor estimado pelo modelo da resposta no eixo vertical para cada valor do preditor no eixo horizontal, mantendo todos os outros preditores constantes. Ele fornece uma visualização útil dos efeitos dos coeficientes de cada preditor na resposta.

Nota: se nenhum preditor for significativo, nenhuma média estimada será produzida.

Sumarização de Construção de Modelo

Quando um algoritmo de seleção de modelo diferente de **Nenhum** é escolhido nas configurações de Seleção de Modelo, isso fornece alguns detalhes do processo de construção do modelo.

Forward stepwise. Quando forward stepwise é o algoritmo de seleção, a tabela exibe os últimos 10 passos no algoritmo stepwise. Para cada passo, o valor do critério de seleção e os efeitos no modelo nesse passo serão mostrados. Isso dará uma ideia do quanto cada passo contribui para com o modelo. Cada coluna permite ordenar as linhas para poder ver mais facilmente quais efeitos estão no modelo em um determinado passo.

Melhores subconjuntos. Quando melhores subconjuntos é o algoritmo de seleção, a tabela exibe os 10 principais modelos. Para cada modelo, o valor do critério de seleção e os efeitos no modelo são mostrados. Isso dará uma ideia da estabilidade dos modelos principais; se eles forem propensos a ter muitos efeitos semelhantes com algumas diferenças, então você poderá confiar razoavelmente no modelo "principal"; se eles forem propensos a ter efeitos muito diferentes, então alguns dos efeitos podem ser muito semelhantes e deverão ser combinados (ou algum deles deverá ser removido). Cada coluna permite ordenar as linhas para poder ver mais facilmente quais efeitos estão no modelo em um determinado passo.

Configurações

Observe que o valor predito é calculado sempre que o modelo é escorado. O nome do novo campo é o nome do campo de destino, prefixado com *$L-*. Por exemplo, para um campo de destino chamado *sales*, o novo campo se chamará *$L-sales*.

Gerar SQL para este modelo Ao usar dados de um banco de dados, código SQL pode ser enviado por push de volta para o banco de dados para execução, fornecendo desempenho superior para muitas operações.

Selecione uma das opções a seguir para especificar como a geração de SQL é executada.

- **Padrão: Escorar usando o Server Scoring Adapter (se instalado) no processo** Se conectado a um banco de dados com um adaptador de escoragem instalado, gera a SQL utilizando o adaptador de escoragem e funções definidas pelo usuário (UDF) associadas e escora seu modelo no banco de dados. Quando nenhum adaptador de escoragem estiver disponível, essa opção busca seus dados novamente a partir do banco de dados e os escora no SPSS Modeler.
- **Escorar ao converter para SQL nativo** Se selecionada, gera SQL nativo para escorar o modelo no banco de dados.

  Nota: Embora essa opção possa fornecer resultados mais rápidos, o tamanho e a complexidade do SQL nativo aumentam conforme a complexidade do modelo aumenta.
- **Escorar fora do Banco de dados** Se selecionada, esta opção busca seus dados novamente a partir de seu banco de dados e os escora no SPSS Modeler.

Nó Linear do AS

O IBM SPSS Modeler possui duas versões diferentes do nó Linear:

- **Linear** é o nó tradicional que é executado no IBM SPSS Modeler Server.
- **Linear do AS** é executado apenas quando conectado ao IBM SPSS Analytic Server.

A regressão linear é uma técnica estatística comum para classificar registros com base nos valores de campos de entrada numéricos. A regressão linear ajusta uma linha ou uma superfície reta que minimiza as discrepâncias entre os valores de saída preditos e reais.

Requisitos. Apenas campos numéricos e preditores categóricos podem ser utilizados em um modelo de regressão linear. Deve-se ter exatamente um campo de destino (com o papel configurado como **Destino**) e um ou mais preditores (com o papel configurado como **Entrada**). Os campos com um papel de **Ambos** ou **Nenhum** são ignorados, já que eles são campos não numéricos. (Se necessário, campos não numéricos podem ser recodificaos utilizando um nó Derivar).

Intensidades. Os modelos de regressão linear são relativamente simples e fornecem uma fórmula matemática facilmente interpretada para gerar predições. Como a regressão linear é um procedimento estatístico consagrado, as propriedades desses modelos são bem entendidas. Em geral, os modelos lineares também são muito rápidos para treinar. O nó Linear fornece métodos para seleção automática de campo para eliminar campos de entrada não significativos da equação.

Nota: Nos casos em que o campo de destino é categórico ao invés de um intervalo contínuo, como **yes/no** ou **churn/don't churn**, a regressão logística poderá ser utilizada como uma alternativa. A regressão logística também fornece suporte para entradas não numéricas, removendo a necessidade de recodificação destes campos. Consulte o tópico “Nó de Logística” na página 176 para obter mais informações.

Modelos Lineares do AS

Os modelos lineares preveem uma variável resposta contínua com base em relacionamentos lineares entre o destino e um ou mais preditores.

Os modelos lineares são relativamente simples e fornecem uma fórmula matemática facilmente interpretada para escoragem. As propriedades desses modelos são bem entendidas e geralmente podem ser construídas muito rapidamente em comparação com outros tipos de modelo (como redes neurais ou árvores de decisão) no mesmo conjunto de dados.

Exemplo. Uma empresa de seguros com recursos limitados para investigar reclamações de seguro residencial quer construir um modelo para estimar os custos da reclamação. Ao implementar esse modelo nos centros de serviço, os representantes poderão inserir informações da reclamação enquanto estiver no telefone com um cliente e obter imediatamente o custo "esperado" da reclamação com base nos dados passados.

Requisitos de campo. Deve haver um Destino e pelo menos uma Entrada. Por padrão, os campos com papéis predefinidos de Ambos ou Nenhum não são utilizados. O destino deve ser contínuo (escala). Não há restrições de nível de medição nos preditores (entradas); os campos categóricos (flag, nominal e ordinal) são utilizados como fatores no modelo e os campos contínuos são utilizados como covariáveis.

Básicos

Incluir intercepto. Essa opção inclui um offset no eixo y quando o eixo x é 0. O intercepto geralmente é incluído no modelo. Se você conseguir presumir a passagem de dados por meio da origem, será possível excluir o intercepto.

Considerar interação de duas vias. Essa opção instrui o modelo a comparar cada par de entradas possível para ver se a tendência de uma afeta a outra. Se for o caso, então é mais provável que essas entradas sejam incluídas na matriz de design.

Intervalo de confiança para estimativa de coeficiente (%). Este é o intervalo de confiança utilizado para calcular estimativas dos coeficientes do modelo na visualização Coeficientes. Especifique um valor maior que 0 e menor que 100. O padrão é 95.

Ordenação para preditores categóricos. Esses controles determinam a ordem das categorias para os fatores (entradas categóricas) a fim de determinar a "última" categoria. A configuração da ordenação será ignorada se a entrada não for categórica ou se uma categoria de referência customizada for especificada.

Seleção de Modelo

Método de seleção de modelo. Escolha um dos métodos de seleção de modelo (detalhes abaixo) ou **Incluir todos os preditores**, que simplesmente insere todos os preditores disponíveis como termos modelo de efeitos principais. Por padrão, o **Forward stepwise** é utilizado.

Seleção de Forward Stepwise. Isso é iniciado sem efeitos no modelo e inclui e remove os efeitos um passo por vez até que mais nenhum possa ser incluído ou removido de acordo com os critérios de stepwise.

- **Critérios de entrada/remoção.** Essa é a estatística utilizada para determinar se um efeito deve ser incluído ou removido do modelo. O **Critério de Informações (AICC)** baseia-se na probabilidade do conjunto de treinamento dado o modelo, e é ajustado para penalizar modelos excessivamente complexos. As **Estatísticas de F** baseiam-se em um teste estatístico da melhoria no erro do modelo. O **R-quadrado ajustado** baseia-se no ajuste do conjunto de treinamento, e é ajustado para penalizar modelos excessivamente complexos. O **Critério de Prevenção ao Super Ajuste (ASE)** baseia-se no ajuste (erro quadrático médio, ou ASE) do conjunto de prevenção ao super ajuste. O conjunto de prevenção ao super ajuste é uma subamostra aleatória de aproximadamente 30% do conjunto de dados original que não é utilizado para treinar o modelo.

  Se qualquer critério diferente de **Estatísticas de F** for escolhido, então, em cada passo, o efeito que corresponder ao maior aumento positivo no critério será incluído no modelo. Quaisquer efeitos no modelo que corresponderem a uma diminuição no critério serão removidos.

  Se **Estatísticas de F** forem escolhidas como o critério, então, em cada passo, o efeito que tiver um valor de *p* menor que o limite especificado em **Incluir efeitos com valores de p menores que** será incluído

no modelo. O padrão é 0,05. Quaisquer efeitos no modelo com um valor de *p* maior que o limite especificado em **Remover efeitos com valores de p maiores que** serão removidos. O padrão é 0,10.

- **Customizar o número máximo de efeitos no modelo final.** Por padrão, todos os efeitos disponíveis podem ser inseridos no modelo. Como alternativa, se o algoritmo stepwise terminar um passo com o número máximo de efeitos especificado, o algoritmo parará com o conjunto atual de efeitos.
- **Customizar número máximo de passos.** O algoritmo stepwise para após um determinado número de passos. Por padrão, isso é 3 vezes o número de efeitos disponíveis. Como alternativa, especifique um número inteiro máximo positivo de passos.

Seleção dos Melhores Subconjuntos. Isso verifica "todos os modelos possíveis", ou pelo menos um subconjunto maior de modelos possíveis do que o forward stepwise, para escolher o melhor modelo de acordo com o critério de melhores subconjuntos. O **Critério de Informações (AICC)** baseia-se na probabilidade do conjunto de treinamento dado o modelo, e é ajustado para penalizar modelos excessivamente complexos. O **R-quadrado ajustado** baseia-se no ajuste do conjunto de treinamento, e é ajustado para penalizar modelos excessivamente complexos. O **Critério de Prevenção ao Super Ajuste (ASE)** baseia-se no ajuste (erro quadrático médio, ou ASE) do conjunto de prevenção ao super ajuste. O conjunto de prevenção ao super ajuste é uma subamostra aleatória de aproximadamente 30% do conjunto de dados original que não é utilizado para treinar o modelo.

O modelo com o maior valor do critério é escolhido como o melhor modelo.

Nota: A seleção dos melhores subconjuntos requer cálculo computacional intensivo maior do que a seleção forward stepwise. Quando os melhores subconjuntos são executados em conjunto com boosting, bagging ou conjuntos de dados muito grandes, o tempo de construção será maior do que um modelo padrão construído utilizando a seleção forward stepwise.

Opções de Modelo

Nome do Modelo. É possível gerar o nome do modelo automaticamente com base nos campos de destino ou especificar um nome customizado. O nome gerado automaticamente é o nome do campo de destino.

Observe que o valor predito é calculado sempre que o modelo é escorado. O nome do novo campo é o nome do campo de destino, prefixado com *$L-*. Por exemplo, para um campo de destino chamado *sales*, o novo campo se chamará *$L-sales*.

Resultado interativo

Após executar um modelo Linear do AS, a seguinte saída está disponível.

Informações do modelo

A visualização Informações do Modelo fornece informações chave sobre o modelo. A tabela identifica algumas configurações de modelo de alto nível, como:

- O nome do destino especificado na guia Campos
- O campo de ponderação de regressão
- O método de construção de modelo especificado nas configurações de Seleção de Modelo.
- O número de entrada preditores
- O número de preditores no modelo final
- Akaike Information Criterion Corrected (AICC). O AICC é uma medida para selecionar e comparar modelos combinados com base no log da verossimilhança -2 (Restrito). Valores menores indicam melhores modelos. O AICC "corrige" o AIC para tamanhos de amostra pequenos. Conforme o tamanho da amostra aumenta, o AICC converge para o AIC.
- R-quadrado. Esta é a medida de qualidade de ajuste de um modelo linear, às vezes chamada de coeficiente de determinação. É a proporção de variação na variável dependente explicada pelo modelo de regressão. Ela varia de 0 a 1. Valores pequenos indicam que o modelo não ajusta bem os dados.
- R-quadrado ajustado

Sumarização de registros

A visualização Sumarização de Registros fornece informações sobre o número e a porcentagem de registros (casos) incluídos e excluídos do modelo.

Importância do preditor

Geralmente você desejará focar seus esforços de modelagem nos campos preditores que forem mais importantes e considerar descartar ou ignorar aqueles que forem menos importantes. O gráfico de importância do preditor ajuda a fazer isso ao indicar a importância relativa de cada preditor na estimativa do modelo. Como os valores são relativos, a soma dos valores para todos os preditores na tela é 1,0. A importância do preditor não tem relação com a precisão do modelo. Ela está relacionada apenas com a importância de cada preditor em fazer uma predição, e não se a predição é precisa ou não.

Predito por Observado

Isso exibe um gráfico de dispersão categorizado dos valores preditos no eixo vertical em função dos valores observados no eixo horizontal. Idealmente, os pontos devem estar em uma linha de 45 graus, e essa visualização poderá informar se algum registro foi particularmente mal predito pelo modelo.

Configurações

Observe que o valor predito é calculado sempre que o modelo é escorado. O nome do novo campo é o nome do campo de destino, prefixado com *$L-*. Por exemplo, para um campo de destino chamado *sales*, o novo campo se chamará *$L-sales*.

Gerar SQL para este modelo Ao usar dados de um banco de dados, código SQL pode ser enviado por push de volta para o banco de dados para execução, fornecendo desempenho superior para muitas operações.

Selecione uma das opções a seguir para especificar como a geração de SQL é executada.

- **Padrão: Escorar usando o Server Scoring Adapter (se instalado) de outra forma no processo.** Se conectado a um banco de dados com um adaptador de escoragem instalado, gera a SQL utilizando o adaptador de escoragem e funções definidas pelo usuário (UDF) associadas e escora seu modelo no banco de dados. Quando nenhum adaptador de escoragem estiver disponível, essa opção busca seus dados novamente a partir do banco de dados e os escora no SPSS Modeler.
- **Escorar fora do Banco de Dados.** Se selecionada, essa opção busca seus dados novamente a partir do banco de dados e os escora no SPSS Modeler.

Nó de Logística

A **Regressão logística**, também conhecida como **regressão nominal**, é uma técnica estatística para classificar registros com base nos valores de campos de entrada. Ela é semelhante a uma regressão linear, mas usa um campo de destino categórico ao invés de um campo numérico. Os modelos binomiais (para destinos com duas categorias discretas) e modelos multinomiais (para destinos com mais de duas categorias) são suportados.

A regressão logística funciona ao construir um conjunto de equações que relacionam os valores do campo de entrada com as probabilidades associadas a cada uma das categorias do campo de saída. Quando o modelo é gerado, ele pode ser utilizado para estimar as probabilidades para os novos dados. Para cada registro, uma probabilidade de associação é calculada para cada categoria de saída possível. A categoria de destino com a probabilidade mais alta é designada como o valor de saída predito para esse registro.

Exemplo binomial. Um provedor de telecomunicações está preocupado com o número de clientes que está perdendo para a concorrência. Utilizando dados de uso de serviço, é possível criar um modelo binomial para prever quais clientes são susceptíveis a mudarem para outro provedor e customizar ofertas

a fim de reter o máximo de clientes possível. Um modelo binomial é utilizado porque o destino possui duas categorias distintas (propensas à transferência ou não).

Nota: apenas para modelos binomiais, os campos de sequência de caracteres devem ser limitados a oito caracteres. Se necessário, sequências de caracteres mais longas podem ser recodificaas utilizando um nó Reclassificar.

Exemplo multinomial. Um provedor de telecomunicações segmentou sua base de clientes por padrões de uso de serviço, categorizando os clientes em quatro grupos. Utilizando dados demográficos para prever a associação ao grupo, é possível criar um modelo multinomial para classificar os possíveis clientes em grupos e, em seguida, customizar ofertas para clientes individuais.

Requisitos. Um ou mais campos de entrada e exatamente um campo de destino categórico com duas ou mais categorias. Para um modelo binomial, o destino deve ter um nível de medição de *Flag*. Para um modelo multinomial, o destino pode ter um nível de medição de *Flag* ou de *Nominal* com duas ou mais categorias. Campos configurados para *Ambos* ou *Nenhum* são ignorados. Os campos utilizados no modelo devem ter seus tipos totalmente instanciados.

Intensidades. Os modelos de regressão logística geralmente são muito precisos. Eles podem manipular campos de entrada simbólicos e numéricos. Eles podem fornecer probabilidades preditas para todas as categorias de destino para que uma segunda melhor suposição possa ser facilmente identificada. Os modelos de logística são mais eficientes quando a associação ao grupo for um campo totalmente categórico. Se a associação ao grupo for baseada em valores de um campo de intervalo contínuo (por exemplo, alto QI versus baixo QI), deve-se considerar a utilização de regressão linear para aproveitar as informações mais ricas oferecidas pelo intervalo completo de valores. Os modelos de logística também podem desempenhar a seleção de campo automática, embora outras abordagens, como modelos de árvore ou Seleção de Variável, possam fazer isto mais rapidamente em grandes conjuntos de dados. Por fim, quando os modelos de logística são bem entendidos por muitos analistas e mineradores de dados, eles poderão ser utilizados por alguns como uma linha de base com relação à qual outras técnicas de modelagem podem ser comparadas.

Ao processar grandes conjuntos de dados, é possível melhorar o desempenho consideravelmente ao desativar o teste de razão de verossimilhança, uma opção de saída avançada. Consulte o tópico “Saída Avançada de Regressão Logística” na página 182 para obter mais informações.

Opções de Modelo do Nó Logística

Nome do modelo É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Utilizar dados particionados. Se um campo de partição for definido, essa opção assegurará que apenas os dados da partição de treinamento sejam utilizados para construir o modelo.

Criar modelos de divisão. Constrói um modelo separado para cada valor possível de campos de entrada que são especificados como campos de divisão. Consulte o tópico “Construindo Modelos de Divisão” na página 28 para obter mais informações.

Procedimento. Especifica se um modelo binomial ou multinomial é criado. As opções disponíveis na caixa de diálogo variam, dependendo do tipo de procedimento de modelagem que for selecionado.

- **Binomial.** Usado quando o campo de destino é um flag ou um campo nominal com dois valores discretos (dicotômicos), como *sim/não*, *ligado/desligado*, *macho/fêmea*.
- **Multinomial.** Usado quando o campo de destino for um campo nominal com mais de dois valores. É possível especificar **Efeitos Principais**, **Fatorial completo** ou **Customizado**.

Incluir constante na equação. Esta opção determina se as equações resultantes incluem um termo constante. Na maioria das situações, deve-se deixar esta opção selecionada.

Modelos Binomiais

Para modelos binomiais, os métodos e opções a seguir estão disponíveis:

Método. Especifique o método a ser utilizado na construção do modelo de regressão logística.

- **Inserir.** Este é o método padrão que insere todos os termos na equação diretamente. Nenhuma seleção de campo é executada na construção do modelo.
- **Forwards Stepwise.** O método Forwards Stepwise de seleção de campo constrói a equação em passos, tal como o nome implica. O modelo inicial é o modelo mais simples possível, sem termos modelo (exceto a constante) na equação. Em cada passo, os termos que ainda não tiverem sido incluídos no modelo são avaliados e, se o melhor desses termos aumentar significativamente o poder preditivo do modelo, ele será incluído. Além disso, os termos que estiverem atualmente no modelo são reavaliados para determinar se algum deles pode ser removido sem reduzir significativamente o modelo. Caso positivo, eles serão removidos. O processo se repete e outros termos são incluídos e/ou removidos. Quando mais nenhum termo puder ser incluído para melhorar o modelo, e mais nenhum termo puder ser removido sem reduzir o modelo, o modelo final será gerado.
- **Backwards Stepwise.** O método Backwards Stepwise é essencialmente o oposto do método Forwards Stepwise. Com esse método, o modelo inicial contém todos os termos como preditores. Em cada passo, os termos no modelo são avaliados e todos os termos que puderem ser removidos sem reduzir significativamente o modelo são removidos. Além disso, os termos removidos anteriormente são reavaliados para determinar se o melhor desses termos aumenta significativamente o poder preditivo do modelo. Caso positivo, ele é incluído novamente no modelo. Quando mais nenhum termo puder ser removido sem reduzir significativamente o modelo e mais nenhum termo puder ser incluído para melhorar o modelo, o modelo final será gerado.

Entradas categóricas. Lista os campos que são identificados como categóricos, ou seja, aqueles com um nível de medição de flag, nominal ou ordinal. É possível especificar o contraste e a categoria base para cada campo categórico.

- **Nome do Campo.** Esta coluna contém os nomes de campo das entradas categóricas e é preenchida com todos os valores flag e nominais nos dados. Para incluir entradas contínuas ou numéricas nesta coluna, clique no ícone Incluir Campos à direita da lista e selecione as entradas necessárias.
- **Contraste.** A interpretação dos coeficientes de regressão para um campo categórico depende dos contrastes que são utilizados. O contraste determina como os testes de hipótese são configurados para comparar as médias estimadas. Por exemplo, se você souber que um campo categórico tem uma ordem implícita, como um padrão ou agrupamento, será possível usar o contraste para modelar essa ordem. Os contrastes disponíveis são:

  Indicador. Os contrastes indicam a presença ou a ausência de associação de categoria. Este é o método padrão.

  Simples. Cada categoria do campo preditor, exceto a categoria de referência, é comparada com a categoria de referência.

  Diferença. Cada categoria do campo preditor, exceto a primeira categoria, é comparada com o efeito médio de categorias anteriores. Também conhecido como contrastes de Helmert reversos.

  Helmert. Cada categoria do campo preditor, exceto a última categoria, é comparada com o efeito médio de categorias subsequentes.

  Repetida. Cada categoria do campo preditor, exceto a primeira categoria, é comparada com a categoria que a precede.

  Polinomial. Contrastes polinomiais ortogonais. As categorias são consideradas igualmente espaçadas. Os contrastes polinomiais estão disponíveis somente para campos numéricos.

  Desvio. Cada categoria do campo preditor, exceto a categoria de referência, é comparada com o efeito geral.

- **Categoria Base.** Especifica como a categoria de referência é determinada para o tipo de contraste selecionado. Selecione **Primeira** para utilizar a primeira categoria para o campo de entrada, classificadas em ordem alfabética, ou selecione **Última** para utilizar a última categoria. O valor padrão é First.

  Nota: esse campo não estará disponível se a configuração de contraste for Diferença, Helmert, Repetida ou Polinomial.

A estimativa do efeito de cada campo na resposta geral é calculada como um aumento ou uma diminuição da probabilidade de cada uma das outras categorias com relação à categoria de referência. Isso pode ajudá-lo a identificar os campos e os valores mais propensos a darem uma resposta específica.

A categoria base é mostrada na saída como 0,0. Isso ocorre porque compará-la com si mesma produz um resultado vazio. Todas as outras categorias são mostradas como equações relevantes à categoria base. Consulte o tópico "Detalhes do Modelo do Nugget de Logística" na página 184 para obter mais informações.

Modelos Multinomiais

Para modelos multinomiais, os métodos e opções a seguir estão disponíveis:

Método. Especifique o método a ser utilizado na construção do modelo de regressão logística.

- **Inserir.** Este é o método padrão que insere todos os termos na equação diretamente. Nenhuma seleção de campo é executada na construção do modelo.
- **Stepwise.** O método Stepwise de seleção de campo constrói a equação em passos, tal como o nome implica. O modelo inicial é o modelo mais simples possível, sem termos modelo (exceto a constante) na equação. Em cada passo, os termos que ainda não tiverem sido incluídos no modelo são avaliados e, se o melhor desses termos aumentar significativamente o poder preditivo do modelo, ele será incluído. Além disso, os termos que estiverem atualmente no modelo são reavaliados para determinar se algum deles pode ser removido sem reduzir significativamente o modelo. Caso positivo, eles serão removidos. O processo se repete e outros termos são incluídos e/ou removidos. Quando mais nenhum termo puder ser incluído para melhorar o modelo, e mais nenhum termo puder ser removido sem reduzir o modelo, o modelo final será gerado.
- **Forwards.** O método de seleção do campo Forwards é semelhante ao método Stepwise pelo fato de que o modelo é construído em passos. No entanto, com este método, o modelo inicial é o modelo mais simples e apenas a constante e os termos podem ser incluídos no modelo. Em cada passo, os termos que ainda não estiverem no modelo são testados sobre o quanto eles podem melhorar o modelo, e o melhor desses termos é incluído no modelo. Quando mais nenhum termo puder ser incluído, ou se o melhor termo candidato não produzir uma melhoria grande o suficiente no modelo, o modelo final será gerado.
- **Backwards.** O método Backwards é essencialmente o oposto do método Forwards. Com esse método, o modelo inicial contém todos os termos como preditores, e os termos somente podem ser removidos do modelo. Os termos modelo que pouco contribuírem para com o modelo são removidos um por um até que mais nenhum termo possa ser removido sem reduzir significativamente o modelo, gerando o modelo final.
- **Backwards Stepwise.** O método Backwards Stepwise é essencialmente o oposto do método Stepwise. Com esse método, o modelo inicial contém todos os termos como preditores. Em cada passo, os termos no modelo são avaliados e todos os termos que puderem ser removidos sem reduzir significativamente o modelo são removidos. Além disso, os termos removidos anteriormente são reavaliados para determinar se o melhor desses termos aumenta significativamente o poder preditivo do modelo. Caso positivo, ele é incluído novamente no modelo. Quando mais nenhum termo puder ser removido sem reduzir significativamente o modelo e mais nenhum termo puder ser incluído para melhorar o modelo, o modelo final será gerado.

Nota: os métodos automáticos, incluindo Stepwise, Forwards e Backwards, são métodos de aprendizado altamente adaptáveis e que possuem uma forte tendência a super ajuste dos dados de treinamento. Ao utilizar esses métodos, é essencialmente importante verificar a validade do modelo resultante, seja com novos dados ou com uma amostra de teste de validação criada utilizando o nó Partição.

Categoria base para resposta. Especifica como a categoria de referência é determinada. Isso é utilizado como a linha de base com relação a quais equações de regressão de todas as outras categorias na resposta são estimadas. Selecione **Primeira** para utilizar a primeira categoria para o campo de destino atual, classificada em ordem alfabética, ou selecione **Última** para utilizar a última categoria. Como alternativa, é possível selecionar **Especificar** para escolher uma categoria específica e selecionar o valor desejado na lista. Os valores disponíveis podem ser definidos para cada campo em um nó Tipo.

Geralmente você especifica a categoria que seria a menos desejada para ser a categoria base, por exemplo, um produto "boi de piranha". As outras categorias são, então, relacionadas a esta categoria base de modo a identificar aquilo que mais provavelmente as faz estar em sua própria categoria. Isso pode ajudá-lo a identificar os campos e os valores mais propensos a darem uma resposta específica.

A categoria base é mostrada na saída como 0,0. Isso ocorre porque compará-la com si mesma produz um resultado vazio. Todas as outras categorias são mostradas como equações relevantes à categoria base. Consulte o tópico “Detalhes do Modelo do Nugget de Logística” na página 184 para obter mais informações.

Tipo de modelo. Há três opções para definir os termos no modelo. Os modelos de **Efeitos Principais** incluem apenas os campos de entrada individualmente e não testam as interações (efeitos multiplicadores) entre os campos de entrada. Os modelos **Fatorial completo** incluem todas as interações e também os efeitos principais do campo de entrada. Os modelos fatoriais completos são melhores para capturar relacionamentos complexos, mas também são muito mais difíceis de interpretar e podem sofrer mais com super ajuste. Devido ao número potencialmente grande de combinações possíveis, os métodos de seleção automática de campo (métodos diferentes de Enter) são desativados para modelos fatoriais completos. Os modelos **Customizados** incluem apenas os termos (efeitos principais e interações) que você especificar. Ao selecionar essa opção, utilize a lista Termos Modelo para incluir ou remover termos no modelo.

Termos Modelo. Ao construir um modelo Customizado, será necessário especificar explicitamente os termos no modelo. A lista mostra o conjunto atual de termos para o modelo. Os botões do lado direito da lista Termos Modelo permitem incluir e remover os termos modelo.

- Para incluir termos no modelo, clique no botão *Incluir novos termos modelo*.
- Para excluir termos, selecione os termos desejados e clique no botão *Excluir termos modelo selecionados*.

Incluindo Termos em um Modelo de Regressão Logística

Ao solicitar um modelo de regressão logística customizado, é possível incluir termos no modelo clicando no botão *Incluir novos termos modelo* na guia Modelo de Regressão Logística. Uma caixa de diálogo Novos Termos é aberta na qual é possível especificar termos.

Tipo de termo a incluir. Há várias maneiras de incluir termos no modelo, com base na seleção de campos de entrada na lista Campos disponíveis.

- **Interação única.** Insere o termo que representa a interação de todos os campos selecionados.
- **Efeitos principais.** Insere um termo de efeito principal (o campo em si) para cada campo de entrada selecionado.
- **Todas as interações de duas vias.** Insere um termo de interação de duas vias (o produto dos campos de entrada) para cada par possível de campos de entrada selecionados. Por exemplo, se você tiver selecionado os campos de entrada *A*, *B* e *C* na lista Campos disponíveis, esse método inserirá os termos *A * B*, *A * C* e *B * C*.

- **Todas as interações de três vias.** Insere um termo de interação de três vias (o produto dos campos de entrada) para cada combinação possível de campos de entrada selecionados, três por vez. Por exemplo, se você tiver selecionado os campos de entrada *A*, *B*, *C* e *D* na lista Campos disponíveis, esse método inserirá os termos *A * B * C*, *A * B * D*, *A * C * D* e *B * C * D*.
- **Todas as interações de quatro vias.** Insere um termo de interação de quatro vias (o produto dos campos de entrada) para cada combinação possível de campos de entrada selecionados, quatro por vez. Por exemplo, se você tiver selecionado os campos de entrada *A*, *B*, *C*, *D* e *E* na lista Campos disponíveis, esse método inserirá os termos *A * B * C * D*, *A * B * C * E*, *A * B * D * E*, *A * C * D * E* e *B * C * D * E*.

Campos disponíveis. Lista os campos de entrada disponíveis a serem utilizados na construção de termos modelo.

Visualizar. Mostra os termos que serão incluídos no modelo se você clicar em **Inserir**, com base nos campos e no tipo de termo selecionados.

Inserir. Insere os termos no modelo (com base na seleção de campos e no tipo de termo atuais) e fecha a caixa de diálogo.

Opções Avançadas do Nó Logística

Se você tiver conhecimento detalhado de regressão logística, as opções avançadas permitirão fazer um ajuste preciso do processo de treinamento. Para acessar as opções avançadas, configure o Modo para **Especialista** na guia Especialista.

Escala (apenas modelos Multinomiais). É possível especificar um valor de escala de dispersão que será utilizado para corrigir a estimativa da matriz de covariâncias de parâmetro. **Pearson** estima o valor de escala utilizando a estatística qui-quadrado de Pearson. **Deviance** estima o valor de escala utilizando a estatística de função de deviance (qui-quadrado de razão de verossimilhança). Também é possível especificar seu próprio valor de escala definido pelo usuário. Ele deve ser um valor numérico positivo.

Incluir todas as probabilidades. Se essa opção for selecionada, as probabilidades para cada categoria do campo de saída serão incluídas em cada registro processado pelo nó. Se essa opção não estiver selecionada, apenas a probabilidade da categoria predita será incluída.

Por exemplo, uma tabela contendo os resultados de um modelo multinomial com três categorias incluirá cinco novas colunas. Uma coluna listará a probabilidade de o resultado ser predito corretamente, a próxima coluna mostrará a probabilidade de que essa predição seja uma ocorrência ou uma perda e mais três colunas mostrarão a probabilidade de que a predição de cada categoria seja uma perda ou uma ocorrência. Consulte o tópico “Nugget do Modelo Logística” na página 184 para obter mais informações.

Nota: essa opção é sempre selecionada para modelos binomiais.

Tolerância à singularidade. Especificar a tolerância usada na verificação de singularidade.

Convergência. Essas opções permitem controlar os parâmetros para convergência do modelo. Ao executar o modelo, as configurações de convergência controlam quantas vezes os diferentes parâmetros são executados repetidamente até poder ver quão bem eles se ajustam. Quanto mais frequentemente os parâmetros forem tentados, mais próximos os resultados serão (ou seja, os resultados convergirão). Consulte o tópico “Opções de Convergência de Regressão Logística” na página 182 para obter mais informações.

Saída. Essas opções permitem solicitar estatísticas adicionais que serão exibidas na saída avançada do nugget do modelo construído pelo nó. Consulte o tópico “Saída Avançada de Regressão Logística” na página 182 para obter mais informações.

Progresso. Essas opções permitem controlar os critérios para incluir e remover campos com os métodos de estimação Stepwise, Forwards, Backwards ou Backwards Stepwise. (O botão estará desativado se o método Inserir for selecionado). Consulte o tópico "Opções de Progresso de Regressão Logística" na página 183 para obter mais informações.

Opções de Convergência de Regressão Logística

É possível configurar os parâmetros de convergência para a estimação de modelo de regressão logística.

Máximo de iterações. Especifique o número máximo de iterações para estimativa do modelo.

Divisão máxima da etapa pela metade. A Etapa pela metade é uma técnica utilizada por uma regressão logística para lidar com as complexidades no processo de estimação. Sob circunstâncias normais, deve-se utilizar a configuração padrão.

Convergência de log da verossimilhança. As iterações pararão se a mudança relativa no log de verossimilhança for menor que esse valor. O critério não será utilizado se o valor for 0.

Convergência de parâmetro. As iterações pararão se a mudança absoluta ou relativa nas estimativas de parâmetro for menor que esse valor. O critério não será utilizado se o valor for 0.

Delta (apenas modelos Multinomiais). É possível especificar um valor entre 0 e 1 a ser incluído em cada célula vazia (combinação de valores do campo de entrada e do campo de saída). Isso pode ajudar o algoritmo de estimação a lidar com dados onde houver muitas combinações possíveis de valores de campo com relação ao número de registros nos dados. O padrão é 0.

Saída Avançada de Regressão Logística

Selecione a saída opcional que você deseja exibir na saída avançada do nugget do modelo Regressão. Para visualizar a saída avançada, procure o nugget do modelo e clique na guia **Avançado**. Consulte o tópico "Saída Avançada do Nugget do Modelo Logística" na página 186 para obter mais informações.

Opções Binomiais

Selecione os tipos de saída a serem gerados para o modelo. Consulte o tópico "Saída Avançada do Nugget do Modelo Logística" na página 186 para obter mais informações.

Exibição. Selecione se deseja exibir os resultados em cada passo ou aguardar até que todos os passos tenham sido trabalhados.

IC para exp(B). Selecione os intervalos de confiança para cada coeficiente (mostrados como Beta) na expressão. Especifique o nível do intervalo de confiança (o padrão é 95%).

Diagnósticos de Resíduo. Solicita uma tabela de Diagnósticos de Casos de resíduos.

- **Valor discrepante externo (desv. padrão).** Lista somente os casos de resíduo para os quais o valor padronizado absoluto da variável listada for pelo menos tão grande quanto o valor que você especificar. O valor padrão é 2.
- **Todos os casos.** Inclui todos os casos na tabela Diagnóstico de Casos de resíduos.

  Nota: como esta opção lista cada um dos registros de entrada, isso pode resultar em uma tabela excepcionalmente grande no relatório, com uma linha para cada registro.

Corte de Classificação. Permite determinar o ponto de corte para classificação de casos. Os casos com valores preditos que excederem o corte de classificação são classificados como positivos, ao passo que aqueles com valores preditos menores que o corte são classificados como negativos. Para alterar o padrão, insira um valor entre 0,01 e 0,99.

Opções Multinomiais

Selecione os tipos de saída a serem gerados para o modelo. Consulte o tópico “Saída Avançada do Nugget do Modelo Logística” na página 186 para obter mais informações.

Nota: selecionar a opção **Testes de razão de verossimilhança** aumenta significativamente o tempo de processamento necessário para construir um modelo de regressão logística. Se o seu modelo estiver demorando muito para construir, considere desativar essa opção ou utilizar as estatísticas Wald e Escore. Consulte o tópico “Opções de Progresso de Regressão Logística” para obter mais informações.

Histórico de iteração para cada. Selecione o intervalo de passos para impressão do status da iteração na saída avançada.

Intervalo de Confiança. Os intervalos de confiança para coeficientes nas equações. Especifique o nível do intervalo de confiança (o padrão é 95%).

Opções de Progresso de Regressão Logística

Essas opções permitem controlar os critérios para incluir e remover campos com os métodos de estimação Stepwise, Forwards, Backwards ou Backwards Stepwise.

Número de termos no modelo (apenas modelos Multinomiais). É possível especificar o número mínimo de termos no modelo para modelos Backwards e Backwards Stepwise e o número máximo de termos para os modelos Forwards e Stepwise. Se você especificar um valor mínimo maior que 0, o modelo incluirá todos esses termos, mesmo se alguns dos termos tiverem sido removidos com base em critérios estatísticos. A configuração mínima é ignorada para os modelos Forwards, Stepwise e Enter. Se você especificar um máximo, alguns termos poderão ser omitidos do modelo, mesmo se eles tiverem sido selecionados com base em critérios estatísticos. A configuração **Especificar Máximo** é ignorada para os modelos Backwards, Backwards Stepwise e Enter.

Critério de entrada (apenas modelos Multinomiais). Selecione **Escore** para aumentar a velocidade de processamento. A opção **Razão de Verossimilhança** pode fornecer estatísticas um tanto mais robustas, mas poderá demorar mais tempo para calcular. A configuração padrão é utilizar a estatística Escore.

Critério de remoção. Selecione **Razão de Verossimilhança** para um modelo mais robusto. Para reduzir o tempo necessário para construir o modelo, é possível tentar selecionar **Wald**. Entretanto, se você tiver uma separação completa ou quase completa nos dados (que você pode determinar utilizando a guia Avançado no nugget do modelo), a estatística Wald se tornará particularmente não confiável e não deverá ser utilizada. A configuração padrão é utilizar a estatística de razão de verossimilhança. Para modelos binomiais, há a opção adicional **Condicional**. Isso fornece um teste de remoção com base na probabilidade da estatística de razão de verossimilhança com base nas estimativas de parâmetro condicional.

Limites de significância para critérios. Esta opção permite especificar critérios de seleção com base na probabilidade estatística (o valor *p*) associada a cada campo. Os campos serão incluídos no modelo somente se o valor de *p* associado for menor que o valor de **Entrada** e será removido apenas se o valor de *p* for maior que o valor de **Remoção**. O valor de **Entrada** deve ser menor que o valor de **Remoção**.

Requisitos para entrada ou remoção (somente modelos Multinomiais). Para alguns aplicativos, não faz sentido matemático incluir termos de interação no modelo, a menos que o modelo também contenha os termos de ordem inferior para os campos envolvidos no termo de interação. Por exemplo, poderá não fazer sentido incluir *A * B* no modelo, a menos que *A* e *B* também estejam incluídos no modelo. Estas opções permitem determinar como tais dependências são manipuladas durante a seleção do termo stepwise.

- **Hierarquia para efeitos discretos.** Os efeitos de ordem superior (interações que envolvem mais campos) inserirão o modelo somente se todos os efeitos de ordem inferior (efeitos principais ou interações que envolvem menos campos) para os campos relevantes já estiverem no modelo, e os efeitos de ordem inferior não serão removidos se os efeitos de ordem superior que envolvem os mesmos campos estiverem no modelo. Esta opção se aplica apenas a campos categóricos.
- **Hierarquia para todos os efeitos.** Essa opção funciona da mesma forma que a opção anterior, exceto que ela se aplica a todos os campos de entrada.
- **Contenção para todos os efeitos.** Os efeitos poderão ser incluídos no modelo apenas se todos os efeitos contidos no efeito também estiverem incluídos no modelo. Essa opção é semelhante à opção **Hierarquia para todos os efeitos**, exceto que os campos contínuos são tratados de forma um pouco diferente. Para que um efeito contenha outro efeito, o efeito contido (ordem inferior) deverá incluir *todos* os campos contínuos envolvidos no efeito de contenção (ordem superior), e os campos categóricos do efeito contido deverão ser um subconjunto daqueles no efeito de contenção. Por exemplo, se *A* e *B* forem campos categóricos e *X* for um campo contínuo, o termo *A * B * X* conterá os termos *A * X* e *B * X*.
- **Nenhum.** Nenhum relacionamento é aplicado; os termos são incluídos e removidos do modelo independentemente.

Nugget do Modelo Logística

Um nugget do modelo Logística representa a equação estimada por um nó Logística. Ele contém todas as informações capturadas pelo modelo de regressão logística, bem como informações sobre a estrutura e o desempenho do modelo. Esse tipo de equação também pode ser gerado por outros modelos, como Oracle SVM.

Ao executar um fluxo que contém um nugget do modelo Logística, o nó inclui dois novos campos contendo a predição do modelo e a probabilidade associada. Os nomes dos novos campos são derivados do nome do campo de saída que está sendo predito, prefixados com *$L-* para a categoria predita e com *$LP-* para a probabilidade associada. Por exemplo, para um campo de saída denominado *colorpref*, os novos campos serão denominados *$L-colorpref* e *$LP-colorpref*. Além disso, se você selecionou a opção **Incluir todas as probabilidades** no nó Logística, um campo adicional será incluído para cada categoria do campo de saída, contendo a probabilidade pertencente à categoria correspondente de cada registro. Estes campos adicionais nomeados com base nos valores do campo de saída, prefixados com *$LP-*. Por exemplo, se os valores legais de *colorpref* forem *Red*, *Green* e *Blue*, três novos campos serão incluídos: *$LP-Red*, *$LP-Green* e *$LP-Blue*.

Gerando um nó Filtro. O menu Gerar permite criar um novo nó Filtro para transmitir os campos de entrada com base nos resultados do modelo. Os campos que forem eliminados do modelo devido à multicolinearidade serão filtrados pelo nó gerado, bem como os campos não utilizados no modelo.

Detalhes do Modelo do Nugget de Logística

Para modelos multinomiais, a guia Modelo em um nugget do modelo Logística possui uma exibição de divisão com equações de modelo na área de janela esquerda, e a importância do preditor à direita. Para modelos binomiais, a guia exibe apenas a importância do preditor. Consulte o tópico “Importância do preditor” na página 44 para obter mais informações.

Equações de Modelo

Para modelos multinomiais, a área de janela à esquerda exibe as equações reais estimadas para o modelo de regressão logística. Há uma equação para cada categoria no campo de destino, exceto a categoria de linha de base. As equações são exibidas em um formato de árvore. Esse tipo de equação também pode ser gerado por alguns outros modelos, como Oracle SVM.

Equação Para. Mostra as equações de regressão utilizadas para derivar as probabilidades da categoria de destino, dado um conjunto de valores do preditor. A última categoria do campo de destino é considerada

a **categoria de linha de base**; as equações mostradas fornecem o log-chance para as outras categorias de destino com relação à categoria de linha de base para um conjunto específico de valores do preditor. A probabilidade predita para cada categoria do padrão do preditor fornecido é derivada desses valores de log-chance.

Como as Probabilidades São Calculadas

Cada equação calcula o log-chance para uma determinada categoria de destino, com relação à categoria de linha de base. O **log-chance**, também chamado de **logit**, é a razão da probabilidade para a categoria de destino especificada com a da categoria de linha de base, com a função de logaritmo natural aplicada ao resultado. Para a categoria de linha de base, as chances da categoria com relação a si mesma é 1,0 e, portanto, o log-chance é 0. Isso pode ser considerado como uma equação implícita para a categoria de linha de base em que todos os coeficientes são 0.

Para derivar a probabilidade do log-chance para uma categoria de destino específica, pegue o valor logit calculado pela equação para essa categoria e aplique a seguinte fórmula:

$P(group_i) = \exp(g_i) / \Sigma_k \exp(g_k)$

em que g é o log-chance calculado, i é o índice de categoria e k vai de 1 até o número de categorias de destino.

Importância do preditor

Opcionalmente, um gráfico que indica a importância relativa de cada preditor na estimativa do modelo também pode ser exibido na guia Modelo. Geralmente você desejará focar seus esforços de modelagem nos preditores que forem mais importantes e considerar descartar ou ignorar aqueles que forem menos importantes. Observe que este gráfico estará disponível apenas se **Calcular a importância do preditor** estiver selecionada na guia Análise antes de gerar o modelo. Consulte o tópico "Importância do preditor" na página 44 para obter mais informações.

Nota: a importância do preditor pode levar mais tempo para calcular para regressão logística do que para outros tipos de modelos, e não é selecionada na guia Análise por padrão. Selecionar essa opção pode diminuir o desempenho, principalmente com grandes conjuntos de dados.

Sumarização do Nugget do Modelo de Logística

A sumarização de um modelo de regressão logística exibe os campos e as configurações usados para gerar o modelo. Além disso, se você tiver executado um nó Análise anexado a este nó de modelagem, as informações dessa análise também serão exibidas nesta seção. Para obter informações gerais sobre como utilizar o navegador do modelo, consulte "Procurando Nuggets do Modelo" na página 42.

Configurações do Nugget do Modelo de Logística

A guia Configurações em um nugget do modelo Logística especifica opções para confianças, probabilidades, escores de propensão e geração de SQL durante a escoragem do modelo. Esta guia estará disponível somente após o nugget do modelo ter sido incluído em um fluxo e exibe diferentes opções, dependendo do tipo de modelo e da resposta.

Modelos Multinomiais

Para modelos multinomiais, as opções a seguir estão disponíveis.

Calcular confianças Especifica se as confianças são calculadas durante a escoragem.

Calcular escores de propensão bruta (apenas resposta de flag) Para modelos apenas com respostas de flag, é possível solicitar escores de propensão bruta que indicam a probabilidade do resultado `true`

especificado para o campo de destino. Esses são um complemento dos valores de predição e de confiança padrão. Os escores de propensão ajustada não estão disponíveis. Consulte o tópico "Opções de Análise do Nó de Modelagem" na página 35 para obter mais informações.

Incluir todas as probabilidades Especifica se as probabilidades de cada categoria do campo de saída são incluídas em cada registro processado pelo nó. Se essa opção não estiver selecionada, apenas a probabilidade da categoria predita será incluída. Para uma resposta nominal com três categorias, por exemplo, a saída da escoragem incluirá uma coluna para cada uma das três categorias, além de uma quarta coluna indicando a probabilidade de qualquer categoria ser predita. Por exemplo, se as probabilidades para categorias *Vermelho*, *Verde* e *Azul* forem 0,6, 0,3 e 0,1 respectivamente, a categoria predita será *Vermelho*, com uma probabilidade de 0,6.

Gerar SQL para este modelo Ao usar dados de um banco de dados, código SQL pode ser enviado por push de volta para o banco de dados para execução, fornecendo desempenho superior para muitas operações.

Selecione uma das opções a seguir para especificar como a geração de SQL é executada.

- **Padrão: Escorar usando o Server Scoring Adapter (se instalado) no processo** Se conectado a um banco de dados com um adaptador de escoragem instalado, gera a SQL utilizando o adaptador de escoragem e funções definidas pelo usuário (UDF) associadas e escora seu modelo no banco de dados. Quando nenhum adaptador de escoragem estiver disponível, essa opção busca seus dados novamente a partir do banco de dados e os escora no SPSS Modeler.
- **Escorar ao converter para SQL nativo** Se selecionada, gera SQL nativo para escorar o modelo no banco de dados.

  Nota: Embora essa opção possa fornecer resultados mais rápidos, o tamanho e a complexidade do SQL nativo aumentam conforme a complexidade do modelo aumenta.
- **Escorar fora do Banco de dados** Se selecionada, esta opção busca seus dados novamente a partir de seu banco de dados e os escora no SPSS Modeler.

Nota: Para modelos multinomiais, a geração de SQL estará indisponível se **Incluir todas as probabilidades** for selecionada ou, para modelos com respostas nominais, se **Calcular confianças** tiver sido selecionado. A geração de SQL com cálculos de confiança é suportada para modelos multinomiais apenas com campos de flag. A geração de SQL não está disponível para modelos binomiais.

Modelos Binomiais

Para modelos binomiais, as confianças e as probabilidades são sempre ativadas, e as configurações que permitiriam desativar essas opções não estão disponíveis. A geração de SQL não está disponível para modelos binomiais. A única configuração que pode ser alterada para modelos binomiais é a possibilidade de calcular escores de propensão bruta. Conforme mencionado anteriormente para modelos multinomiais, isso se aplica aos modelos apenas com respostas de flag. Consulte o tópico "Opções de Análise do Nó de Modelagem" na página 35 para obter mais informações.

Saída Avançada do Nugget do Modelo Logística

A saída avançada para regressão logística (também conhecida como **regressão nominal**) fornece informações detalhadas sobre o modelo de estimativa e seu desempenho. A maioria das informações contidas na saída avançada é muito técnica e um conhecimento amplo da análise de regressão logística é necessário para interpretar corretamente esta saída.

Avisos. Indica quaisquer avisos ou problemas em potencial com os resultados.

Sumarização do processamento de caso. Lista o número de registros processados, divididos por cada campo simbólico no modelo.

Sumarização do passo (opcional). Lista os efeitos incluídos ou removidos em cada passo da criação do modelo, quando utilizar a seleção de campo automática.

Nota: mostrado apenas para os métodos Stepwise, Forwards, Backwards ou Backwards Stepwise.

Histórico de iteração (opcional). Mostra o histórico de iteração das estimativas de parâmetro para cada *n* iterações que começam com as estimativas iniciais, em que *n* é o valor do intervalo de impressão. O padrão é para imprimir cada iteração (n=1).

Informações de ajuste do modelo (modelos Multinomiais). Mostra o teste de razão de verossimilhança (final) com relação ao qual todos os coeficientes de parâmetro são 0 (Apenas Intercepto).

Classificação (opcional). Mostra a matriz de valores de campo de saída preditos e reais com porcentagens.

Estatística qui-quadrado de Qualidade do ajuste (opcional). Mostra estatísticas de Pearson e de qui-quadrado de razão de verossimilhança. Essas estatísticas testam o ajuste geral do modelo para os dados de treinamento.

Qualidade do ajuste Hosmer e Lemeshow (opcional). Mostra os resultados de casos de agrupamento em decis de risco e compara a probabilidade observada com a probabilidade esperada em cada decil. Essa estatística de qualidade do ajuste é mais robusta que a estatística de qualidade do ajuste tradicional usada em modelos multinomiais, especialmente para modelos com covariáveis e estudos contínuos com tamanhos pequenos de amostras.

Pseudo R-quadrado (opcional). Mostra as medidas *R* quadrado de Cox e Snell, Nagelkerke e McFadden de ajuste do modelo. Essas estatísticas são de alguma forma análogas às estatísticas *R* quadrado na regressão linear.

Medidas de monotonicidade (opcional). Mostra o número de pares concordantes, pares discordantes e pares empatados nos dados, bem como a porcentagem do número total de pares que cada um representa. O Somers' D, Gama de Goodman e Kruskal, tau-a de Kendall e o Índice C de Concordância também são exibidos nessa tabela.

Critérios de informações (opcional). Mostra o Akaike Information Criterion (AIC) e o Critério de Informação Bayesiano (BIC) de Schwarz.

Testes de razão de verossimilhança (opcional). Mostra estatísticas de teste para determinar se os coeficientes dos efeitos de modelo são estatisticamente diferentes de 0. Os campos de entrada significativos são aqueles com níveis de significância muito pequenos na saída (rotulado como *Sig.*) .

Estimativas de parâmetro (opcional). Mostra as estimativas dos coeficientes da equação, os testes desses coeficientes, as razões de chance derivadas dos coeficientes denominadas *Exp(B)* e os intervalos de confiança para as razões de chances.

Matriz de covariância/correlações assintótica (opcional). Mostra as covariâncias assintóticas e/ou correlações de estimativas de coeficiente.

Frequências observadas e preditas (opcional). Para cada padrão de covariável, mostra as frequências observadas e preditas para cada valor de campo de saída. Esta tabela pode ficar muito grande, principalmente para modelos com campos de entrada numéricos. Se a tabela resultante for muito grande para ser prática, ela será omitida e um aviso será exibido.

Nó PCA/Fator

O nó PCA/Fator fornece técnicas poderosas de redução de dados para reduzir a complexidade de seus dados. Duas abordagens semelhantes, porém distintas, são fornecidas.

- A **análise de componentes principais (PCA)** localiza combinações lineares dos campos de entrada que realizam as melhores tarefas de captura de variância no conjunto de campos inteiro, no qual os componentes são ortogonais (perpendiculares) uns aos outros. O PCA foca em todas as variâncias, incluindo variância compartilhada e exclusiva.
- A **análise fatorial** tenta identificar os conceitos, ou **fatores**, subjacentes que explicam o padrão de correlações dentro de um conjunto de campos observados. A análise fatorial foca apenas na variância compartilhada. A variância que é exclusiva para os campos específicos não é considerada na estimativa do modelo. Vários métodos de análise fatorial são fornecidos pelo nó Fator/PCA.

Para ambas as abordagens, o objetivo é localizar um pequeno número de campos derivados que sumarizem efetivamente as informações no conjunto de campos original.

Requisitos. Apenas campos numéricos podem ser utilizados em um modelo PCA-Fator. Para estimar uma análise fatorial ou PCA, um ou mais campos são necessários com o papel configurado como campos de *Entrada*. Os campos com o papel configurado para *Destino*, *Ambos* ou *Nenhum* são ignorados porque são campos não numéricos.

Intensidades. A análise fatorial e o PCA podem reduzir efetivamente a complexidade de seus dados sem sacrificar grande parte do conteúdo das informações. Essas técnicas podem ajudar a construir modelos mais robustos que executam mais rapidamente do que seria possível com os campos de entrada brutos.

Opções de Modelo do Nó PCA/Fator

Nome do modelo É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Utilizar dados particionados. Se um campo de partição for definido, essa opção assegurará que apenas os dados da partição de treinamento sejam utilizados para construir o modelo.

Método de extração. Especifique o método a ser utilizado para redução de dados.

- **Componentes Principais.** Este é o método padrão que utiliza o PCA para localizar os componentes que sumarizem os campos de entrada.
- **Quadrados Mínimos Não Ponderados.** Esse método de análise fatorial funciona localizando o conjunto de fatores que for mais bem capaz de reproduzir o padrão de relacionamentos (correlações) entre os campos de entrada.
- **Quadrados Mínimos Generalizados.** Este método de análise fatorial é semelhante aos quadrados mínimos não ponderados, com a exceção de que ele utiliza a ponderação para desenfatizar os campos com muita variância exclusiva (não compartilhada).
- **Máxima Verossimilhança.** Este método de análise fatorial produz equações fatoriais que mais provavelmente produziram o padrão de relacionamentos (correlações) observado nos campos de entrada, com base em suposições sobre a forma desses relacionamentos. Especificamente, o método supõe que os dados de treinamento seguem uma distribuição multivariada normal.
- **Fatoração do Eixo Principal.** Este método de análise fatorial é muito semelhante ao método de componentes principais, com a exceção de que ele foca apenas na variância compartilhada.
- **Fatoração Alpha.** Este método de análise fatorial considera os campos na análise como uma amostra do universo de possíveis campos de entrada. Ele maximiza a confiabilidade estatística dos fatores.
- **Fatoração de Imagem.** Este método de análise fatorial usa a estimação de dados para isolar a variância comum e localizar os fatores que descrevem essa variância.

Opções Avançadas do Nó PCA/Fator

Se você tiver conhecimento detalhado da análise fatorial e de PCA, as opções avançadas permitirão fazer um ajuste preciso do processo de treinamento. Para acessar as opções avançadas, configure o Modo para **Especialista** na guia Especialista.

Valores omissos. Por padrão, o IBM SPSS Modeler utiliza apenas registros que tiverem valores válidos para todos os campos utilizados no modelo. (Às vezes isso é chamado de **exclusão de lista** de valores omissos). Se houver muitos dados omissos, você poderá achar que essa abordagem elimina muitos registros, deixando-o sem dados suficientes para gerar um bom modelo. Nesses casos, é possível desmarcar a opção **Usar somente registros completos**. O IBM SPSS Modeler, em seguida, tenta usar o máximo de informações possível para estimar o modelo, incluindo registros nos quais alguns dos campos possuem valores omissos. (Às vezes isso é chamado de **exclusão dos pares** de valores omissos). No entanto, em algumas situações, usar registros incompletos dessa maneira pode levar a problemas computacionais durante a estimativa do modelo.

Campos. Especifique se deseja utilizar a matriz de correlações (o padrão) ou a matriz de covariâncias dos campos de entrada na estimativa do modelo.

Máximo de iterações por convergência. Especifique o número máximo de iterações para estimativa do modelo.

Fatores de extração. Há duas maneiras para selecionar o número de fatores a serem extraídos dos campos de entrada.

- **Autovalor acima.** Esta opção reterá todos os fatores ou componentes com autovalores maiores que o critério especificado. Os **autovalores** medem a capacidade de cada fator ou componente para sumarizar a variância no conjunto de campos de entrada. O modelo reterá todos os fatores ou componentes com autovalores maiores que o valor especificado ao utilizar a matriz de correlações. Ao utilizar a matriz de covariâncias, o critério é o valor especificado vezes o autovalor médio. Esse ajuste de escala fornece a essa opção um significado semelhante para ambos os tipos de matriz.
- **Número máximo.** Essa opção reterá o número especificado de fatores ou componentes em ordem decrescente de autovalores. Em outras palavras, os fatores ou componentes correspondentes aos *n* valores mais altos são retidos, em que *n* é o critério especificado. O critério de extração padrão é cinco fatores/componentes.

Formato de matriz de componente/fator. Estas opções controlam o formato da matriz de fatores (ou matriz de componente para modelos PCA).

- **Ordenar valores.** Se essa opção for selecionada, os carregamentos de fator na saída do modelo serão ordenados numericamente.
- **Ocultar valores abaixo.** Se essa opção for selecionada, os escores abaixo do limite especificado serão ocultados na matriz para facilitar a visualização do padrão na matriz.

Rotação. Essas opções permitem controlar o método de rotação para o modelo. Consulte o tópico "Opções de Rotação do Nó PCA/Fator" para obter mais informações.

Opções de Rotação do Nó PCA/Fator

Em muitos casos, girar matematicamente o conjunto de fatores retidos pode aumentar a utilidade deles e, em particular, a capacidade de interpretação. Selecione um método rotação:

- **Nenhuma rotação.** A opção padrão. Nenhuma rotação é utilizada.
- **Varimax.** Um método de rotação ortogonal que minimiza o número de campos altamente carregados em cada fator. Ele simplifica a interpretação dos fatores.
- **Oblimin direta.** Um método de rotação oblíqua (não ortogonal). Quando **Delta** é igual a 0 (o padrão), as soluções são oblíquas. Conforme delta se torna mais negativo, os fatores se tornam menos oblíquos. Para substituir o delta padrão de 0, insira um número menor ou igual a 0,8.

- **Quartimax.** Um método ortogonal que minimiza o número de fatores necessários para explicar cada campo. Ele simplifica a interpretação dos campos observados.
- **Equamax.** Um método de rotação é uma combinação do método Varimax, que simplifica os fatores, e do método Quartimax, que simplifica os campos. O número de campos que são altamente carregados em um fator e o número de fatores necessários para explicar um campo são minimizados.
- **Promax.** Um rotação oblíqua que permite que fatores sejam correlacionados. Ele pode ser calculado mais rapidamente do que uma rotação oblimin direta, podendo, portanto, ser útil para grandes conjuntos de dados. O **Kappa** controla a obliquidade da solução (a extensão até a qual os fatores podem ser correlacionados).

Nugget do Modelo PCA/Fator

Um nugget do modelo PCA/Fator representa a análise fatorial e a análise de componentes principais (PCA) criadas por um nó PCA/Fator. Elas contêm todas as informações capturadas pelo modelo treinado, bem como informações sobre o desempenho e as características do modelo.

Ao executar um fluxo contendo um modelo de equação fatorial, o nó inclui um novo campo para cada fator ou componente no modelo. Os novos nomes de campos são derivados do nome do modelo, prefixados com *$F-* e sufixados com *-n*, em que *n* é o número do fator ou componente. Por exemplo, se o seu modelo for denominado *Factor* e contiver três fatores, os novos campos serão denominados *$F-Factor-1*, *$F-Factor-2* e *$F-Factor-3*.

Para ter uma ideia melhor dos itens que o modelo de fator codificou, é possível fazer uma análise mais detalhada.Uma maneira útil para visualizar o resultado do modelo fator é visualizar as correlações entre os fatores e os campos de entrada utilizando um nó Estatísticas. Isso mostra quais campos de entrada são altamente carregados em quais fatores e pode ajudar a descobrir se seus fatores possuem algum significado ou interpretação subjacente.

Também é possível avaliar o modelo de fator usando as informações disponíveis na saída avançada. Para visualizar a saída avançada, clique na guia **Avançado** do navegador do nugget do modelo. A saída avançada contém muitas informações detalhadas e é destinada a usuários com conhecimento amplo da análise fatorial ou PCA. Consulte o tópico "Saída Avançada do Nugget do Modelo PCA/Factor" para obter mais informações.

Equações do Nugget do Modelo PCA/Fator

A guia Modelo de um nugget do modelo Fator exibe a equação do escore de fator para cada fator. Os escores de fator ou de componente são calculados multiplicando cada valor do campo de entrada pelo seu coeficiente e somando os resultados.

Sumarização do Nugget do Modelo PCA/Factor

A guia Sumarização de um modelo de fator exibe o número de fatores retidos no modelo Fator/PCA, com informações adicionais sobre os campos e as configurações usados para gerar o modelo. Consulte o tópico "Procurando Nuggets do Modelo" na página 42 para obter mais informações.

Saída Avançada do Nugget do Modelo PCA/Factor

A saída avançada para análise fatorial fornece informações detalhadas sobre o modelo estimado e seu desempenho. A maioria das informações contidas na saída avançada é muito técnica e um conhecimento amplo de análise fatorial é necessário para interpretar corretamente esta saída.

Avisos. Indica quaisquer avisos ou problemas em potencial com os resultados.

Comunalidades. Mostra a proporção de variância de cada campo que é considerada pelos fatores ou componentes. *Inicial* fornece comunalidades iniciais com o conjunto completo de fatores (o modelo começa com tantos fatores quanto houver campos de entrada) e *Extração* fornece as comunalidades com base no conjunto de fatores retidos.

Variância total explicada. Mostra a variância total explicada pelos fatores no modelo. *Autovalores iniciais* mostra a variância explicada pelo conjunto completo de fatores iniciais. *Somas de Extração de Carregamentos Quadrados* mostra a variância explicada pelos fatores retidos no modelo. *Somas de Rotação de Carregamentos Quadrados* mostra a variância explicada pelos fatores girados. Observe que para rotações oblíquas, *Somas de Rotação de Carregamentos Quadrados* mostra apenas as somas de carregamentos ao quadrado e não mostram porcentagens de variância.

Matriz de fator (ou componente). Mostra as correlações entre os campos de entrada e os fatores não girados.

Matriz de fator (ou componente) girado. Mostra as correlações entre os campos de entrada e os fatores girados para rotações ortogonais.

Matriz de padrão. Mostra as correlações parciais entre os campos de entrada e os fatores girados para rotações oblíquas.

Matriz de estrutura. Mostra as correlações simples entre os campos de entrada e os fatores girados para rotações oblíquas.

Matriz de correlações de fatores. Mostra as correlações entre os fatores para rotações oblíquas.

Nó Discriminante

A análise discriminante constrói um modelo preditivo para associação ao grupo. O modelo é composto de uma função discriminante (ou, para mais de dois grupos, um conjunto de funções discriminantes) com base nas combinações lineares das variáveis preditoras que fornecem a melhor discriminação entre os grupos. As funções são geradas a partir de uma amostra de casos para os quais a associação ao grupo é conhecida e podem, em seguida, ser aplicadas aos novos casos que tiverem medições para as variáveis preditoras e possuírem associação ao grupo desconhecida.

Exemplo. Uma empresa de telecomunicações pode utilizar a análise discriminante para classificar clientes em grupos com base nos dados de uso. Isso permite escorar possíveis clientes e destinar aqueles que mais provavelmente estarão nos grupos mais valiosos.

Requisitos. Um ou mais campos de entrada e exatamente um campo de destino são necessários. O destino deve ser um campo categórico (com um nível de medição do *Flag* ou *Nominal*) com armazenamento de sequência de caracteres ou de número inteiro. (O armazenamento pode ser convertido utilizando um nó Preenchimento ou Derivar, se necessário). Os campos configurados para *Ambos* ou *Nenhum* são ignorados. Os campos utilizados no modelo devem ter seus tipos totalmente instanciados.

Intensidades. A análise discriminante e a Regressão Logística são modelos de classificação adequados. No entanto, a análise discriminante faz mais suposições sobre os campos de entrada, por exemplo, eles normalmente são distribuídos e devem ser contínuos, e fornecerão melhores resultados se esses requisitos forem atendidos, especialmente se o tamanho da amostra for pequeno.

Opções de Modelo do Nó Discriminante

Nome do modelo É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Utilizar dados particionados. Se um campo de partição for definido, essa opção assegurará que apenas os dados da partição de treinamento sejam utilizados para construir o modelo.

Criar modelos de divisão. Constrói um modelo separado para cada valor possível de campos de entrada que são especificados como campos de divisão. Consulte o tópico "Construindo Modelos de Divisão" na página 28 para obter mais informações.

Método. As opções a seguir estão disponíveis para inserir preditores no modelo:

- **Inserir.** Este é o método padrão que insere todos os termos na equação diretamente. Os termos que não aumentarem significativamente o poder preditivo do modelo não são incluídos.
- **Stepwise.** O modelo inicial é o modelo mais simples possível, sem termos modelo (exceto a constante) na equação. Em cada passo, os termos que ainda não tiverem sido incluídos no modelo são avaliados e, se o melhor desses termos aumentar significativamente o poder preditivo do modelo, ele será incluído.

Nota: o método Stepwise possui uma forte tendência a super ajuste dos dados de treinamento. Ao utilizar esses métodos, é essencialmente importante verificar a validade do modelo resultante com uma amostra de teste de validação ou com novos dados.

Opções Avançadas do Nó Discriminante

Se você tiver conhecimento detalhado da análise discriminante, as opções avançadas permitirão fazer um ajuste preciso do processo de treinamento. Para acessar as opções avançadas, configure o **Modo** para **Especialista** na guia Especialista.

Probabilidades Anteriores. Esta opção determina se os coeficientes de classificação são ajustados para um conhecimento a priori de associação ao grupo.

- **Todos os grupos iguais.** As probabilidades anteriores iguais são assumidas para todos os grupos; isso não tem efeito sobre os coeficientes.
- **Calcular a partir de tamanhos de grupo.** Os tamanhos de grupo observados em sua amostra determinam as probabilidades anteriores de associação ao grupo. Por exemplo, se 50% das observações incluídas na análise caírem no primeiro grupo, 25% no segundo e 25% no terceiro, os coeficientes de classificação serão ajustados para aumentar a probabilidade de associação no primeiro grupo relativo aos outros dois.

Usar Matriz de Covariâncias. É possível optar por classificar casos utilizando uma matriz de covariâncias dentro de grupos ou uma matriz de covariâncias separada de grupos.

- *Dentro de grupos*. A matriz de covariâncias dentro de grupos em conjunto é usada para classificar casos.
- *Separado de grupos*. As matrizes de covariância grupos-separados são usadas para classificação. Como a classificação é baseada nas funções discriminantes (não com base nas variáveis originais), essa opção nem sempre é equivalente à discriminação quadrática.

Saída. Essas opções permitem solicitar estatísticas adicionais que serão exibidas na saída avançada do nugget do modelo construído pelo nó. Consulte o tópico "Opções de Saída do Nó Discriminante" na página 193 para obter mais informações.

Progresso. Essas opções permitem controlar os critérios para incluir e remover campos com o método de estimativa Stepwise. (O botão estará desativado se o método Inserir for selecionado). Consulte o tópico "Opções de Progresso do Nó Discriminante" na página 194 para obter mais informações.

Opções de Saída do Nó Discriminante

Selecione a saída opcional que você deseja exibir na saída avançada do nugget do modelo de regressão logística. Para visualizar a saída avançada, procure o nugget do modelo e clique na guia **Avançado**. Consulte o tópico "Saída Avançada do Nugget do Modelo Discriminante" na página 195 para obter mais informações.

Descritivos. As opções disponíveis são médias (incluindo desvios padrão), ANOVAs univariados e teste *M* de Box.

- *Médias*. Exibe as médias totais e de grupo, bem como os desvios padrão para as variáveis independentes.
- *ANOVAs Univariadas*. Executa um teste de análise de variância unidirecional de igualdade de médias de grupo para cada variável independente.
- *M de Box*. Um teste para a igualdade das matrizes de covariâncias de grupo. Para amostras suficientemente grandes, um valor p não significativo representa que não há evidência suficiente de que as matrizes diferem. O teste é sensível a partidas da normalidade multivariada.

Coeficientes de Função. As opções disponíveis são coeficientes de classificação de Fisher e coeficientes não padronizados.

- *de Fisher*. Exibe os coeficientes da função de classificação de Fisher que podem ser utilizados diretamente para classificação. Um conjunto separado de coeficientes de função de classificação é obtido para cada grupo, e um caso é designado ao grupo para o qual ele possui o maior escore discriminante (valor da função de classificação).
- *Não padronizado*. Exibe os coeficientes de função discriminante não padronizadas.

Matrizes. As matrizes de coeficientes disponíveis para variáveis independentes são matriz de correlações dentro de grupos, matriz de covariâncias dentro de grupos, matriz de covariâncias separada de grupos e o total de matrizes de covariância.

- *Correlação dentro de grupos*. Exibe uma matriz de correlações dentro de grupos em conjunto que é obtida pela média das matrizes de covariâncias separadas de todos os grupos antes de calcular as correlações.
- *Covariância dentro de grupos*. Exibe uma matriz de covariâncias dentro de grupos em conjunto, que pode diferir da matriz de covariâncias totais. A matriz é obtida pela média das matrizes covariâncias separadas para todos os grupos.
- *Covariância separada de grupos*. Exibe matrizes de covariância separadas para cada grupo.
- *Total de covariâncias*. Exibe uma matriz de covariâncias a partir de todos os casos como se fossem de uma única amostra.

Classificação. A saída a seguir pertence aos resultados de classificação.

- *Resultados do caso*. Os códigos para o grupo real, o grupo predito, as probabilidades posteriores e os escores discriminantes são exibidos para cada caso.
- *Tabela de sumarização*. O número de casos designados correta e incorretamente para cada um dos grupos com base na análise discriminante. Às vezes é chamado de "Matriz de Confusão".
- *Classificação com exclusão de um item*. Cada caso na análise é classificado pelas funções derivadas de todos os outros casos diferentes desse caso. Também é conhecida como "Método U".
- *Mapa territorial*. Um gráfico dos limites utilizados para classificar os casos em grupos com base nos valores de função. Os números correspondem aos grupos nos quais os casos são classificados. A média de cada grupo é indicada por um asterisco dentro de seus limites. O mapa não será exibido se houver apenas uma função discriminante.
- *Grupos combinados*. Cria um gráfico de dispersão de todos os grupos dos dois primeiros valores da função discriminante. Se houver apenas uma função, um histograma será exibido.
- *Separado de grupos*. Cria gráficos de dispersão de grupos separados dos dois primeiros valores da função discriminante. Se houver apenas uma função, histogramas serão exibidos.

Stepwise. A Sumarização de Passos exibe as estatísticas de todas as variáveis após cada passo, e **F para distâncias pairwise** exibe uma matriz de razões *F* de pairwise para cada par de grupos. As razões *F* podem ser utilizadas para testes de significância das distâncias Mahalanobis entre os grupos.

Opções de Progresso do Nó Discriminante

Método. Selecione a estatística a ser utilizada para inserir ou remover novas variáveis. As alternativas disponíveis são lambda de Wilks, variância não explicada, distância de Mahalanobis, razão de *F* menor e *V* de Rao. Com o *V* de Rao, é possível especificar o aumento mínimo em *V* para uma variável a ser inserida.

- *Lambda de Wilks*. Um método de seleção de variáveis para análise discriminante stepwise que escolhe variáveis a serem inseridas na equação com base no quanto elas diminuem o lambda de Wilks. Em cada passo, a variável que minimiza o lambda geral de Wilks é inserida.
- *Variância não explicada*. Em cada passo, a variável que minimiza a soma da variação não explicada entre os grupos é inserida.
- *Distância de Mahalanobis*. A medida do quanto os valores de um caso nas variáveis independentes diferem da média de todos os casos. Uma distância de Mahalanobis grande identifica um caso como tendo valores extremos em uma ou mais variáveis independentes.
- *Razão F menor*. Um método de seleção de variáveis na análise stepwise com base na maximização de uma razão F calculada a partir da distância de Mahalanobis entre os grupos.
- *V de Rao*. A medida das diferenças entre as médias de grupo. Também chamada de rastreio de Lawley-Hotelling. Em cada passo, a variável que maximiza o aumento no V de Rao é inserida. Após selecionar esta opção, insira o valor mínimo que uma variável deve ter para inserir na análise.

Critérios. As alternativas disponíveis são **Usar valor F** e **Usar probabilidade de F**. Insira valores para inserção e remoção de variáveis.

- *Usar valor F*. Uma variável será inserida no modelo se seu valor F for maior que o valor de Entrada e será removida se o valor F for menor que o valor de Remoção. A Entrada deve ser maior que Remoção, e ambos os valores devem ser positivos. Para inserir mais variáveis no modelo, diminua o valor de Entrada. Para remover mais variáveis do modelo, aumente o valor de Remoção.
- *Usar probabilidade de F*. Uma variável será inserida no modelo se o nível de significância de seu valor F for menor que o valor de Entrada e será removida se o nível de significância for maior que o valor de Remoção. A Entrada deve ser menor que Remoção, e ambos os valores devem ser positivos. Para inserir mais variáveis no modelo, aumente o valor de Entrada. Para remover mais variáveis do modelo, diminua o valor de Remoção.

Nugget do Modelo Discriminante

Os nuggets do modelo Discriminante representam as equações estimadas pelos nós Discriminantes. Elas contêm todas as informações capturadas pelo modelo discriminante, bem como informações sobre a estrutura e o desempenho do modelo.

Ao executar um fluxo que contém um nugget do modelo Discriminante, o nó inclui dois novos campos contendo a predição do modelo e a probabilidade associada. Os nomes dos novos campos são derivados do nome do campo de saída que está sendo predito, prefixados com *$D-* para a categoria predita e com *$DP-* para a probabilidade associada. Por exemplo, para um campo de saída denominado *colorpref*, os novos campos serão denominados *$D-colorpref* e *$DP-colorpref*.

Gerando um nó Filtro. O menu Gerar permite criar um novo nó Filtro para transmitir os campos de entrada com base nos resultados do modelo.

Importância do preditor

Opcionalmente, um gráfico que indica a importância relativa de cada preditor na estimativa do modelo também pode ser exibido na guia Modelo. Geralmente você desejará focar seus esforços de modelagem

nos preditores que forem mais importantes e considerar descartar ou ignorar aqueles que forem menos importantes. Observe que este gráfico estará disponível apenas se **Calcular a importância do preditor** estiver selecionada na guia Análise antes de gerar o modelo. Consulte o tópico "Importância do preditor" na página 44 para obter mais informações.

Saída Avançada do Nugget do Modelo Discriminante

A saída avançada para análise discriminante fornece informações detalhadas sobre o modelo estimado e seu desempenho. A maioria das informações contidas na saída avançada é muito técnica e um conhecimento amplo de análise discriminante é necessário para interpretar corretamente esta saída. Consulte o tópico "Opções de Saída do Nó Discriminante" na página 193 para obter mais informações.

Configurações do Nugget do Modelo Discriminante

A guia Configurações em um nugget do modelo Discriminante permite obter os escores de propensão quando escorar o modelo. Esta guia está disponível para modelos apenas com destinos de flag, e somente após o nugget do modelo ter sido incluído em um fluxo.

Calcular escores de propensão bruta. Para modelos com uma resposta de flag (que retornam uma predição de sim ou não), é possível solicitar os escores de propensão que indicam a probabilidade do resultado real especificado para o campo de destino. Esses são um complemento dos outros valores de predição e de confiança que podem ser gerados durante a escoragem.

Calcular escores de propensão ajustada. Os escores de propensão bruta baseiam-se apenas nos dados de treinamento e esses podem ser altamente otimistas devido à tendência de muitos modelos a super ajustar desses dados. As propensões ajustadas tentam compensar ao avaliar o desempenho do modelo com relação à partição de teste ou de validação. Essa opção requer que um campo de partição seja definido no fluxo e que os escores de propensão ajustada sejam ativados no nó de modelagem antes de gerar o modelo.

Gerar SQL para este modelo Ao usar dados de um banco de dados, código SQL pode ser enviado por push de volta para o banco de dados para execução, fornecendo desempenho superior para muitas operações.

Selecione uma das opções a seguir para especificar como a geração de SQL é executada.

- **Padrão: Escorar usando o Server Scoring Adapter (se instalado) no processo** Se conectado a um banco de dados com um adaptador de escoragem instalado, gera a SQL utilizando o adaptador de escoragem e funções definidas pelo usuário (UDF) associadas e escora seu modelo no banco de dados. Quando nenhum adaptador de escoragem estiver disponível, essa opção busca seus dados novamente a partir do banco de dados e os escora no SPSS Modeler.
- **Escorar fora do Banco de dados** Se selecionada, esta opção busca seus dados novamente a partir do banco de dados e os escora no SPSS Modeler.

Sumarização do Nugget do Modelo Discriminante

A guia Sumarização de um nugget do modelo Discriminante exibe os campos e as configurações usados para gerar o modelo. Além disso, se você tiver executado um nó Análise anexado a este nó de modelagem, as informações dessa análise também serão exibidas nesta seção. Para obter informações gerais sobre como utilizar o navegador do modelo, consulte "Procurando Nuggets do Modelo" na página 42.

Nó GenLin

O modelo linear generalizado expande o modelo linear geral para que a variável dependente esteja linearmente relacionada aos fatores e às covariáveis por meio de uma função de ligação especificada. Além disso, o modelo permite à variável dependente ter uma distribuição não normal. Ele cobre modelos estatísticos amplamente utilizados, como regressão linear para respostas distribuídas normalmente, modelos logísticos para dados binários, modelos log-linear para dados de contagem, modelos de log-log

complementares para dados de sobrevivência censurados por intervalo, além de muitos outros modelos estatísticos por meio de sua formulação de modelo amplamente generalizada.

Exemplos. Uma companhia de navegação pode utilizar modelos lineares generalizados para ajustar uma regressão de Poisson para contagens de danos de vários tipos de navios construídos em diferentes períodos de tempo, e o modelo resultante pode ajudar a determinar quais tipos de navios são mais propensos a danos.

Uma empresa de seguros de automóveis pode utilizar modelos lineares generalizados para ajustar uma regressão gama para reclamações de sinistros de automóveis, e o modelo resultante pode ajudar a determinar os fatores que mais contribuem com o tamanho da reclamação.

Os pesquisadores médicos podem utilizar modelos lineares generalizados para ajustar uma regressão log-log complementar de dados de sobrevivência censurados por intervalo para prever o tempo de recorrência de uma condição médica.

Os modelos lineares generalizados funcionam ao construir uma equação que relaciona os valores do campo de entrada com os valores do campo de saída. Quando o modelo é gerado, ele pode ser utilizado para estimar valores para os novos dados. Para cada registro, uma probabilidade de associação é calculada para cada categoria de saída possível. A categoria de destino com a probabilidade mais alta é designada como o valor de saída predito para esse registro.

Requisitos. Um ou mais campos de entrada e exatamente um campo de destino (que pode ter um nível de medição de *Contínuo* ou de *Flag*) com duas ou mais categorias são necessários. Os campos utilizados no modelo devem ter seus tipos totalmente instanciados.

Intensidades. O modelo linear generalizado é extremamente flexível, mas o processo de escolha da estrutura do modelo não é automatizado, o que exige um nível de familiaridade com os dados que não forem necessários pelos algoritmos "caixa preta".

Opções de Campo do Nó GenLin

Além das opções customizadas de destino, de entrada e de partição geralmente oferecidas nas guias Campos do nó de modelagem (consulte “Opções de Campos do Nó de Modelagem” na página 31), o nó GenLin oferece a funcionalidade extra a seguir.

Usar campo de ponderação. O parâmetro de escala é um parâmetro de modelo estimado relacionado à variância da resposta. As ponderações de escala são valores "conhecidos" que podem variar de observação para observação. Se a variável de ponderação de escala for especificada, o parâmetro de escala, que está relacionado à variância da resposta, será dividido por essa variável para cada observação. Os registros com valores de ponderação de escala que forem menores ou iguais a 0 ou estiverem omissos não são utilizados na análise.

O campo de destino representa o número de eventos que ocorrem em um conjunto de avaliações. Quando a resposta for um número de eventos que ocorrem em um conjunto de avaliações, o campo de destino conterá o número de eventos e será possível selecionar uma variável adicional contendo o número de avaliações. Como alternativa, se o número de avaliações for o mesmo em todos os sujeitos, então as avaliações poderão ser especificadas utilizando um valor fixo. O número de avaliações deve ser maior ou igual ao número de eventos para cada registro. Os eventos devem ser números inteiros não negativos e as avaliações devem ser números inteiros positivos.

Opções de Modelo do Nó GenLin

Nome do modelo É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Utilizar dados particionados. Se um campo de partição for definido, essa opção assegurará que apenas os dados da partição de treinamento sejam utilizados para construir o modelo.

Criar modelos de divisão. Constrói um modelo separado para cada valor possível de campos de entrada que são especificados como campos de divisão. Consulte o tópico "Construindo Modelos de Divisão" na página 28 para obter mais informações.

Tipo de modelo. Há duas opções para o tipo de modelo a ser construído. **Apenas efeitos principais** faz com que o modelo inclua apenas os campos de entrada individualmente, e não testar as interações (efeitos multiplicativos) entre os campos de entrada. **Efeitos principais e todas as interações de duas vias** inclui todas as interações de duas vias, bem como os efeitos principais do campo de entrada.

Offset. O termo de offset é um preditor "estrutural". Seu coeficiente não é estimado pelo modelo, mas supõe-se que seu valor seja 1, portanto, os valores do offset são apenas incluídos no preditor linear da resposta. Isso é útil especialmente em modelos de regressão de Poisson, em que cada caso pode ter diferentes níveis de exposição para o evento de interesse.

Por exemplo, ao modelar as taxas de acidentes para motoristas individuais, há uma diferença significativa entre um motorista que causou um acidente em três anos de experiência e um motorista que causou um acidente em 25 anos de experiência! O número de acidentes pode ser modelado como uma resposta de Poisson ou binomial negativo com uma ligação de log, se o log natural da experiência do motorista for incluído como um termo do offset.

Outras combinações de tipos de distribuição e de ligação podem requerer outras transformações da variável de offset.

Nota: se um campo de offset variável for utilizado, o campo especificado também não deverá ser utilizado como uma entrada. Configure o papel para o campo de offset para **Nenhum** em uma origem ou nó Tipo de envio de dados, se necessário.

Categoria base para resposta de flag.

Para resposta binária, é possível escolher a categoria de referência para a variável dependente. Isso pode afetar uma determinada saída, como estimativas de parâmetros e valores salvos, mas não deve alterar o ajuste do modelo. Por exemplo, se sua resposta binária utilizar os valores 0 e 1:

- Por padrão, o procedimento torna a última categoria (valor mais alto), ou 1, a categoria de referência. Nesta situação, as probabilidades de modelo salvas estimam a chance de um determinado caso assumir o valor 0, e as estimativas de parâmetro devem ser interpretadas como relacionadas à probabilidade da categoria 0.
- Se você especificar a primeira categoria (valor mais baixo), ou 0, como a categoria de referência, então as probabilidades de modelo salvas estimarão a chance de um determinado caso assumir o valor 1.
- Se você especificar a categoria customizada e sua variável tiver rótulos definidos, será possível configurar a categoria de referência ao escolher um valor na lista. Isso poderá ser conveniente quando, no meio da especificação de um modelo, você não se lembrar exatamente como uma determinada variável foi codificada.

Incluir intercepto no modelo. O intercepto geralmente é incluído no modelo. Se você conseguir presumir a passagem de dados por meio da origem, será possível excluir o intercepto.

Opções Avançadas do Nó GenLin

Se você tiver conhecimento detalhado dos modelos lineares generalizados, as opções avançadas permitirão fazer um ajuste preciso do processo de treinamento. Para acessar as opções avançadas, configure o **Modo** para **Especialista** na guia Especialista.

Distribuição de Campo de Destino e Função de Ligação

Distribuição.

Esta seleção especifica a distribuição da variável dependente. A capacidade de especificar uma distribuição não normal e uma função de ligação de não identidade é a melhoria essencial do modelo linear generalizado sobre o modelo linear geral. Há muitas combinações de função de distribuição e de ligação possíveis e várias delas podem ser apropriadas para qualquer conjunto de dados específico, portanto, sua opção poderá ser orientada pelas considerações teóricas a priori ou pela combinação que for mais bem adequada.

- **Binomial.** Essa distribuição é apropriada apenas para as variáveis que representarem uma resposta binária ou um número de eventos.
- **Gama.** Essa distribuição é apropriada para as variáveis com valores de escala positivos que são voltados para valores maiores positivos. Se um valor de dados for menor ou igual a 0 ou estiver omisso, então o caso correspondente não será utilizado na análise.
- **Gaussiana inversa.** Essa distribuição é apropriada para as variáveis com valores de escala positivos que são voltados para valores maiores positivos. Se um valor de dados for menor ou igual a 0 ou estiver omisso, então o caso correspondente não será utilizado na análise.
- **Binomial negativo.** Esta distribuição pode ser considerada como o número de avaliações necessárias para observar o *k* sucessos e é apropriada para as variáveis com os valores de número inteiro não negativos. Se um valor de dados for um número não inteiro, menor que 0 ou omisso, então o caso correspondente não será utilizado na análise. O valor fixo do parâmetro auxiliar da distribuição binomial negativa pode ser qualquer número inteiro maior ou igual a 0. Quando o parâmetro auxiliar é configurado para 0, utilizar essa distribuição é equivalente a utilizar a distribuição de Poisson.
- **Normal.** Isso é apropriado para variáveis de escala cujos valores utilizam uma distribuição simétrica bem-formada sobre um valor central (média). A variável dependente deve ser numérica.
- **Poisson.** Esta distribuição pode ser considerada como o número de ocorrências de um evento de interesse em um período de tempo fixo e é apropriada para variáveis com valores de número inteiro não negativos. Se um valor de dados for um número não inteiro, menor que 0 ou omisso, então o caso correspondente não será utilizado na análise.
- **Tweedie.** Essa distribuição é apropriada para variáveis que podem ser representadas pelas combinações de Poissom de distribuições gama; a distribuição é "combinada" no sentido de que ela combina propriedades de distribuições contínuas (aceita valores reais não negativos) e discretas (massa de probabilidade positiva em um valor único, 0). A variável dependente deve ser numérica, com valores de dados maiores ou iguais a zero. Se um valor de dados for menor que zero ou omisso, então o caso correspondente não será utilizado na análise. O valor fixo do parâmetro de distribuição Tweedie pode ser qualquer número maior que um e menor que dois.
- **Multinomial.** Essa distribuição é apropriada para variáveis que representam uma resposta ordinal. A variável dependente pode ser numérica ou de sequência de caracteres e deve ter pelo menos dois valores de dados válidos distintos.

Funções de Ligação.

A função de ligação é uma transformação da variável dependente que permite estimação do modelo. As funções a seguir estão disponíveis:

- **Identidade.** $f(x)=x$. A variável dependente não é transformada. Essa ligação pode ser utilizada com qualquer distribuição.
- **Log-log complementar.** $f(x)=\log(-\log(1-x))$. Isso é apropriado apenas com a distribuição binomial.
- **Cauchit acumulativo.** $f(x) = \tan(\pi\ (x - 0.5))$, aplicado à probabilidade acumulativa de cada categoria da resposta. Isso é apropriado apenas com a distribuição multinomial.
- **Log-log complementar acumulativo.** $f(x)=\ln(-\ln(1-x))$, aplicado à probabilidade acumulativa de cada categoria da resposta. Isso é apropriado apenas com a distribuição multinomial.
- **Logit acumulativo.** $f(x)=\ln(x\ /\ (1-x))$, aplicado à probabilidade acumulativa de cada categoria da resposta. Isso é apropriado apenas com a distribuição multinomial.

- **Log-log negativo acumulativo.** $f(x)=-\ln(-\ln(x))$, aplicado à probabilidade acumulativa de cada categoria da resposta. Isso é apropriado apenas com a distribuição multinomial.
- **Probito acumulativo.** $f(x)=\Phi^{-1}(x)$, aplicado à probabilidade acumulativa de cada categoria da resposta, em que Φ^{-1} é a função de distribuição cumulativa de norma padrão inversa. Isso é apropriado apenas com a distribuição multinomial.
- **Log.** $f(x)=\log(x)$. Essa ligação pode ser utilizada com qualquer distribuição.
- **Log complementar.** $f(x)=\log(1-x)$. Isso é apropriado apenas com a distribuição binomial.
- **Logit.** $f(x)=\log(x\ /\ (1-x))$. Isso é apropriado apenas com a distribuição binomial.
- **Binomial negativo.** $f(x)=\log(x\ /\ (x+k^{-1}))$, em que k é o parâmetro auxiliar da distribuição binomial negativa. Isso é apropriado apenas com a distribuição binomial negativa.
- **Log-log negativo.** $f(x)=-\log(-\log(x))$. Isso é apropriado apenas com a distribuição binomial.
- **Potência das chances.** $f(x)=[(x/(1-x))^{\alpha}-1]/\alpha$, if $\alpha \neq 0$. $f(x)=\log(x)$, se $\alpha=0$. O α é a especificação do número necessário e deve ser um número real. Isso é apropriado apenas com a distribuição binomial.
- **Probito.** $f(x)=\Phi^{-1}(x)$, em que Φ^{-1} é a função de distribuição cumulativa padrão normal inversa. Isso é apropriado apenas com a distribuição binomial.
- **Potência.** $f(x)=x^{\alpha}$, if $\alpha \neq 0$. $f(x)=\log(x)$, if $\alpha=0$. O α é a especificação do número necessário e deve ser um número real. Essa ligação pode ser utilizada com qualquer distribuição.

Parâmetros. Os controles neste grupo permitem especificar os valores de parâmetros quando determinadas opções de distribuição são escolhidas.

- **Parâmetro para binomial negativo.** Para distribuição binomial negativa, escolha um para especificar um valor ou para permitir que o sistema forneça um valor estimado.
- **Parâmetro para Tweedie.** Para distribuição Tweedie, especifique um número entre 1,0 e 2,0 para o valor fixo.

  Estimação de Parâmetro. Os controles nesse grupo permitem especificar métodos de estimação e fornecer valores iniciais para as estimativas de parâmetro.

  – **Método.** É possível selecionar um método de estimação de parâmetros. Escolha entre Newton-Raphson, escoragem de Fisher ou um método híbrido no qual as iterações de escoragem de Fisher são executadas antes de alternar para o método Newton-Raphson. Se a convergência for alcançada durante a fase de escoragem de Fisher do método híbrido antes de o número máximo de iterações Fisher ser alcançado, o algoritmo continua com o método Newton-Raphson.

  – **Método de parâmetro de escala.** É possível selecionar o método de estimação de parâmetro de escala. Em conjunto com a máxima verossimilhança estima o parâmetro de escala com os efeitos do modelo; observe que essa opção não será válida se a resposta tiver um binomial negativo, Poisson, ou distribuição binomial . As opções de deviance e de qui-quadrado de Pearson estimam o parâmetro de escala a partir do valor dessas estatísticas. Como alternativa, é possível especificar um valor fixo para o parâmetro da escala.
- **Matriz de covariâncias.** O estimador baseado em modelo é o negativo da inversa generalizada da matriz Hessiana. O estimador robusto (também chamado de Huber/White/sandwich) é um estimador baseado em modelo "corrigido" que fornece uma estimativa consistente da covariância, mesmo quando a especificação das funções de variância e de ligação estiver incorreta.

Iterações. Essas opções permitem controlar os parâmetros para convergência do modelo. Consulte o tópico “Iterações de Modelos Lineares Generalizados” na página 200 para obter mais informações.

Saída. Essas opções permitem solicitar estatísticas adicionais que serão exibidas na saída avançada do nugget do modelo construído pelo nó. Consulte o tópico “Saída Avançada de Modelos Lineares Generalizados” na página 200 para obter mais informações.

Tolerância à singularidade. Matrizes singulares (ou não inversíveis) possuem colunas linearmente dependentes que podem causar sérios problemas para o algoritmo de estimação. Como até mesmo

matrizes quase singulares podem levar a resultados ruins, o procedimento tratará como singular uma matriz cuja determinante for menor que a tolerância. Especifique um valor positivo.

Iterações de Modelos Lineares Generalizados

É possível configurar os parâmetros de convergência para estimar o modelo linear generalizado.

Iterações. As opções a seguir estão disponíveis:

- **Máximo de iterações.** O número máximo de iterações que o algoritmo executará. Especifique um número inteiro não negativo.
- **Divisão máxima da etapa pela metade.** Em cada iteração, o tamanho do passo é reduzido por um fator de 0,5 até que o log da verossimilhança aumente ou a divisão máxima da etapa pela metade seja atingida. Especifique um número inteiro positivo.
- **Verificar separação de pontos de dados.** Quando selecionada, o algoritmo executa testes para assegurar que as estimativas do parâmetro tenham valores exclusivos. A separação ocorre quando o procedimento pode produzir um modelo que classifica corretamente cada caso. Esta opção está disponível para respostas binomiais com formato binário .

Critérios de Convergência. As opções a seguir estão disponíveis

- **Convergência de parâmetro.** Quando selecionada, o algoritmo é interrompido após uma iteração na qual uma mudança absoluta ou relativa nas estimativas de parâmetro for menor que o valor especificado, que deve ser positivo.
- **Convergência de log da verossimilhança.** Quando selecionada, o algoritmo é interrompido após uma iteração na qual uma mudança absoluta ou relativa na função de log da verossimilhança for menor que o valor especificado, que deve ser positivo.
- **Convergência da Hessiana.** Para uma especificação Absoluta, a convergência será assumida se uma estatística baseada na convergência da Hessiana for menor que o valor positivo especificado. Para a especificação Relativa, a convergência será assumida se a estatística for menor que o produto entre o valor positivo especificado e o valor absoluto do log da verossimilhança.

Saída Avançada de Modelos Lineares Generalizados

Selecione a saída opcional que deseja exibir na saída avançada do nugget do modelo linear generalizado. Para visualizar a saída avançada, procure o nugget do modelo e clique na guia **Avançado**. Consulte o tópico “Saída Avançada do Nugget do Modelo GenLin” na página 202 para obter mais informações.

A saída a seguir está disponível:

- **Sumarização do processamento de caso.** Exibe o número e a porcentagem de casos incluídos e excluídos da análise e da tabela Sumarização de Dados Correlacionados.
- **Estatísticas descritivas.** Exibe estatísticas descritivas e informações de sumarização sobre a variável dependente, covariáveis e fatores.
- **Informações de modelo.** Exibe o nome do conjunto de dados, a variável dependente ou as variáveis de eventos e de avaliações, a variável de offset, a variável de ponderação de escala, a distribuição de probabilidade e a função de ligação.
- **Estatísticas de Qualidade do ajuste.** Exibe o deviance e o deviance escalado, o qui-quadrado de Pearson e o qui-quadrado de Pearson escalado, log da verossimilhança, o Akaike Information Criterion (AIC), o Finite Sample Corrected AIC (AICC), o Critério de Informação Bayesiano (BIC) e o Consistent AIC (CAIC).
- **Estatísticas de sumarização do modelo.** Exibe testes de ajuste do modelo, incluindo estatísticas de razão de verossimilhança para o teste de omnibus de ajuste do modelo e estatísticas para contrastes do Tipo I ou III para cada efeito.
- **Estimativas de parâmetro.** Exibe estimativas de parâmetro, estatísticas do teste correspondentes e intervalos de confiança. Opcionalmente, é possível exibir estimativas de parâmetro exponeciado e também estimativas de parâmetro bruto.

- **Matriz de covariâncias para estimativas de parâmetro.** Exibe a matriz de covariâncias de parâmetro estimado.
- **Matriz de correlações para estimativas de parâmetro.** Exibe a matriz de correlações de parâmetro estimado.
- **Matrizes de coeficiente de contraste (L).** Exibe coeficientes de contraste para os efeitos padrão e para as médias marginais estimadas, se solicitado na guia Médias de EM.
- **Funções estimáveis gerais.** Exibe as matrizes para gerar as matrizes de coeficiente de contraste (L).
- **Histórico de iteração.** Exibe o histórico de iteração para as estimativas de parâmetro e para o log da verossimilhança e imprime a última avaliação do vetor gradiente e da matriz Hessiana. A tabela de históricos de iteração exibe estimativas de parâmetro para cada $n^{ésima}$ iteração, começando com a $0^{ésima}$ iteração (as estimativas iniciais), em que *n* é o valor do intervalo de impressão. Se o histórico de iteração for solicitado, então a última iteração será sempre exibida, independentemente de *n*.
- **Teste de multiplicadores de Lagrange.** Exibe as estatísticas do teste de multiplicadores de Lagrange para avaliar a validade de um parâmetro de escala que é calculado utilizando o deviance ou o qui-quadrado de Pearson, ou para configurar, a um número fixo, as distribuições normais, de gama e Gaussiana inversa. Para a distribuição binomial negativa, isso testa o parâmetro auxiliar fixo.

Efeitos do Modelo. As opções a seguir estão disponíveis:

- **Tipo de análise.** Especifique o tipo de análise a produzir. A análise Tipo I é apropriada geralmente quando você tiver motivos a priori para ordenar preditores no modelo, ao passo que o Tipo III é mais geralmente aplicável. As estatísticas Wald ou de razão de verossimilhança são calculadas com base na seleção feita no grupo Estatísticas Qui-quadrado.
- **Intervalos de confiança.** Especifique um nível de confiança maior que 50 e menor que 100. Os intervalos de Wald baseiam-se na suposição de que os parâmetros têm uma distribuição normal assintótica; os intervalos de probabilidade de perfil são mais precisos, mas podem ser dispendiosos em termos computacionais. O nível de tolerância para intervalos de probabilidade de perfil é o critério utilizado para parar o algoritmo iterativo utilizado para calcular os intervalos.
- **Função de log da verossimilhança.** Isso controla o formato de exibição da função de log-verossimilhança. A função completa inclui um termo adicional que é constante com relação às estimativas de parâmetro; ela não tem efeito sobre a estimativa de parâmetro e é deixada fora da exibição em alguns produtos de software.

Nugget do Modelo GenLin

Um nugget do modelo GenLin representa as equações estimadas por um nó GenLin. Elas contêm todas as informações capturadas pelo modelo, bem como informações sobre a estrutura e o desempenho do modelo.

Ao executar um fluxo que contém um nugget do modelo GenLin, o nó inclui novos campos cujo conteúdo depende da natureza do campo de destino:

- **Campo de flag.** Inclui campos contendo a categoria predita, a probabilidade associada e as probabilidades de cada categoria. Os nomes dos dois primeiros novos campos são derivados do nome do campo de saída que está sendo predito, prefixados com *$G-* para a categoria predita e com *$GP-* para a probabilidade associada. Por exemplo, para um campo de saída denominado *default*, os novos campos serão denominados *$G-default* e *$GP-default*. Os dois últimos campos adicionais são nomeados com base nos valores do campo de saída, prefixados por *$GP-*. Por exemplo, se os valores legais de *default* forem *Yes* e *No*, os novos campos serão denominados *$GP-Yes* e *$GP-No*.
- **Variável resposta contínua.** Inclui campos contendo a média predita e o erro padrão.
- **Variável resposta contínua, representando o número de eventos em uma série de avaliações.** Inclui campos contendo a média predita e o erro padrão.

- **Resposta ordinal.** Inclui campos contendo a categoria predita e a probabilidade associada para cada valor do conjunto ordenado. Os nomes dos campos são derivados do valor do conjunto ordenado que está sendo predito, prefixados com *$G-* para a categoria predita e com *$GP-* para a probabilidade associada.

Gerando um nó Filtro. O menu Gerar permite criar um novo nó Filtro para transmitir os campos de entrada com base nos resultados do modelo.

Importância do preditor

Opcionalmente, um gráfico que indica a importância relativa de cada preditor na estimativa do modelo também pode ser exibido na guia Modelo. Geralmente você desejará focar seus esforços de modelagem nos preditores que forem mais importantes e considerar descartar ou ignorar aqueles que forem menos importantes. Observe que este gráfico estará disponível apenas se **Calcular a importância do preditor** estiver selecionada na guia Análise antes de gerar o modelo. Consulte o tópico "Importância do preditor" na página 44 para obter mais informações.

Saída Avançada do Nugget do Modelo GenLin

A saída avançada para o modelo linear generalizado fornece informações detalhadas sobre o modelo estimado e seu desempenho. A maioria das informações contidas na saída avançada é muito técnica e um conhecimento amplo desse tipo de análise é necessário para interpretar corretamente esta saída. Consulte o tópico "Saída Avançada de Modelos Lineares Generalizados" na página 200 para obter mais informações.

Configurações do Nugget do Modelo GenLin

A guia Configurações de um nugget do modelo GenLin permite obter os escores de propensão quando escorar o modelo, e também para a geração de SQL durante a escoragem do modelo. Esta guia está disponível para modelos apenas com destinos de flag, e somente após o nugget do modelo ter sido incluído em um fluxo.

Calcular escores de propensão bruta. Para modelos com uma resposta de flag (que retornam uma predição de sim ou não), é possível solicitar os escores de propensão que indicam a probabilidade do resultado real especificado para o campo de destino. Esses são um complemento dos outros valores de predição e de confiança que podem ser gerados durante a escoragem.

Calcular escores de propensão ajustada. Os escores de propensão bruta baseiam-se apenas nos dados de treinamento e esses podem ser altamente otimistas devido à tendência de muitos modelos a super ajustar desses dados. As propensões ajustadas tentam compensar ao avaliar o desempenho do modelo com relação à partição de teste ou de validação. Essa opção requer que um campo de partição seja definido no fluxo e que os escores de propensão ajustada sejam ativados no nó de modelagem antes de gerar o modelo.

Gerar SQL para este modelo Ao usar dados de um banco de dados, código SQL pode ser enviado por push de volta para o banco de dados para execução, fornecendo desempenho superior para muitas operações.

Selecione uma das opções a seguir para especificar como a geração de SQL é executada.

- **Padrão: Escorar usando o Server Scoring Adapter (se instalado) no processo** Se conectado a um banco de dados com um adaptador de escoragem instalado, gera a SQL utilizando o adaptador de escoragem e funções definidas pelo usuário (UDF) associadas e escora seu modelo no banco de dados. Quando nenhum adaptador de escoragem estiver disponível, essa opção busca seus dados novamente a partir do banco de dados e os escora no SPSS Modeler.
- **Escorar fora do Banco de dados** Se selecionada, esta opção busca seus dados novamente a partir do banco de dados e os escora no SPSS Modeler.

Sumarização do Nugget do Modelo GenLin

A guia Sumarização de um nugget do modelo GenLin exibe os campos e as configurações usados para gerar o modelo. Além disso, se você tiver executado um nó Análise anexado a este nó de modelagem, as informações dessa análise também serão exibidas nesta seção. Para obter informações gerais sobre como utilizar o navegador do modelo, consulte "Procurando Nuggets do Modelo" na página 42.

Modelos Mistos Lineares Generalizados

Nó GLMM

Utilize esse nó para criar um modelo linear generalizado misto (GLMM).

modelos lineares generalizados mistos

Os modelos lineares generalizados mistos estendem o modelo linear para que:

- A resposta esteja linearmente relacionada aos fatores e covariáveis por meio de uma função de ligação especificada.
- A resposta possa ter uma distribuição não normal.
- As observações possam ser correlacionadas.

Os modelos lineares generalizados mistos abrangem uma ampla variedade de modelos, desde regressão linear simples até modelos multiníveis complexos para dados longitudinais não normais.

Exemplos. A diretoria de uma escola pode utilizar um modelo linear generalizado misto para determinar se um método de ensino experimental é eficiente para melhorar os escores de matemática. Os alunos da mesma classe devem ser correlacionados, já que eles são ensinados pelo mesmo professor, e as classes da mesma escola também podem ser correlacionadas, de modo que podemos incluir efeitos aleatórios em níveis de escola e de classe para considerar diferentes origens de variabilidade. Consulte o tópico para obter mais informações.

Os pesquisadores médicos podem utilizar um modelo linear generalizado misto para determinar se uma nova medicação anticonvulsiva pode reduzir a taxa de ataques epiléticos de um paciente. As medições repetidas do mesmo paciente são geralmente correlacionadas positivamente de modo que um modelo combinado com alguns efeitos aleatórios deve ser apropriado. O campo de destino, nesse caso, o número de ataques, aceita valores de número inteiro positivo, de modo que um modelo linear generalizado misto com uma distribuição de Poisson e ligação de log possam ser apropriados. Consulte o tópico para obter mais informações.

Os executivos de um provedor de serviços de TV, telefone e internet a cabo podem utilizar um modelo linear generalizado misto para saber mais sobre possíveis clientes. Como as possíveis respostas possuem níveis de medição nominais, o analista da empresa utiliza um modelo logit combinado generalizado com um intercepto aleatório para capturar a correlação entre as respostas das questões de uso de serviço entre os tipos de serviço (TV, telefone, Internet) e as respostas de um determinado respondente de pesquisa de opinião. Consulte o tópico para obter mais informações.

A guia Estrutura de Dados permite especificar os relacionamentos estruturais entre os registros em seu conjunto de dados quando as observações forem correlacionadas. Se os registros no conjunto de dados representarem observações independentes, não será necessário especificar nada nessa guia.

Assuntos. A combinação de valores dos campos categóricos especificados deve definir exclusivamente sujeitos no conjunto de dados. Por exemplo, um único campo *ID do Paciente* deve ser suficiente para definir sujeitos em um hospital único, mas a combinação de *ID do Hospital* e *ID do Paciente* poderá ser necessária se os números de identificação do paciente não forem exclusivos entre os hospitais. Em uma configuração de medidas repetidas, várias observações são registradas para cada sujeito, de modo que cada sujeito poderá ocupar diversos registros no conjunto de dados.

Um **sujeito** é uma unidade de observação que pode ser considerada independente de outros sujeitos. Por exemplo, as leituras de pressão arterial de um paciente em um estudo médico podem ser consideradas independentes das leituras de outros pacientes. A definição de sujeitos se torna particularmente importante quando houver medidas repetidas por sujeito e você quiser modelar a correlação entre essas observações. Por exemplo, você pode desejar que as leituras da pressão arterial de um determinado paciente sejam correlacionadas durante os retornos consecutivos ao médico.

Todos os campos especificados como Assuntos na guia Estrutura de Dados são utilizados para definir sujeitos para a estrutura de covariâncias de resíduo e fornecem a lista de possíveis campos para definir sujeitos para estruturas de covariâncias de efeitos aleatórios no Bloco de Efeito Aleatório.

Medidas repetidas. Os campos especificados aqui são utilizados para identificar observações repetidas. Por exemplo, uma única variável *Semana* pode identificar as 10 semanas de observações em um estudo médico, ou *Mês* e *Dia* podem ser utilizados juntos para identificar observações diárias ao longo de um ano.

Definir grupos de covariâncias por. Os campos categóricos especificados aqui definem conjuntos independentes de parâmetros de covariância de efeitos repetidos, um para cada categoria definida pela classificação cruzada dos campos de agrupamento. Todos os sujeitos possuem o mesmo tipo de covariância e sujeitos no mesmo agrupamento de covariância terão os mesmos valores para os parâmetros.

Tipo de covariância repetida. Especifica a estrutura de covariâncias para os resíduos. As estruturas disponíveis são:

- Autorregressivo de primeira ordem (AR1)
- Média móvel autorregressiva (1,1) (ARMA11)
- Simetria composta
- Diagonal
- Identidade dimensionada
- Toeplitz
- Não estruturado
- Componente de variância

Resposta: Essas configurações definem a resposta, sua distribuição e seu relacionamento com os preditores por meio da função de ligação.

Resposta. A resposta é necessária. Ela pode ter qualquer nível de medição e o nível de medição da resposta restringe quais distribuições e funções de ligação são apropriadas.

- **Usar número de avaliações como denominador.** Quando a resposta for um número de eventos que ocorrem em um conjunto de avaliações, o campo de destino contém o número de eventos e é possível selecionar um campo adicional contendo o número de avaliações. Por exemplo, ao testar um novo pesticida, é possível expor amostras de formigas para diferentes concentrações desse pesticida e, em seguida, registrar o número de formigas mortas e o número de formigas em cada amostra. Neste caso, o campo que registra o número de formigas mortas deve ser especificado como o campo de destino (eventos) e o campo que registra o número de formigas em cada amostra deve ser especificado como o campo de avaliações. Se o número de formigas for o mesmo para cada amostra, então o número de avaliações poderá ser especificado utilizando um valor fixo.

  O número de avaliações deve ser maior ou igual ao número de eventos para cada registro. Os eventos devem ser números inteiros não negativos e as avaliações devem ser números inteiros positivos.
- **Customizar categoria de referência.** Para uma variável resposta categórica, é possível escolher a categoria de referência. Isso pode afetar uma determinada saída, como as estimativas de parâmetro, mas ela não deverá alterar o ajuste do modelo. Por exemplo, se a sua resposta assumir os valores 0, 1 e 2, por padrão, o procedimento tornará a última categoria (de valor mais alto), ou 2, a categoria de

referência. Nesta situação, as estimativas de parâmetro devem ser interpretadas como relacionando a probabilidade da categoria 0 ou 1 *relativo* à probabilidade da categoria 2. Se você especificar uma categoria customizada e sua resposta tiver rótulos definidos, será possível configurar a categoria de referência ao escolher um valor na lista. Isso poderá ser conveniente quando, no meio da especificação de um modelo, você não se lembrar exatamente como um campo específico foi codificado.

Distribuição e Relacionamento de Resposta (Ligação) com o Modelo Linear. Dados os valores dos preditores, o modelo espera que a distribuição de valores de resposta siga o formato especificado e que os valores de resposta estejam linearmente relacionados aos preditores por meio da função de ligação especificada. Atalhos para vários modelos comuns são fornecidos, ou escolha uma configuração **Customizado** se houver uma combinação específica de distribuição e de função de ligação que você deseja ajustar e que não estiver na lista breve.

- **Modelo linear.** Especifica uma distribuição normal com uma ligação de identidade, que é útil quando a resposta puder ser predita usando um modelo de regressão linear ou ANOVA.
- **Regressão gama.** Especifica uma distribuição Gama com uma ligação de log, que deve ser usada quando a resposta contiver todos os valores positivos e for voltada para valores maiores.
- **Log-linear.** Especifica uma distribuição de Poisson com uma ligação de log, que deve ser usada quando a resposta representar uma contagem de ocorrências em um período de tempo fixo.
- **Regressão binomial negativa.** Especifica uma distribuição binomial negativa com uma ligação de log, que deve ser usada quando a resposta e o denominador representarem o número de avaliações necessárias para observar *k* sucessos.
- **Regressão logística multinomial.** Especifica uma distribuição multinomial, que deve ser usada quando a resposta for uma resposta de diversas categorias. Usa uma ligação logit acumulativa (resultados ordinais) ou uma ligação logit generalizada (respostas nominais de diversas categorias).
- **Regressão logística binária.** Especifica uma distribuição binomial com uma ligação logit, que deve ser usada quando a resposta for uma resposta binária predita por um modelo de regressão logística.
- **Probito binário.** Especifica uma distribuição binomial com uma ligação probito, que deve ser usada quando a resposta for uma resposta binária com uma distribuição normal subjacente.
- **Sobrevivência censurada por intervalo.** Especifica uma distribuição binomial com uma ligação log-log complementar, que é útil em análises de sobrevivência quando algumas observações não possuírem um evento de encerramento.

Distribuição

Esta seleção especifica a distribuição da resposta. A capacidade de especificar uma distribuição não normal e uma função de ligação de não identidade é a melhoria essencial do modelo linear generalizado misto sobre o modelo linear combinado. Há muitas combinações de função de distribuição e de ligação possíveis e várias delas podem ser apropriadas para qualquer conjunto de dados específico, portanto, sua opção poderá ser orientada pelas considerações teóricas a priori ou pela combinação que for mais bem adequada.

- **Binomial.** Essa distribuição é apropriada apenas para uma resposta que representa uma resposta binária ou um número de eventos.
- **Gama.** Essa distribuição é apropriada para uma resposta com valores de escala positivos que são voltados para valores maiores positivos. Se um valor de dados for menor ou igual a 0 ou estiver omisso, então o caso correspondente não será utilizado na análise.
- **Gaussiana inversa.** Essa distribuição é apropriada para uma resposta com valores de escala positivos que são voltados para valores maiores positivos. Se um valor de dados for menor ou igual a 0 ou estiver omisso, então o caso correspondente não será utilizado na análise.
- **Multinomial.** Essa distribuição é adequada para uma resposta que representa uma resposta de diversas categorias. O formato do modelo dependerá do nível de medição da resposta.

  Uma **resposta nominal** resultará em um modelo multinomial nominal em que um conjunto separado de parâmetros de modelo é estimado para cada categoria da resposta (exceto a categoria de referência).

As estimativas de parâmetro para um determinado preditor mostram o relacionamento entre esse preditor e a probabilidade de cada categoria da resposta, com relação à categoria de referência.

Uma **resposta ordinal** resultará em um modelo multinomial ordinal na qual o termo de intercepção tradicional é substituído por um conjunto de parâmetros de **limite** que estão relacionados à probabilidade acumulativa das categorias de resposta.

- **Binomial negativo.** A regressão binomial negativa usa uma distribuição binomial negativa com uma ligação de log, que deve ser usada quando a resposta representar uma contagem de ocorrências com alta variância.
- **Normal.** Isso é apropriado para uma variável resposta contínua cujos valores utilizam uma distribuição simétrica bem-formada sobre um valor central (média).
- **Poisson.** Esta distribuição pode ser considerada como o número de ocorrências de um evento de interesse em um período de tempo fixo e é apropriada para variáveis com valores de número inteiro não negativos. Se um valor de dados for um número não inteiro, menor que 0 ou omisso, então o caso correspondente não será utilizado na análise.

Funções de Ligação

A função de ligação é uma transformação da resposta que permite a estimativa do modelo. As funções a seguir estão disponíveis:

- **Identidade.** $f(x)=x$. A resposta não é transformada. Essa ligação pode ser utilizada com qualquer distribuição, exceto a multinomial.
- **Log-log complementar.** $f(x)=\log(-\log(1-x))$. Isso é apropriado apenas com a distribuição binomial ou multinomial.
- **Cauchit.** $f(x) = \tan(\pi\ (x - 0.5))$. Isso é apropriado apenas com a distribuição binomial ou multinomial.
- **Log.** $f(x)=\log(x)$. Essa ligação pode ser utilizada com qualquer distribuição, exceto a multinomial.
- **Log complementar.** $f(x)=\log(1-x)$. Isso é apropriado apenas com a distribuição binomial.
- **Logit.** $f(x)=\log(x\ /\ (1-x))$. Isso é apropriado apenas com a distribuição binomial ou multinomial.
- **Log-log negativo.** $f(x)=-\log(-\log(x))$. Isso é apropriado apenas com a distribuição binomial ou multinomial.
- **Probito.** $f(x)=\Phi^{-1}(x)$, em que Φ^{-1} é a função de distribuição cumulativa padrão normal inversa. Isso é apropriado apenas com a distribuição binomial ou multinomial.
- **Potência.** $f(x)=x^{\alpha}$, se $\alpha \neq 0$. $f(x)=\log(x)$, se $\alpha=0$. O α é a especificação do número necessário e deve ser um número real. Essa ligação pode ser utilizada com qualquer distribuição, exceto a multinomial.

Efeitos Fixos: Os fatores de efeitos fixos geralmente são considerados como campos cujos valores de interesse são todos representados no conjunto de dados e podem ser usados para escoragem. Por padrão, os campos com um papel de entrada predefinido que não estiverem especificados em nenhum outro lugar no diálogo são inseridos na parte de efeitos fixos do modelo. Os campos categóricos (flag, nominal e ordinal) que são utilizados como fatores nos campos contínuos e de modelo são utilizados como covariáveis.

Insira os efeitos no modelo ao selecionar um ou mais campos na lista de origem e arrastá-los para a lista de efeitos. O tipo de efeito criado depende do hotspot que você soltar na seleção.

- **Principal.** Os campos eliminados aparecem como efeitos principais separados na parte inferior da lista de efeitos.
- **Duas vias.** Todos os pares possíveis dos campos eliminados aparecem como interações de duas vias na parte inferior da lista de efeitos.
- **Três vias.** Todos os trios possíveis dos campos eliminados aparecem como interações de três vias na parte inferior da lista de efeitos.
- ***.** A combinação de todos os campos eliminados aparece como uma interação única na parte inferior da lista de efeitos.

Os botões à direita do Construtor de Efeito permitem executar várias ações.

Tabela 10. Descrições de botão do construtor de efeito.

Ícone	Descrição
	Exclui termos do modelo de efeitos fixos ao selecionar os termos que você deseja excluir e clicar no botão Excluir.
	Reordena os termos dentro do modelo de efeitos fixos ao selecionar os termos que você deseja reordenar e clicar na seta para cima ou para baixo.
	Inclui termos aninhados no modelo utilizando o diálogo do "Incluir um Termo Customizado" ao clicar no botão Incluir um termo customizado.

Incluir Intercepto. O intercepto é geralmente incluído no modelo. Se você conseguir presumir a passagem de dados por meio da origem, será possível excluir o intercepto.

Incluir um Termo Customizado: É possível construir termos aninhados para seu modelo neste procedimento. Os termos aninhados são úteis para modelar o efeito de um fator ou covariável cujos valores não interagem com os níveis de outro fator. Por exemplo, uma rede de supermercados pode seguir os hábitos de gastos de seus clientes em vários locais de loja. Como cada cliente frequenta apenas um desses locais, pode-se dizer que o efeito *Cliente* é **aninhado no** efeito *local da Loja*.

Além disso, é possível incluir os efeitos de interação, como termos polinomiais envolvendo a mesma covariável, ou incluir diversos níveis de aninhamento no termo aninhado.

Limitações. Os termos aninhados possuem as restrições a seguir:

- Todos os fatores em uma interação devem ser exclusivos. Assim, se *A* for um fator, então especificar *A*A* será inválido.
- Todos os fatores dentro de um efeito aninhado devem ser exclusivos. Assim, se *A* for um fator, então especificar *A(A)* será inválido.
- Nenhum efeito pode ser aninhado dentro de uma covariável. Assim, se *A* for um fator e *X* uma covariável, então especificar *A(X)* será inválido.

Construindo um termo aninhado

1. Selecione um fator ou uma covariável que esteja aninhada dentro de outro fator e, em seguida, clique no botão de seta.
2. Clique em **(Dentro)**.
3. Selecione o fator dentro do qual o fator ou covariável anterior está aninhada e, em seguida, clique no botão de seta.
4. Clique em **Incluir Termo**.

Opcionalmente, é possível incluir efeitos de interações ou incluir diversos níveis de aninhamento no termo aninhado.

Efeitos Aleatórios: Os fatores de efeitos aleatórios são campos cujos valores no arquivo de dados podem ser considerados uma amostra aleatória de uma população maior de valores. Eles são úteis para explicar a variabilidade em excesso na resposta. Por padrão, se você tiver selecionado mais de um sujeito na guia Estrutura de Dados, um bloco de Efeito Aleatório será criado para cada sujeito além do sujeito mais interno. Por exemplo, se você selecionou Escola, Classe e Aluno como sujeitos na guia Estrutura de Dados, os seguintes blocos de efeito aleatório serão criados automaticamente:

- Efeito Aleatório 1: o sujeito é escola (sem efeitos, apenas interceptos)
- Efeito Aleatório 2: o sujeito é escola * classe (sem efeitos, apenas interceptos)

É possível trabalhar com blocos de efeitos aleatórios das seguintes maneiras:

1. Para incluir um novo bloco, clique em **Incluir Bloco...**. Isso abre o diálogo "Bloco de Efeito Aleatório"
2. Para editar um bloco existente, selecione o bloco que deseja editar e clique em **Editar Bloco...** Este abre o diálogo "Bloco de Efeito Aleatório"
3. Para excluir um ou mais blocos, selecione os blocos que deseja excluir e clique no botão Excluir.

Bloco de Efeito Aleatório: Insira os efeitos no modelo ao selecionar um ou mais campos na lista de origem e arrastá-los para a lista de efeitos. O tipo de efeito criado depende do hotspot que você soltar na seleção. Os campos categóricos (flag, nominal e ordinal) que são utilizados como fatores nos campos contínuos e de modelo são utilizados como covariáveis.

- **Principal.** Os campos eliminados aparecem como efeitos principais separados na parte inferior da lista de efeitos.
- **Duas vias.** Todos os pares possíveis dos campos eliminados aparecem como interações de duas vias na parte inferior da lista de efeitos.
- **Três vias.** Todos os trios possíveis dos campos eliminados aparecem como interações de três vias na parte inferior da lista de efeitos.
- ***.** A combinação de todos os campos eliminados aparece como uma interação única na parte inferior da lista de efeitos.

Os botões à direita do Construtor de Efeito permitem executar várias ações.

Tabela 11. Descrições de botão do construtor de efeito.

Ícone	Descrição
	Exclui termos do modelo ao selecionar os termos que você deseja excluir e clicar no botão Excluir.
	Reordena os termos dentro do modelo ao selecionar os termos que você deseja reordenar e clicar na seta para cima ou para baixo.
	Inclui termos aninhados no modelo utilizando o diálogo do "Incluir um Termo Customizado" na página 207 ao clicar no botão Incluir um termo customizado.

Incluir Intercepto. O intercepto não é incluído no modelo de efeitos aleatórios por padrão. Se você conseguir presumir a passagem de dados por meio da origem, será possível excluir o intercepto.

Definir grupos de covariâncias por. Os campos categóricos especificados aqui definem conjuntos independentes de parâmetros de covariância de efeitos aleatórios, um para cada categoria definida pela classificação cruzada dos campos de agrupamento. Um conjunto diferente de campos de agrupamento pode ser especificado para cada bloco de efeito aleatório. Todos os sujeitos possuem o mesmo tipo de covariância e sujeitos no mesmo agrupamento de covariância terão os mesmos valores para os parâmetros.

Combinação de sujeito. Permite especificar os sujeitos de efeito aleatório de combinações pré-configuradas de sujeitos na guia Estrutura de Dados. Por exemplo, se *Escola*, *Classe* e *Aluno* forem definidos respectivamente como sujeitos na guia Estrutura de Dados, a lista suspensa de combinação de Assunto terá **Nenhum**, **Escola**, **Escola * Classe** e **Escola * Classe * Aluno** como opções.

Tipo de covariância de efeito aleatório. Especifica a estrutura de covariâncias para os resíduos. As estruturas disponíveis são:

- Autorregressivo de primeira ordem (AR1)
- Média móvel autorregressiva (1,1) (ARMA11)
- Simetria composta
- Diagonal
- Identidade dimensionada
- Toeplitz
- Não estruturado
- Componente de variância

Ponderação e Offset: Ponderação de análise. O parâmetro de escala é um parâmetro de modelo estimado relacionado à variância da resposta. As ponderações de análise são valores "conhecidos" que podem variar de observação para observação. Se o campo de ponderação de análise for especificado, o parâmetro de escala, que está relacionado à variância da resposta, será dividido pelos valores de ponderação de análise para cada observação. Os registros com valores de ponderação de análise que forem menores ou iguais a 0 ou estiverem omissos não são utilizados na análise.

Offset. O termo de offset é um preditor "estrutural". Seu coeficiente não é estimado pelo modelo, mas supõe-se que seu valor seja 1, portanto, os valores do offset são apenas incluídos no preditor linear da resposta. Isso é útil especialmente em modelos de regressão de Poisson, em que cada caso pode ter diferentes níveis de exposição para o evento de interesse.

Por exemplo, ao modelar as taxas de acidentes para motoristas individuais, há uma diferença significativa entre um motorista que causou um acidente em três anos de experiência e um motorista que causou um acidente em 25 anos de experiência! O número de acidentes pode ser modelado como uma resposta de Poisson ou binomial negativo com uma ligação de log, se o log natural da experiência do motorista for incluído como um termo do offset.

Outras combinações de tipos de distribuição e de ligação podem requerer outras transformações da variável de offset.

Opções de Criação Gerais: Essas seleções especificam critérios um pouco mais avançados utilizados para construir o modelo.

Ordenação. Esses controles determinam a ordem das categorias para a resposta e os fatores (entradas categóricas) para propósitos de determinar a "última" categoria. A configuração da ordenação de resposta será ignorada se a resposta não for categórica ou se uma categoria de referência customizada estiver especificada nas configurações do “Resposta” na página 204.

Regras de Parada. É possível especificar o número máximo de iterações que o algoritmo executará. O algoritmo utiliza um processo duplamente iterativo que consiste em um loop interno e em um loop externo. O valor que é especificado para o número máximo de iterações se aplica a ambos os loops. Especifique um número inteiro não negativo. O padrão é 100.

Configurações de Pós-Estimação. Essas configurações determinam como alguma das saídas de modelo é calculada para visualização.

- **Nível de confiança.** Este é o nível de confiança utilizado para calcular estimativas de intervalo dos coeficientes do modelo. Especifique um valor maior que 0 e menor que 100. O padrão é 95.
- **Graus de liberdade.** Especifica como graus de liberdade são calculados para testes de significância. Escolha **Fixo para todos os testes (método Residual)** se o tamanho da amostra for suficientemente grande, ou se os dados estiverem balanceados ou se o modelo utilizar um tipo de covariância mais simples, por exemplo, identidade ou diagonal escalado. Este é o padrão. Escolha **Variados entre os**

testes (aproximação Satterthwaite) se o tamanho da amostra for pequeno, ou se os dados estiverem desbalanceados ou se o modelo utilizar um tipo de covariância complicado, por exemplo, não estruturado.

- **Testes de efeitos e coeficientes fixos.** Este é o método para calcular a matriz de covariância de estimativas de parâmetro. Escolha a estimativa robusta se você estiver preocupado que as suposições de modelo sejam violadas.

Estimação: O algoritmo de construção de modelo utiliza um processo duplamente iterativo que consiste em um loop interno e em um loop externo. As seguintes configurações se aplicam ao loop interno.

Convergência de Parâmetro.
A convergência será assumida se a mudança máxima absoluta ou a mudança máxima relativa nas estimativas de parâmetro for menor que o valor especificado, que deve ser não negativo. O critério não será utilizado se o valor especificado for igual a 0.

Convergência de log da verossimilhança.
A convergência será assumida se a mudança absoluta ou a mudança relativa na função de log da verossimilhança for menor que o valor especificado, que deve ser não negativo. O critério não será utilizado se o valor especificado for igual a 0.

Convergência da Hessiana.
Para uma especificação **Absoluta**, a convergência será assumida se uma estatística baseada na Hessiana for menor que o valor especificado. Para a especificação **Relativa**, a convergência será assumida se a estatística for menor que o produto do valor especificado e do valor absoluto do log da verossimilhança. O critério não será utilizado se o valor especificado for igual a 0.

Máximo de passos de escoragem de Fisher.
Especifique um número inteiro não negativo. Um valor 0 especifica o método Newton-Raphson. Valores maiores que 0 especificam utilizar o algoritmo de escoragem de Fisher até o número *n* de iteração, em que *n* é o número inteiro especificado e, a partir daí, o Newton-Raphson.

Tolerância à singularidade.
Este valor é utilizado como a tolerância na verificação de singularidade. Especifique um valor positivo.

Nota: Por padrão, a Convergência de Parâmetro é utilizada, em que a mudança máxima **Absoluta** em uma tolerância de 1E-6 é verificada. Essa configuração pode produzir resultados que diferem dos resultados que são obtidos em versões anteriores à versão 22. Para produzir resultados de versões anteriores a 22, utilize **Relativo** para o critério de Convergência de Parâmetro e mantenha o valor de tolerância padrão de 1E-6.

Geral: Nome do Modelo. É possível gerar o nome do modelo automaticamente com base nos campos de destino ou especificar um nome customizado. O nome gerado automaticamente é o nome do campo de destino. Se houver diversos destinos, então o nome do modelo será os nomes do campo na ordem, ligados pelo e comercial (símbolo &). Por exemplo, se *field1 field2 e field3* forem os destinos, então o nome do modelo será: *field1 & field2 & field3*.

Disponibilizar para Escoragem. Quando o modelo é escorado, os itens selecionados neste grupo devem ser produzidos. O valor predito (para todas as respostas) e a confiança (para variáveis resposta categóricas) são sempre calculados quando o modelo é escorado. A confiança calculada pode ser baseada na probabilidade do valor predito (a probabilidade predita mais alta) ou na diferença entre a probabilidade predita mais alta e a segunda probabilidade predita mais alta.

- **Probabilidade predita para variáveis resposta categóricas.** Isso produz as probabilidades preditas para variáveis resposta categóricas. Um campo é criado para cada categoria.
- **Escores de propensão para destino de sinalização.** Para modelos com uma resposta de flag (que retornam uma predição de sim ou não), é possível solicitar os escores de propensão que indicam a

probabilidade do resultado real especificado para o campo de destino. O modelo produz escores de propensão bruta; se as partições estiverem em vigor, o modelo também produzirá escores de propensão ajustada com base na partição de teste.

Médias Estimadas: Esta guia permite exibir as médias marginais estimadas para os níveis de fatores e interações entre fatores. As médias marginais estimadas não estão disponíveis para modelos multinomiais.

Termos. Os termos modelo nos Efeitos Fixos que são totalmente constituídos de campos categóricos são listados aqui. Verifique cada termo para o qual deseja que o modelo produza médias marginais estimadas.

- **Tipo de Contraste.** Especifica o tipo de contraste a ser utilizado para os níveis do campo de contraste. Se **Nenhum** for selecionado, nenhum contraste será produzido. **Entre pares** produz comparações pairwise para todas as combinações de nível dos fatores especificados. Este é o único contraste disponível para interações entre fatores. Os contrastes de **Desvio** comparam cada nível do fator com a média global. Os contrastes **Simples** comparam cada nível do fator, exceto o último, com o último nível. O nível "último" é determinado pela ordenação dos fatores especificados nas Opções de Criação. Observe que todos estes tipos de contraste não são ortogonais.
- **Campo de Contraste.** Especifica um fator, cujos níveis são comparados usando o tipo de contraste selecionado. Se **Nenhum** for selecionado como o tipo de contraste, nenhum campo de contraste poderá (ou precisará) ser selecionado.

Campos Contínuos. Os campos contínuos listados são extraídos dos termos nos Efeitos Fixos que utilizam campos contínuos. Ao calcular médias marginais estimadas, as covariáveis são fixadas nos valores especificados. Selecione a média ou especifique um valor customizado.

Exibir médias estimadas em termos de. Especifica se é necessário calcular médias marginais estimadas com base na escala original da resposta ou com base na transformação de função de ligação. A **Escala de resposta original** calcula médias marginais estimadas para a resposta. Observe que quando a resposta é especificada utilizando a opção de eventos/avaliações, isso fornece as médias marginais estimadas para a proporção de eventos/avaliações e não para o número de eventos. A **Transformação de função de ligação** calcula as médias marginais estimadas para o preditor linear.

Ajuste para diversas comparações usando. Ao executar teste de hipóteses com diversos contrastes, o nível de significância global poderá ser ajustado a partir dos níveis de significância para os contrastes incluídos. Isso permite escolher o método de ajustamento.

- **Diferença menos significativa.** Esse método não controla a probabilidade geral de rejeitar a hipótese de que alguns contrastes lineares são diferentes dos valores de hipótese nulos.
- *Bonferroni Sequencial*. Este é um procedimento de Bonferroni de rejeição sequencialmente decrescente que tende ser muito menos conservador em termos de rejeição de hipóteses individuais, mas mantém o mesmo nível de significância geral.
- *Sidak Sequencial*. Este é um procedimento de Sidak de rejeição sequencialmente decrescente que tende ser muito menos conservador em termos de rejeição de hipóteses individuais, mas mantém o mesmo nível de significância geral.

O método de diferença menos significativa é menos conservador do que o método Sidak sequencial que, por sua vez, é menos conservador do que Bonferroni sequencial, ou seja, a diferença menos significativa rejeitará pelo menos tantas hipóteses individuais quanto o Sidak sequencial que, por sua vez, rejeitará pelo menos tantas hipóteses individuais quanto o Bonferroni sequencial.

Visualização do Modelo: Por padrão, a visualização Sumarização de Modelo é mostrada. Para ver outra visualização do modelo, selecione-a nas miniaturas de visualização.

Sumarização do Modelo: Essa visualização é uma captura instantânea ou uma sumarização rápida do modelo e de seu ajuste.

Tabela. A tabela identifica a resposta, a distribuição de probabilidade e a função de ligação especificadas nas Configurações de resposta. Se a resposta for definida pelos eventos e avaliações, a célula será dividida para mostrar o campo de eventos e o campo de avaliações ou o número fixo de avaliações. Além disso, o Finite Sample Corrected Akaike Information Criterion (AICC) e o Critério de Informação Bayesiano (BIC) são exibidos.

- *Akaike Corrected*. Uma medida para selecionar e comparar modelos combinados com base no log da verossimilhança -2 (Restrito). Valores menores indicam melhores modelos. O AICC "corrige" o AIC para tamanhos de amostra pequenos. Conforme o tamanho da amostra aumenta, o AICC converge para o AIC.
- *bayesiano*. Uma medida para selecionar e comparar modelos com base no log da verossimilhança -2. Valores menores indicam melhores modelos. O BIC também "penaliza" modelos sobreparametrizados (por exemplo, modelos complexos com um número grande de entradas), mas é mais rígido do que o AIC.

Gráfico. Se a resposta for categórica, um gráfico exibirá a precisão do modelo final, que é a porcentagem das classificações corretas.

Estrutura de Dados: Esta visualização fornece uma sumarização da estrutura de dados que você especificou e ajuda a verificar se os sujeitos e medidas repetidas foram especificados corretamente. As informações observadas para o primeiro sujeito são exibidas para cada campo de sujeito e campo de medidas repetidas, bem como para o destino. Além disso, o número de níveis para cada campo de sujeito e campo de medidas repetidas é exibido.

Predito Por Observado: Para variáveis resposta contínuas, incluindo respostas especificadas como eventos/avaliações, isso exibe um gráfico de dispersão categorizado dos valores preditos no eixo vertical em função dos valores observados no eixo horizontal. Idealmente, os pontos devem estar em uma linha de 45 graus, e essa visualização poderá informar se algum registro foi particularmente mal predito pelo modelo.

Classificação: Para variáveis resposta categóricas, isto exibe a classificação cruzada de valores observados versus valores preditos em um heat map, mais o percentual geral correto.

Estilos da tabela. Há vários estilos diferentes de exibição que podem ser acessados a partir da lista suspensa **Estilo**.

- **Percentuais de linhas.** Isso exibe as porcentagens de linhas (as contagens de células expressas como uma porcentagem dos totais de linhas) nas células. Este é o padrão.
- **Contagens de células.** Isso exibe as contagens de célula nas células. O sombreamento para o heat map ainda baseia-se nas porcentagens de linhas.
- **Heat map.** Isso não exibe nenhum valor nas células, apenas o sombreamento.
- **Compactado.** Isso não exibe nenhuma linha ou título da coluna, nem valores nas células. Este pode ser útil quando a resposta tiver um lote de categorias.

Omisso. Se quaisquer registros tiverem valores omissos na resposta, eles serão exibidos em uma linha **(Omisso)** abaixo de todas as linhas válidas. Os registros com valores omissos não contribuem com o percentual geral correto.

Diversas respostas. Se houver diversas variáveis resposta categóricas, então cada resposta será exibida em uma tabela separada e haverá uma lista suspensa **Resposta** que controla qual resposta exibir.

Tabelas Grandes. Se a resposta exibida possuir mais de 100 categorias, nenhuma tabela será exibida.

Efeitos Fixos: Essa visualização exibe o tamanho de cada efeito fixo no modelo.

Estilos. Há diferentes estilos de exibição que podem ser acessados a partir da lista suspensa **Estilo**.

- **Diagrama.** Esse é um gráfico no qual os efeitos são classificados de cima para baixo na ordem em que eles foram especificados nas configurações de Efeitos Fixos. As linhas de conexão no diagrama são ponderadas com base na significância do efeito, com a largura de linha maior correspondendo aos efeitos mais significativos (valores de *p* menores). Este é o padrão.
- **Tabela.** Esta é uma tabela ANOVA para efeitos do modelo geral e de modelo individual. Os efeitos são classificados de cima para baixo na ordem em que eles foram especificados nas configurações de Efeitos Fixos.

Significância. Há uma régua de controle Significância que controla quais efeitos são mostrados na visualização. Os efeitos com valores de significância maiores que o valor da régua de controle são ocultados. Isso não altera o modelo, apenas permite focar nos efeitos mais importantes. Por padrão, o valor é 1,00, de modo que nenhum efeito seja filtrado com base na significância.

Coeficientes Fixos: Essa visualização exibe o valor de cada coeficiente fixo no modelo. Observe que os fatores (preditores categóricos) são codificados por indicador no modelo, de modo que os **efeitos** que contiverem fatores geralmente possuirão diversos **coeficientes** associados, um para cada categoria, exceto para a categoria correspondente ao coeficiente redundante.

Estilos. Há diferentes estilos de exibição que podem ser acessados a partir da lista suspensa **Estilo**.

- **Diagrama.** Este é um gráfico que exibe o intercepto primeiro e, em seguida, classifica os efeitos de cima para baixo na ordem em que eles foram especificados nas configurações de Efeitos Fixos. Nos efeitos que contêm fatores, os coeficientes são classificados em ordem crescente de valores de dados. As linhas de conexão no diagrama são coloridas e ponderadas com base na significância do coeficiente, com a largura de linha maior correspondendo aos coeficientes mais significativos (valores de *p* menores). Este é o estilo padrão.
- **Tabela.** Mostra os valores, os testes de significância e os intervalos de confiança para coeficientes de modelo individuais. Após o intercepto, os efeitos são classificados de cima para baixo na ordem em que eles foram especificados nas configurações de Efeitos Fixos. Nos efeitos que contêm fatores, os coeficientes são classificados em ordem crescente de valores de dados.

Multinomial. Se a distribuição multinomial estiver em vigor, então a lista suspensa Multinomial controlará qual categoria de destino exibir. A ordenação dos valores na lista é determinada pela especificação das configurações de Opções de Criação.

Exponencial. Exibe estimativas de coeficiente exponencial e intervalos de confiança para determinados tipos de modelo, incluindo regressão logística binária (distribuição binomial e ligação logit), regressão logística nominal (distribuição multinomial e ligação logit), regressão binomial negativa (distribuição binomial negativa e ligação de log) e modelo Log-linear (distribuição de Poisson e ligação de log).

Significância. Há uma régua de controle Significância que controla quais coeficientes são mostrados na visualização. Os coeficientes com valores de significância maiores que o valor da régua de controle são ocultados. Isso não altera o modelo, apenas permite focar nos coeficientes mais importantes. Por padrão, o valor é 1,00, de modo que nenhum coeficiente seja filtrado com base na significância.

Covariâncias de Efeito Aleatório: Esta visualização exibe a matriz de covariâncias de efeitos aleatórios (**G**).

Estilos. Há diferentes estilos de exibição que podem ser acessados a partir da lista suspensa **Estilo**.

- **Valores de covariância.** Esse é um heat map da matriz de covariâncias no qual os efeitos são classificados de cima para baixo na ordem em que eles foram especificados nas configurações de Efeitos Fixos. As cores no diagrama de correlação correspondem aos valores da célula conforme mostrado na chave. Este é o padrão.
- **Corrgram.** Este é um heat map da matriz de covariâncias.
- **Compactado.** Este é um heat map da matriz covariâncias sem a linha e títulos de coluna.

Blocos. Se houver diversos blocos de efeito aleatório, haverá uma lista suspensa Bloco para selecionar o bloco para exibir.

Grupos. Se um bloco de efeito aleatório possuir uma especificação de grupo, então haverá uma lista suspensa Grupo para selecionar o nível de grupo para exibir.

Multinomial. Se a distribuição multinomial estiver em vigor, então a lista suspensa Multinomial controlará qual categoria de destino exibir. A ordenação dos valores na lista é determinada pela especificação das configurações de Opções de Criação.

Parâmetros de Covariância: Esta visualização exibe as estimativas de parâmetro de covariância e estatísticas relacionadas para efeitos residuais e aleatórios. Esses resultados avançados, porém fundamentais, fornecem informações sobre se a estrutura de covariâncias é apropriada.

Tabela de sumarização. Essa é uma referência rápida para o número de parâmetros nas matrizes de covariâncias de efeito residual (**R**) e aleatório (**G**), para o ranqueamento (número de colunas) nas matrizes de design de efeito fixo (**X**) e efeito aleatório (**Z**) e para o número de sujeitos definidos pelos campos de sujeito que definem a estrutura de dados.

Tabela de parâmetro de covariância. Para o efeito selecionado, a estimativa, o erro padrão e o intervalo de confiança são exibidos para cada parâmetro de covariância. O número de parâmetros mostrados depende da estrutura de covariâncias para o efeito e, para blocos de efeito aleatórios, o número de efeitos no bloco. Se achar que os parâmetros fora da diagonal não são significativos, será possível utilizar uma estrutura de covariâncias mais simples.

Efeitos. Se houver blocos de efeito aleatório, haverá uma lista suspensa Efeito para selecionar o bloco de efeito residual ou aleatório para exibir. O efeito residual está sempre disponível.

Grupos. Se um bloco de efeito residual ou aleatório possuir uma especificação de grupo, então haverá uma lista suspensa Grupo para selecionar o nível de grupo para exibir.

Multinomial. Se a distribuição multinomial estiver em vigor, então a lista suspensa Multinomial controlará qual categoria de destino exibir. A ordenação dos valores na lista é determinada pela especificação das configurações de Opções de Criação.

Médias Estimadas: Efeitos Significativos: Esses gráficos exibem os 10 efeitos de todos os fatores fixos "mais significativos", começando com interações de três vias, seguido pelas interações de duas vias e, por fim, os efeitos principais. O gráfico exibe o valor estimado pelo modelo da resposta no eixo vertical para cada valor do efeito principal (ou do primeiro efeito listado em uma interação) no eixo horizontal, uma linha separada é produzida para cada valor do segundo efeito listado em uma interação, um gráfico separado é produzido para cada valor do terceiro efeito listado em uma interação de três vias e todos os outros preditores são mantidos constantes. Ele fornece uma visualização útil dos efeitos dos coeficientes de cada preditor na resposta. Observe que, se nenhum preditor for significativo, nenhuma média estimada será produzida.

Confiança. Exibe os limites de confiança superior e inferior para as médias marginais, utilizando o nível de confiança especificado como parte das Opções de Criação.

Médias Estimadas: Efeitos Customizados: Essas são tabelas e gráficos para efeitos de todos os fatores fixos solicitados pelo usuário.

Estilos. Há diferentes estilos de exibição que podem ser acessados a partir da lista suspensa **Estilo**.

- **Diagrama.** Esse estilo exibe um gráfico de linha do valor estimado pelo modelo da resposta no eixo vertical para cada valor do efeito principal (ou do primeiro efeito listado em uma interação) no eixo horizontal, uma linha separada é produzida para cada valor do segundo efeito listado em uma

interação, um gráfico separado é produzido para cada valor do terceiro efeito listado em uma interação de três vias e todos os outros preditores são mantidos constantes.

Se contrastes forem solicitados, outro gráfico será exibido para comparar os níveis do campo de contraste; para interações, um gráfico é exibido para cada combinação de nível dos outros efeitos que o campo de contraste. Para contrates de **pares**, é um gráfico de rede de distância, ou seja, uma representação gráfica da tabela de comparações na qual as distâncias entre os nós da rede correspondem às diferenças entre as amostras. As linhas amarelas correspondem às distâncias estatisticamente significantes e as linhas pretas correspondem às diferenças não significantes. Focalizar uma linha na rede exibe uma dica de ferramenta com a significância ajustada da diferença entre os nós conectados pela linha.

Para contrastes de **Desvio**, um gráfico de barras é exibido com o valor estimado pelo modelo da resposta no eixo vertical e com os valores do campo de contraste no eixo horizontal e, para interações, um gráfico é exibido para cada combinação de nível dos efeitos que não sejam do campo de contraste. As barras mostram a diferença entre cada nível do campo de contraste e a média geral, que é representada por uma linha horizontal preta.

Para contrastes **simples**, um gráfico de barras é exibido com o valor estimado pelo modelo da resposta no eixo vertical e com os valores do campo de contraste no eixo horizontal e, para interações, um gráfico é exibido para cada combinação de nível dos efeitos que não sejam do campo de contraste. As barras mostram a diferença entre cada nível do campo de contraste (exceto o último) e o último nível, que é representada por uma linha horizontal preta.

- **Tabela.** Esse estilo exibe uma tabela do valor estimado pelo modelo da resposta, seu erro padrão e o intervalo de confiança para cada combinação de nível dos campos em vigor, e todos os outros preditores são mantidos constantes.

  Se contrastes forem solicitados, outra tabela é exibida com a estimativa, com o erro padrão, com o teste de significância e com o intervalo de confiança para cada contraste e, para interações, há um conjunto separado de linhas para cada combinação de nível dos efeitos que não sejam do campo de contraste. Além disso, uma tabela com os resultados do teste geral é exibida e, para interações, há um teste geral separado para cada combinação de nível dos efeitos que não sejam do campo de contraste.

Confiança. Alterna a exibição de limites de confiança superior e inferior para médias marginais, utilizando o nível de confiança especificado como parte das Opções de Criação.

Layout. Alterna o layout do diagrama de contrastes entre pares. O layout de círculo é menos revelador de contrastes do que o layout da rede, mas evita sobreposição de linhas.

Configurações: Quando o modelo é escorado, os itens selecionados nesta guia devem ser produzidos. O valor predito (para todas as respostas) e a confiança (para variáveis resposta categóricas) são sempre calculados quando o modelo é escorado. A confiança calculada pode ser baseada na probabilidade do valor predito (a probabilidade predita mais alta) ou na diferença entre a probabilidade predita mais alta e a segunda probabilidade predita mais alta.

- **Probabilidade predita para variáveis resposta categóricas.** Isso produz as probabilidades preditas para variáveis resposta categóricas. Um campo é criado para cada categoria.
- **Escores de propensão para destino de sinalização.** Para modelos com uma resposta de flag (que retornam uma predição de sim ou não), é possível solicitar os escores de propensão que indicam a probabilidade do resultado real especificado para o campo de destino. O modelo produz escores de propensão bruta; se as partições estiverem em vigor, o modelo também produzirá escores de propensão ajustada com base na partição de teste.

Gerar SQL para este modelo Ao usar dados de um banco de dados, código SQL pode ser enviado por push de volta para o banco de dados para execução, fornecendo desempenho superior para muitas operações.

Selecione uma das opções a seguir para especificar como a geração de SQL é executada.

- **Padrão: Escorar usando o Server Scoring Adapter (se instalado) no processo** Se conectado a um banco de dados com um adaptador de escoragem instalado, gera a SQL utilizando o adaptador de escoragem e funções definidas pelo usuário (UDF) associadas e escora seu modelo no banco de dados. Quando nenhum adaptador de escoragem estiver disponível, essa opção busca seus dados novamente a partir do banco de dados e os escora no SPSS Modeler.
- **Escorar ao converter para SQL nativo** Se selecionada, gera SQL nativo para escorar o modelo no banco de dados.

  Nota: Embora essa opção possa fornecer resultados mais rápidos, o tamanho e a complexidade do SQL nativo aumentam conforme a complexidade do modelo aumenta.
- **Escorar fora do Banco de dados** Se selecionada, esta opção busca seus dados novamente a partir de seu banco de dados e os escora no SPSS Modeler.

Nó Cox

A Regressão de Cox constrói um modelo preditivo para os dados de tempo de eventos. O modelo produz uma função de sobrevivência que prediz a probabilidade de o evento de interesse ter ocorrido em um determinado momento (*t*) de acordo com os valores fornecidos das variáveis preditoras. A forma da função de sobrevivência e os coeficientes de regressão para os preditores são estimados a partir de sujeitos observados; o modelo poderá, então, ser aplicado aos novos casos que tiverem medições para as variáveis preditoras. Observe que as informações de sujeitos censurados, ou seja, aqueles que não experimentam o evento de interesse durante o tempo de observação, contribuem utilmente para a estimação do modelo.

Exemplo. Como parte de seus esforços para reduzir a migração para o concorrente, uma empresa de telecomunicações deseja modelar o "tempo para migração para o concorrente" para determinar os fatores que estão associados aos clientes que mais rapidamente querem mudar para outro serviço. Para isso, uma amostra aleatória de clientes é selecionada e o tempo permanecido como clientes (independentemente se forem clientes ativos ou não) e vários campos demográficos são extraídos do banco de dados.

Requisitos. São necessários um ou mais campos de entrada, apenas um campo de destino e um campo de tempo de sobrevivência que deverá ser especificado dentro do nó Cox. O campo de destino deverá ser codificado de modo que o valor "false" indique a sobrevivência e o valor "true" indique que o evento de interesse ocorreu, e ter um nível de medição de *Flag*, com o armazenamento de sequência de caracteres ou de número inteiro. (O armazenamento pode ser convertido utilizando um nó Preenchimento ou Derivar, se necessário). Os campos configurados para *Ambos* ou *Nenhum* são ignorados. Os campos utilizados no modelo devem ter seus tipos totalmente instanciados. O tempo de sobrevivência pode ser qualquer campo numérico.

Datas e Horas. Os campos de Data/Hora não podem ser utilizados para definir diretamente o tempo de sobrevivência; se você tiver campos Data/Hora, eles deverão ser utilizados para criar um campo contendo os tempos de sobrevivência, com base na diferença entre a data de entrada e a data do estudo e da observação.

Análise de Kaplan-Meier. A regressão de Cox pode ser executada sem campos de entrada. Isso é equivalente a uma análise de Kaplan-Meier.

Opções de Campo do Nó Cox

Tempo de sobrevivência. Escolha um campo numérico (um com um nível de medição de *Contínuo*) para tornar o nó executável. O tempo de sobrevivência indica o tempo de vida do registro que está sendo predito. Por exemplo, quando modelar o tempo do cliente para migrar para o concorrente, este é o campo que registra quanto tempo o cliente esteve com a organização. A data em que o cliente entrou ou que migrou para o concorrente não afeta o modelo, apenas a duração do aforamento do cliente é relevante.

O tempo de sobrevivência é considerado como uma duração sem unidades. Assegure-se de que os campos de entrada correspondam ao tempo de sobrevivência. Por exemplo, em um estudo para medir a migração para o concorrente por meses, você utilizaria vendas por mês como uma entrada ao invés de vendas por ano. Se os dados tiverem datas de início e de encerramento ao invés de uma duração, essas datas deverão ser recodificadas para uma duração anterior do nó Cox.

Os campos restantes nessa caixa de diálogo são aqueles padrão utilizados em todo o IBM SPSS Modeler. Consulte o tópico "Opções de Campos do Nó de Modelagem" na página 31 para obter mais informações.

Opções de Modelo do Nó Cox

Nome do modelo É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Utilizar dados particionados. Se um campo de partição for definido, essa opção assegurará que apenas os dados da partição de treinamento sejam utilizados para construir o modelo.

Criar modelos de divisão. Constrói um modelo separado para cada valor possível de campos de entrada que são especificados como campos de divisão. Consulte o tópico "Construindo Modelos de Divisão" na página 28 para obter mais informações.

Método. As opções a seguir estão disponíveis para inserir preditores no modelo:

- **Inserir.** Este é o método padrão que insere todos os termos no modelo diretamente. Nenhuma seleção de campo é executada na construção do modelo.
- **Stepwise.** O método Stepwise de seleção de campo constrói o modelo em passos, tal como o nome implica. O modelo inicial é o modelo mais simples possível, sem termos modelo (exceto a constante) no modelo. Em cada passo, os termos que ainda não tiverem sido incluídos no modelo são avaliados e, se o melhor desses termos aumentar significativamente o poder preditivo do modelo, ele será incluído. Além disso, os termos que estiverem atualmente no modelo são reavaliados para determinar se algum deles pode ser removido sem reduzir significativamente o modelo. Caso positivo, eles serão removidos. O processo se repete e outros termos são incluídos e/ou removidos. Quando mais nenhum termo puder ser incluído para melhorar o modelo, e mais nenhum termo puder ser removido sem reduzir o modelo, o modelo final será gerado.
- **Backwards Stepwise.** O método Backwards Stepwise é essencialmente o oposto do método Stepwise. Com esse método, o modelo inicial contém todos os termos como preditores. Em cada passo, os termos no modelo são avaliados e todos os termos que puderem ser removidos sem reduzir significativamente o modelo são removidos. Além disso, os termos removidos anteriormente são reavaliados para determinar se o melhor desses termos aumenta significativamente o poder preditivo do modelo. Caso positivo, ele é incluído novamente no modelo. Quando mais nenhum termo puder ser removido sem reduzir significativamente o modelo e mais nenhum termo puder ser incluído para melhorar o modelo, o modelo final será gerado.

Nota: os métodos automáticos, incluindo Stepwise e Backwards Stepwise, são métodos de aprendizado altamente adaptáveis e que possuem uma forte tendência a super ajuste dos dados de treinamento. Ao utilizar esses métodos, é essencialmente importante verificar a validade do modelo resultante, seja com novos dados ou com uma amostra de teste de validação criada utilizando o nó Partição.

Grupos. Especificar um campo de grupos faz com que o nó calcule modelos separados para cada categoria do campo. Este pode ser qualquer campo categórico (Flag ou Nominal) com um armazenamento de sequência de caracteres ou de número inteiro.

Tipo de modelo. Há duas opções para definição dos termos no modelo. Os modelos de **Efeitos principais** incluem apenas os campos de entrada individualmente e não testam as interações (efeitos

multiplicadores) entre os campos de entrada. Os modelos **Customizados** incluem apenas os termos (efeitos principais e interações) que você especificar. Ao selecionar essa opção, utilize a lista Termos Modelo para incluir ou remover termos no modelo.

Termos Modelo. Ao construir um modelo Customizado, será necessário especificar explicitamente os termos no modelo. A lista mostra o conjunto atual de termos para o modelo. Os botões do lado direito da lista Termos Modelo permitem incluir e remover termos modelo.

- Para incluir termos no modelo, clique no botão *Incluir novos termos modelo*.
- Para excluir termos, selecione os termos desejados e clique no botão *Excluir termos modelo selecionados*.

Incluindo Termos em um Modelo de Regressão de Cox

Ao solicitar um modelo customizado, é possível incluir termos no modelo clicando no botão *Incluir novos termos modelo* na guia Modelo. Uma nova caixa de diálogo é aberta na qual é possível especificar termos.

Tipo de termo a incluir. Há várias maneiras de incluir termos no modelo, com base na seleção de campos de entrada na lista Campos disponíveis.

- **Interação única.** Insere o termo que representa a interação de todos os campos selecionados.
- **Efeitos principais.** Insere um termo de efeito principal (o campo em si) para cada campo de entrada selecionado.
- **Todas as interações de duas vias.** Insere um termo de interação de duas vias (o produto dos campos de entrada) para cada par possível de campos de entrada selecionados. Por exemplo, se você tiver selecionado os campos de entrada *A*, *B* e *C* na lista Campos disponíveis, esse método inserirá os termos *A * B*, *A * C* e *B * C*.
- **Todas as interações de três vias.** Insere um termo de interação de três vias (o produto dos campos de entrada) para cada combinação possível de campos de entrada selecionados, três por vez. Por exemplo, se você tiver selecionado os campos de entrada *A*, *B*, *C* e *D* na lista Campos disponíveis, esse método inserirá os termos *A * B * C*, *A * B * D*, *A * C * D* e *B * C * D*.
- **Todas as interações de quatro vias.** Insere um termo de interação de quatro vias (o produto dos campos de entrada) para cada combinação possível de campos de entrada selecionados, quatro por vez. Por exemplo, se você tiver selecionado os campos de entrada *A*, *B*, *C*, *D* e *E* na lista Campos disponíveis, esse método inserirá os termos *A * B * C * D*, *A * B * C * E*, *A * B * D * E*, *A * C * D * E* e *B * C * D * E*.

Campos disponíveis. Lista os campos de entrada disponíveis a serem utilizados na construção de termos modelo. Observe que a lista pode incluir campos que não são campos de entrada legais, portanto, assegure-se cuidadosamente de que todos os termos modelo incluam apenas campos de entrada.

Visualizar. Mostra os termos que serão incluídos no modelo se você clicar em **Inserir**, com base nos campos selecionados e no tipo de termo selecionado acima.

Inserir. Insere os termos no modelo (com base na seleção de campos e no tipo de termo atuais) e fecha a caixa de diálogo.

Opções Avançadas do Nó Cox

Convergência. Essas opções permitem controlar os parâmetros para convergência do modelo. Ao executar o modelo, as configurações de convergência controlam quantas vezes os diferentes parâmetros são executados repetidamente até poder ver quão bem eles se ajustam. Quanto mais frequentemente os parâmetros forem tentados, mais próximos os resultados serão (ou seja, os resultados convergirão). Consulte o tópico “Critérios de Convergência de Nó de Cox” na página 219 para obter mais informações.

Saída. Essas opções permitem solicitar estatísticas e gráficos adicionais, incluindo a curva de sobrevivência, que serão exibidos na saída avançada do modelo gerado construído pelo nó. Consulte o tópico “Opções de Saída Avançadas do Nó Cox” na página 219 para obter mais informações.

Progresso. Essas opções permitem controlar os critérios para incluir e remover campos com o método de estimativa Stepwise. (O botão estará desativado se o método Inserir for selecionado). Consulte o tópico "Critérios de Progresso do Nó Cox" para obter mais informações.

Critérios de Convergência de Nó de Cox

Máximo de iterações. Permite especificar o máximo de iterações para o modelo, que controla por quanto tempo o processo procurará uma solução.

Convergência de log da verossimilhança. As iterações pararão se a mudança relativa no log de verossimilhança for menor que esse valor. O critério não será utilizado se o valor for 0.

Convergência de parâmetro. As iterações pararão se a mudança absoluta ou relativa nas estimativas de parâmetro for menor que esse valor. O critério não será utilizado se o valor for 0.

Opções de Saída Avançadas do Nó Cox

Estatísticas. É possível obter estatísticas para seus parâmetros de modelo, incluindo intervalos de confiança para exp(*B*) e correlação de estimativas. É possível solicitar essas estatísticas em cada passo ou apenas no último passo.

Exibir função de linha de base. Permite exibir a função de risco da linha de base e a sobrevivência acumulativa na média das covariáveis.

Gráficos

Os gráficos podem ajudar a avaliar seu modelo estimado e a interpretar os resultados. É possível representar as funções de sobrevivência, de risco, de log menos log e de um menos sobrevivência.

- *Sobrevivência*. Exibe a função de sobrevivência acumulativa em uma escala linear.
- *Risco*. Exibe a função de risco acumulativo em uma escala linear.
- **Log menos log.** Exibe a estimativa de sobrevivência acumulativa após a transformação ln(-ln) ser aplicada à estimativa.
- *Um menos sobrevivência*. Exibe a função um menos a sobrevivência em uma escala linear.

Representa uma linha separada para cada valor. Esta opção está disponível apenas para campos categóricos.

Valor a ser usado para gráficos. Como essas funções dependem de valores dos preditores, deve-se utilizar valores constantes para os preditores representarem as funções versus o tempo. O padrão é utilizar a média de cada preditor como um valor constante, mas é possível inserir seus próprios valores no gráfico utilizando a grade. Para entradas categóricas, como a codificação do indicador é utilizada, há um coeficiente de regressão para cada categoria (exceto a última). Assim, uma entrada categórica possui um valor médio para cada contraste do indicador, igual à proporção dos casos na categoria correspondente ao contraste do indicador.

Critérios de Progresso do Nó Cox

Critério de remoção. Selecione **Razão de Verossimilhança** para um modelo mais robusto. Para reduzir o tempo necessário para construir o modelo, é possível tentar selecionar **Wald**. Também há a opção **Condicional**, que fornece teste de remoção com base na probabilidade da estatística de razão de verossimilhança com base nas estimativas de parâmetro condicional.

Limites de significância para critérios. Esta opção permite especificar critérios de seleção com base na probabilidade estatística (o valor *p*) associada a cada campo. Os campos serão incluídos no modelo somente se o valor de *p* associado for menor que o valor de **Entrada** e será removido apenas se o valor de *p* for maior que o valor de **Remoção**. O valor de **Entrada** deve ser menor que o valor de **Remoção**.

Opções de Configurações do Nó Cox

Prever sobrevivência em tempos futuros. Especifique um ou mais tempos no futuro. A sobrevivência, ou seja, se cada caso deverá sobreviver durante pelo menos esse período de tempo (a partir de agora) sem o evento terminal ocorrer, é predita para cada registro em cada valor de tempo, uma predição por valor de tempo. Observe que a sobrevivência é o valor "false" do campo de destino.

- **Intervalos regulares.** Os valores de tempo de sobrevivência são gerados a partir do **Intervalo de tempo** e do **Número de períodos de tempo para escorar** especificados. Por exemplo, se 3 períodos de tempo forem solicitados com um intervalo de 2 entre cada tempo, a sobrevivência será predita para os tempos futuros 2, 4 e 6. Cada registro é avaliado nos mesmos valores de tempo.
- **Campos de tempo.** Os tempos de sobrevivência são fornecidos para cada registro no campo de tempo escolhido (um campo de predição é gerado), portanto, cada registro poderá ser avaliado em tempos diferentes.

Tempo de sobrevivência passado. Especifique o tempo de sobrevivência do registro até o momento, por exemplo, o aforamento de um cliente existente como um campo. Escorar a probabilidade de sobrevivência em um tempo futuro será condicional sobre o tempo de sobrevivência passado.

Nota: os valores dos tempos de sobrevivência passados e futuros devem estar dentro do intervalo de tempos de sobrevivência nos dados utilizados para treinar o modelo. Os registros cujos tempos estiverem fora desse intervalo são escorados como nulos.

Incluir todas as probabilidades. Especifica se as probabilidades de cada categoria do campo de saída são incluídas em cada registro processado pelo nó. Se essa opção não estiver selecionada, apenas a probabilidade da categoria predita será incluída. As probabilidades são calculadas para cada tempo futuro.

Calcular função de risco acumulativo. Especifica se o valor do risco acumulativo é incluído em cada registro. O risco acumulativo é calculado para cada tempo futuro.

Nugget do Modelo Cox

Os modelos de regressão de Cox representam as equações estimadas pelos nós Cox. Elas contêm todas as informações capturadas pelo modelo, bem como informações sobre a estrutura e o desempenho do modelo.

Ao executar um fluxo que contém um modelo de regressão de Cox gerado, o nó inclui dois novos campos contendo a predição do modelo e a probabilidade associada. Os nomes dos novos campos são derivados do nome do campo de saída que está sendo predito, prefixados com *$C-* para a categoria predita e com *$CP-* para a probabilidade associada, e sufixados com o número do intervalo de tempo futuro ou com o nome do campo de tempo que define o intervalo de tempo. Por exemplo, para um campo de saída denominado *churn* e dois intervalos de tempo futuros definidos em intervalos regulares, os novos campos serão denominados *$C-churn-1*, *$CP-churn-1*, *$C-churn-2* e *$CP-churn-2*. Se tempos futuros forem definidos com um campo de tempo *tenure*, os novos campos serão *$C-churn_tenure* e *$CP-churn_tenure*.

Se você selecionou a opção de configurações **Incluir todas as probabilidades** no nó Cox, dois campos adicionais serão incluídos para cada tempo futuro, contendo as probabilidades de sobrevivência e de falha para cada registro. Esses campos adicionais são nomeados com base no nome do campo de saída, prefixados com *$CP-<false value>–* para a probabilidade de sobrevivência e com *$CP-<true value>–* para a probabilidade de o evento ter ocorrido, e sufixados com o número do intervalo de tempo futuro. Por exemplo, para um campo de saída em que o valor "false" é 0 e o valor "true" é 1, e dois intervalos de tempo futuro definidos em intervalos regulares, os novos campos serão denominados *$CP-0-1*, *$CP-1-1*, *$CP-0-2* e *$CP-1-2*. Se tempos futuros forem definidos com um único campo de tempo *tenure*, os novos campos serão *$CP-0-1* e *$CP-1-1*, já que há um único intervalo futuro.

Se você selecionou a opção de configurações **Calcular função de risco acumulativo** no Nó Cox, um campo adicional será incluído para cada tempo futuro, contendo a função de risco acumulativo para cada registro. Esses campos adicionais são nomeados com base no nome do campo de saída, prefixados com *$CH-* e sufixados com o número do intervalo de tempo futuro ou com o nome do campo de tempo que define o intervalo de tempo. Por exemplo, para um campo de saída denominado *churn* e dois intervalos de tempo futuro definidos em intervalos regulares, os novos campos serão denominados *$CH-churn-1* e *$CH-churn-2*. Se tempos futuros forem definidos com um campo de tempo *tenure*, o novo campo será *$CH-churn-1*.

Configurações da Saída de Regressão de Cox

Exceto para geração de SQL, a guia Configurações do nugget contém os mesmos controles que a guia Configurações do nó do modelo. Os valores padrão dos controles do nugget são determinados pelos valores configurados no nó de modelo. Consulte o tópico "Opções de Configurações do Nó Cox" na página 220 para obter mais informações.

Gerar SQL para este modelo Ao usar dados de um banco de dados, código SQL pode ser enviado por push de volta para o banco de dados para execução, fornecendo desempenho superior para muitas operações.

Selecione uma das opções a seguir para especificar como a geração de SQL é executada.

- **Padrão: Escorar usando o Server Scoring Adapter (se instalado) no processo** Se conectado a um banco de dados com um adaptador de escoragem instalado, gera a SQL utilizando o adaptador de escoragem e funções definidas pelo usuário (UDF) associadas e escora seu modelo no banco de dados. Quando nenhum adaptador de escoragem estiver disponível, essa opção busca seus dados novamente a partir do banco de dados e os escora no SPSS Modeler.
- **Escorar fora do Banco de dados** Se selecionada, esta opção busca seus dados novamente a partir do banco de dados e os escora no SPSS Modeler.

Saída Avançada de Regressão de Cox

A saída avançada para regressão de Cox fornece informações detalhadas sobre o modelo de estimativa e seu desempenho, incluindo a curva de sobrevivência. A maioria das informações contidas na saída avançada é muito técnica e um conhecimento amplo da regressão de Cox é necessário para interpretar corretamente esta saída.

Capítulo 11. Modelos de Armazenamento em Cluster

Os modelos de armazenamento em cluster focam na identificação de grupos de registros semelhantes e na rotulagem dos registros de acordo com o grupo ao qual eles pertencem. Isso é feito sem o benefício de conhecer previamente os grupos e suas características. Na realidade, talvez você nem saiba exatamente quantos grupos procurar. Isso é o que diferencia os modelos de armazenamento em cluster de outras técnicas de aprendizado por máquina por não haver nenhum campo de saída ou de destino predefinido para o modelo prever. Esses modelos geralmente são referidos como modelos de **aprendizado não supervisionado**, já que não há um padrão externo pelo qual avaliar o desempenho de classificação do modelo. Não há respostas *certas* ou *erradas* para esses modelos. Seu valor é determinado pela sua capacidade de capturar agrupamentos interessantes nos dados e fornecer descrições úteis desses agrupamentos.

Os métodos de armazenamento em cluster baseiam-se na medição das distâncias entre os registros e os clusters. Os registros são designados aos clusters de modo a minimizar a distância entre os registros pertencentes ao mesmo cluster.

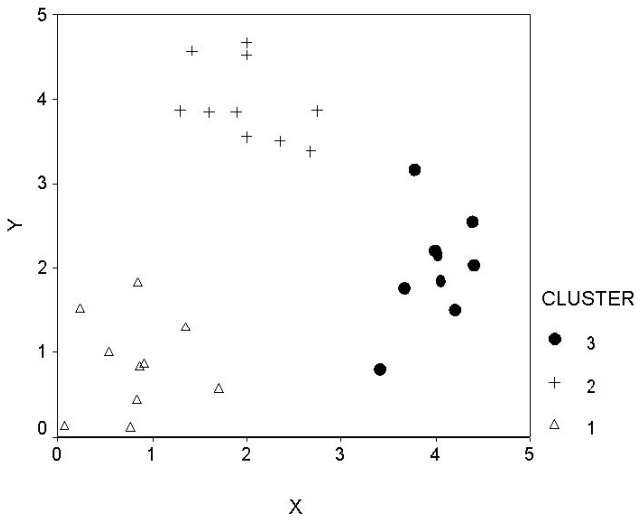


Figura 44. Modelo de armazenamento em cluster simples

Três métodos de armazenamento em cluster são fornecidos:

O nó K-Médias armazena em cluster o conjunto de dados em grupos distintos (ou clusters). O método define um número fixo de clusters, designa iterativamente registros para clusters e ajusta os centros do cluster até que o refinamento adicional não consiga mais melhorar o modelo. Ao invés de tentar prever um resultado, o *k-médias* utiliza um processo conhecido como aprendizado não supervisionado para descobrir padrões no conjunto de campos de entrada.

O nó TwoStep utiliza um método de clusterização em dois passos. O primeiro passo faz uma única passagem através dos dados para compactar os dados de entrada brutos em um conjunto gerenciável de subclusters. O segundo passo usa um método de armazenamento em cluster hierárquico para mesclar progressivamente os subclusters em clusters cada vez maiores. O TwoStep tem a vantagem de estimar automaticamente o número ideal de clusters para os dados de treinamento. Ele pode manipular tipos de campo combinados e grandes conjuntos de dados de maneira eficiente.

O nó Kohonen gera um tipo de rede neural que pode ser utilizado para armazenar em cluster o conjunto de dados em grupos distintos. Quando a rede for totalmente treinada, os registros similares deverão estar próximos no mapa de saída, ao passo que registros que forem diferentes estarão distantes. É possível examinar o número de observações capturadas por cada unidade no nugget do modelo para identificar as unidades fortes. Isto poderá dar uma ideia do número apropriado de clusters.

Os modelos de armazenamento em cluster geralmente são usados para criar clusters ou segmentos que, em seguida, são usados como entradas em análises subsequentes. Um exemplo comum disso são os segmentos de mercado usados pelos comerciantes para particionar seu mercado geral em subgrupos homogêneos. Cada segmento tem características especiais que afetam o sucesso dos esforços de marketing direcionados a ele. Se você estiver usando mineração de dados para otimizar sua estratégia de marketing, geralmente é possível melhorar seu modelo significativamente identificando os segmentos apropriados e usando as informações do segmento em seus modelos preditivos.

Nó Kohonen

As redes Kohonen são um tipo de rede neural que executa armazenamento em cluster, também conhecido como um **knet** ou um **mapa de auto-organização**. Este tipo de rede pode ser utilizado para armazenar em cluster o conjunto de dados em grupos distintos quando você não souber quais desses grupos estão no início. Os registros são agrupados de modo que os registros em um grupo ou cluster sejam similares entre si, e que registros em grupos diferentes sejam dissimilares.

As unidades básicas são **neurônios**, que são organizados em duas camadas: a **camada de entrada** e a **camada de saída** (também chamada de **mapa de saída**). Todos os neurônios de entrada são conectados a todos os neurônios de saída, e essas conexões possuem **intensidades** ou **ponderações** associadas a eles. Durante o treinamento, cada unidade concorre com todas as outras para "vencer" cada registro.

O mapa de saída é uma grade bidimensional de neurônios, sem conexões entre as unidades.

Os dados de entrada são apresentados à camada de entrada, e os valores são propagados para a camada de saída. O neurônio de saída com a resposta mais forte é considerado o **vencedor** e é a resposta para essa entrada.

Inicialmente, todas as ponderações são aleatórias. Quando uma unidade vence um registro, suas ponderações (junto com as ponderações de outras unidades próximas, referidas coletivamente como uma **vizinhança**) são ajustadas para melhor corresponder ao padrão de valores de preditor para esse registro. Todos os registros de entrada são mostrados, e as ponderações são atualizadas apropriadamente. Esse processo é repetido muitas vezes até que as mudanças se tornem muito pequenas. Conforme o treinamento avança, as ponderações nas unidades de grade são ajustadas para que elas formem um "mapa" bidimensional dos clusters (daí o termo **mapa de auto-organização**).

Quando a rede for totalmente treinada, os registros que forem semelhantes deverão estar próximos no mapa de saída, ao passo que registros que forem diferentes estarão mais distantes.

Ao contrário da maioria dos métodos de aprendizado no IBM SPSS Modeler, as redes Kohonen *não* utilizam um campo de destino. Esse tipo de aprendizado sem um campo de destino é chamado de **aprendizado não supervisionado**. Ao invés de tentar prever um resultado, a rede Kohonen tenta descobrir padrões no conjunto de campos de entrada. Geralmente, uma rede Kohonen terminará com algumas unidades que sumarizam muitas observações (unidades **fortes**) e várias unidades que realmente não correspondam a nenhuma das observações (unidades **fracas**). As unidades fortes (e às vezes outras unidades adjacentes a elas na grade) representam centros de cluster prováveis.

Outro uso das redes Kohonen está na **redução de dimensão**. A característica espacial da grade bidimensional fornece um mapeamento dos *k* preditores originais para duas variáveis derivadas que preservam os relacionamentos de similaridade nos preditores originais. Em alguns casos, isso pode fornecer o mesmo tipo de benefícios que a análise de fator ou PCA.

Observe que o método para calcular o tamanho padrão da grade de saída foi alterado a partir de versões anteriores do IBM SPSS Modeler. O novo método geralmente produzirá camadas de saída menores que são mais rápidas de treinar e melhores de generalizar. Se achar que está obtendo resultados ruins com o tamanho padrão, tente aumentar o tamanho da grade de saída na guia Especialista. Consulte o tópico "Opções Avançadas do Nó Kohonen" na página 226 para obter mais informações.

Requisitos. Para treinar uma rede Kohonen, um ou mais campos com o papel configurado para *Entrada* são necessários. Os campos com o papel configurado para *Destino*, *Ambos* ou *Nenhum* são ignorados.

Intensidades. Não é necessário fazer com que os dados na associação ao grupo construam um modelo de rede Kohonen. Não é necessário nem mesmo saber o número de grupos para procurar. As redes Kohonen iniciam com um número grande de unidades e, conforme o treinamento avança, as unidades fluem em direção aos clusters naturais nos dados. É possível observar o número de observações capturadas por cada unidade no nugget do modelo para identificar as unidades fortes, que podem dar uma ideia melhor do número apropriado de clusters.

Opções de Modelo do Nó Kohonen

Nome do modelo É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Utilizar dados particionados. Se um campo de partição for definido, essa opção assegurará que apenas os dados da partição de treinamento sejam utilizados para construir o modelo.

Continuar treinando o modelo existente. Por padrão, toda vez que você executar um nó Kohonen, uma rede completamente nova é criada. Se selecionar essa opção, o treinamento continuará com a última rede produzida com sucesso pelo nó.

Mostrar gráfico de feedback. Se essa opção for selecionada, uma representação visual da matriz bidimensional será exibida durante o treinamento. A intensidade de cada nó é representada por cor. Vermelho indica uma unidade está vencendo muitos registros (uma unidade **forte**) e branco indica uma unidade que está vencendo poucos ou nenhum registro (uma unidade **fraca**). O feedback poderá não ser exibido se o tempo gasto para construir o modelo for relativamente curto. Observe que esse recurso pode reduzir o tempo de treinamento. Para acelerar o tempo de treinamento, desmarque essa opção.

Parar em. O critério de parada padrão interrompe o treinamento com base nos parâmetros internos. Também é possível especificar o tempo como o critério de parada. Insira o tempo (em minutos) para a rede treinar.

Configurar semente aleatória. Se nenhuma semente aleatória for configurada, a sequência de valores aleatórios utilizados para inicializar as ponderações de rede será diferente toda vez que o nó for executado. Isso poderá fazer com que o nó crie modelos diferentes em execuções diferentes, mesmo se as configurações do nó e os valores de dados forem exatamente os mesmos. Ao selecionar essa opção, será possível configurar a semente aleatória para um valor específico, para que o modelo resultante seja reproduzido exatamente igual. Uma semente aleatória específica sempre gera a mesma sequência de valores aleatórios, em que, nesse caso, a execução do nó sempre produz o mesmo modelo gerado.

Nota: ao usar a opção **Configurar semente aleatória** com registros lidos a partir de um banco de dados, um nó Ordenar poderá ser necessário antes da amostragem para assegurar o mesmo resultado sempre

que o nó for executado. Isso ocorre porque a semente aleatória depende da ordem de registros, que não é garantido que ela permaneça a mesma em um banco de dados relacional.

Nota: se desejar incluir campos nominais (conjunto) em seu modelo, mas estiver tendo problemas de memória na construção do modelo ou o modelo está demorando muito tempo para construir, considere recodificar campos de conjunto grandes para reduzir o número de valores, ou considere usar um campo diferente com menos valores como um proxy para o conjunto grande. Por exemplo, se você estiver tendo um problema com um campo *product_id* que contém valores para produtos individuais, será possível considerar removê-lo do modelo e incluir um campo *product_category* menos detalhado.

Otimizar. Selecione as opções projetadas para aumentar o desempenho durante a construção de modelo com base em suas necessidades específicas.

- Selecione **Velocidade** para instruir o algoritmo a nunca utilizar spilling de disco para melhorar o desempenho.
- Selecione **Memória** para instruir o algoritmo a utilizar spilling de disco quando apropriado em algum sacrificar de velocidade. Essa opção é selecionada por padrão.

  Nota: ao executar no modo distribuído, essa configuração poderá ser substituída pelas opções do administrador especificadas em *options.cfg*.

Anexar rótulo do cluster. Selecionada por padrão para novos modelos, mas desmarcada para modelos carregados a partir de versões anteriores do IBM SPSS Modeler, esta opção cria um campo de escoragem categórica único do mesmo tipo que é criado por ambos os nós K-Médias e TwoStep. Este campo de sequência de caracteres é utilizado no nó Cluster Automático quando calcular medidas de classificação para os diferentes tipos de modelo. Consulte o tópico "Nó Cluster Automático" na página 74 para obter mais informações.

Opções Avançadas do Nó Kohonen

Para aqueles que possuem um conhecimento detalhado de redes Kohonen, as opções avançadas permitem fazer um ajuste preciso do processo de treinamento. Para acessar as opções avançadas, configure o Modo para **Especialista** na guia Especialista.

Largura e Comprimento. Especifique o tamanho (largura e comprimento) do mapa de saída bidimensional como o número de unidades de saída com cada dimensão.

Declínio da taxa de aprendizado. Selecione o declínio da taxa de aprendizado linear ou exponencial. A **taxa de aprendizado** é um fator de ponderação que diminui ao longo do tempo, em que a rede começa codificando variáveis em larga escala dos dados, passando a focar gradativamente em detalhes de nível mais refinado.

Fase 1 e a Fase 2. O treinamento da rede Kohonen é dividido em duas fazes. A Fase 1 é uma fase de estimação aproximada, utilizada para capturar padrões brutos nos dados. A Fase 2 é uma fase de ajuste, utilizada para ajustar o mapa para modelar as variáveis mais finas dos dados. Para cada fase, há três parâmetros:

- **Vizinhança.** Configura o tamanho inicial (raio) da vizinhança. Isto determina o número de unidades "vizinhas" que são atualizadas com a unidade vencedora durante o treinamento. Durante a fase 1, o tamanho da vizinhança começa em *Phase 1 Neighborhood* e diminui para (*Phase 2 Neighborhood* + 1). Durante a fase 2, o tamanho da vizinhança começa em *Phase 2 Neighborhood* e diminui para 1,0. *Phase 1 Neighborhood* deve ser maior que *Phase 2 Neighborhood*.
- **Eta. Inicial** Configura o valor inicial para a taxa de aprendizado **eta**. Durante a fase 1, o eta começa em *Phase 1 Initial Eta* e diminui para *Phase 2 Initial Eta*. Durante a fase 2, o eta inicia em *Phase 2 Initial Eta* e diminui para 0. O *Phase 1 Initial Eta* deve ser maior que *Phase 2 Initial Eta*.
- **Ciclos.** Configura o número de ciclos para cada fase do treinamento. Cada fase continua para o número especificado de passagens dos dados.

Nuggets do Modelo de Kohonen

Os nuggets do modelo de Kohonen contêm todas as informações capturadas pela rede Kohenen treinada, bem como informações sobre a arquitetura da rede.

Ao executar um fluxo contendo um nugget do modelo Kohonen, o nó inclui dois novos campos contendo as coordenadas *X* e *Y* da unidade na grade de saída de Kohonen que responderam mais fortemente a esse registro. Os novos nomes de campos são derivados do nome do modelo, prefixados por *$KX-* e *$KY-*. Por exemplo, se seu modelo for denominado *Kohonen*, os novos campos serão denominados *$KX-Kohonen* e *$KY-Kohonen*.

Para ter uma ideia melhor dos itens que a rede Kohonen codificou, clique na guia Modelo no navegador do nugget do modelo. Isso exibe o Visualizador de Cluster, que fornece uma representação gráfica de clusters, campos e níveis de importância. Consulte o tópico "Visualizador de Cluster – Guia Modelo" na página 238 para obter mais informações.

Se preferir visualizar os clusters como uma grade, será possível visualizar o resultado da rede Kohonen ao representar os campos *$KX-* e *$KY-* utilizando um nó Gráfico. (Deve-se selecionar **Agitação X** e **Agitação Y** no nó Gráfico para evitar que todos os registros de cada unidade sejam representados um em cima do outro). No gráfico, também é possível sobrepor um campo simbólico para investigar como a rede Kohonen armazenou os dados em cluster.

Outra técnica poderosa para obter insight da rede Kohonen é utilizar a indução de regra para descobrir as características que distinguem os clusters localizados pela rede. Consulte o tópico "Nó C5.0" na página 105 para obter mais informações.

Para obter informações gerais sobre como utilizar o navegador do modelo, consulte "Procurando Nuggets do Modelo" na página 42

Sumarização do Modelo de Kohonen

A guia Sumarização de um nugget do modelo de Kohonen exibe informações sobre a arquitetura ou topologia da rede. O comprimento e largura do mapa de variáveis Kohonen bidimensional (a camada de saída) são mostrados como **$KX-** *model_name* e **$KY-** *model_name*. Para as camadas de entrada e de saída, o número de unidades nessa camada é listado.

Nó K-Médias

O nó K-Médias fornece um método de **análise de cluster**. Ele pode ser utilizado para armazenar em cluster o conjunto de dados em grupos distintos quando você não souber quais desses grupos estão no início. Ao contrário da maioria dos métodos de aprendizado no IBM SPSS Modeler, os modelos K-Médias *não* utilizam um campo de destino. Esse tipo de aprendizado sem um campo de destino é chamado de **aprendizado não supervisionado**. Ao invés de tentar prever um resultado, o K-Médias tenta descobrir padrões no conjunto de campos de entrada. Os registros são agrupados de modo que os registros em um grupo ou cluster sejam similares entre si, mas que registros em grupos diferentes sejam dissimilares.

O K-Médias funciona ao definir um conjunto de centros de cluster de início derivados dos dados. Em seguida, ele designa cada registro para o cluster ao qual ele for mais similar, com base nos valores do campo de entrada do registro. Após todos os casos terem sido designados, os centros do cluster são atualizados para refletir o novo conjunto de registros designados para cada cluster. Os registros são, então, verificados novamente para ver se eles devem ser redesignados para um cluster diferente, e o processo de designação de registro/iteração do cluster continua até que o número máximo de iterações seja atingido ou até que a mudança entre uma iteração e a próxima falhe ao exceder um limite especificado.

Nota: o modelo resultante depende, até certo ponto, da ordem dos dados de treinamento. Reordenar os dados e reconstruir o modelo pode levar a um modelo de cluster final diferente.

Requisitos. Para treinar um modelo K-Médias, um ou mais campos com o papel configurado para *Entrada* são necessários. Os campos com o papel configurado para *Saída*, *Ambos* ou *Nenhum* são ignorados.

Intensidades. Não é necessário fazer com que os dados na associação ao grupo construam um modelo K-Médias. O modelo K-Médias geralmente é o método mais rápido de armazenar em cluster grandes conjuntos de dados.

Opções de Modelo do Nó K-Médias

Nome do modelo É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Utilizar dados particionados. Se um campo de partição for definido, essa opção assegurará que apenas os dados da partição de treinamento sejam utilizados para construir o modelo.

Número de clusters especificados. Especifique o número de clusters a serem gerados. O padrão é 5.

Gerar campo distância. Se essa opção for selecionada, o nugget do modelo incluirá um campo contendo a distância de cada registro a partir do centro do seu cluster designado.

Rótulo do cluster. Especifique o formato para os valores no campo de associação de cluster gerado. A associação de cluster pode ser indicada como uma **Sequência de caracteres** com o **Prefixo de rótulo** especificado (por exemplo, `"Cluster 1"`, `"Cluster 2"`, e assim por diante) ou como um **Número**.

Nota: se desejar incluir campos nominais (conjunto) em seu modelo, mas estiver tendo problemas de memória na construção do modelo ou o modelo está demorando muito tempo para construir, considere recodificar campos de conjunto grandes para reduzir o número de valores, ou considere usar um campo diferente com menos valores como um proxy para o conjunto grande. Por exemplo, se você estiver tendo um problema com um campo *product_id* que contém valores para produtos individuais, será possível considerar removê-lo do modelo e incluir um campo *product_category* menos detalhado.

Otimizar. Selecione as opções projetadas para aumentar o desempenho durante a construção de modelo com base em suas necessidades específicas.

- Selecione **Velocidade** para instruir o algoritmo a nunca utilizar spilling de disco para melhorar o desempenho.
- Selecione **Memória** para instruir o algoritmo a utilizar spilling de disco quando apropriado em algum sacrificar de velocidade. Essa opção é selecionada por padrão.

  Nota: ao executar no modo distribuído, essa configuração poderá ser substituída pelas opções do administrador especificadas em *options.cfg*.

Opções Avançadas do Nó K-Médias

Para aqueles que possuem um conhecimento detalhado de armazenamento em cluster de *k*-médias, as opções avançadas permitem fazer um ajuste preciso do processo de treinamento. Para acessar as opções avançadas, configure o Modo para **Especialista** na guia Especialista.

Parar em. Especifique o critério de parada a ser utilizado no treinamento do modelo. O critério de parada **Padrão** é 20 iterações ou uma mudança < 0,000001, ou o que ocorrer primeiro. Selecione **Customizado** para especificar seus próprios critérios de parada.

- **Máximo de Iterações.** Esta opção permite parar o treinamento do modelo após o número de iterações especificado.

- **Alterar tolerância.** Esta opção permite parar o treinamento do modelo quando a maior mudança nos centros de cluster para uma iteração for menor que o nível especificado.

Codificando valor para conjuntos. Especifique um valor entre 0 e 1,0 a ser usado para recodificar os campos de conjunto como grupos de campos numéricos. O valor padrão é a raiz quadrada de 0,5 (aproximadamente 0,707107), que fornece a ponderação adequada para campos de flag recodificados. Os valores mais próximos de 1,0 atribuem uma ponderação maior para os campos de conjunto do que para os campos numéricos.

Nuggets do Modelo de K-Médias

Os nuggets do modelo K-Médias contêm todas as informações capturadas pelo modelo de armazenamento em cluster, bem como informações sobre os dados de treinamento e o processo de estimação.

Ao executar um fluxo contendo um nó de modelagem K-Médias, o nó inclui dois novos campos contendo a associação de cluster e distância do centro do cluster designado para esse registro. Os novos nomes de campo são derivados do nome do modelo, prefixados com *$KM-* para a associação de cluster e com *$KMD-* para a distância do centro do cluster. Por exemplo, se o seu modelo for denominado *Kmeans*, os novos campos serão denominados *$KM-Kmeans* e *$KMD-Kmeans*.

Uma técnica poderosa para obter insight do modelo do K-Médias é utilizar a indução de regra para descobrir as características que distinguem os clusters localizados pelo modelo. Consulte o tópico “Nó C5.0” na página 105 para obter mais informações. Também é possível clicar na guia Modelo do navegador do nugget do modelo para exibir o Visualizador de Cluster, que fornece uma representação gráfica de clusters, campos e níveis de importância. Consulte o tópico “Visualizador de Cluster – Guia Modelo” na página 238 para obter mais informações.

Para obter informações gerais sobre como utilizar o navegador do modelo, consulte “Procurando Nuggets do Modelo” na página 42

Sumarização do Modelo de K-Médias

A guia Sumarização de um nugget do modelo K-Médias contém informações sobre os dados de treinamento, sobre o processo de estimação e sobre os clusters definidos pelo modelo. O número de clusters é mostrado, bem como o histórico de iteração. Se você tiver executado um nó Análise anexado a este nó de modelagem, as informações dessa análise também serão exibidas nesta seção.

Nó do Cluster TwoStep

O nó do Cluster TwoStep fornece uma forma de **análise de cluster**. Ele pode ser utilizado para armazenar em cluster o conjunto de dados em grupos distintos quando você não souber quais desses grupos estão no início. Assim como os nós Kohonen e os nós K-Médias, os modelos de cluster TwoStep *não* utilizam um campo de destino. Ao invés de tentar prever um resultado, o Cluster TwoStep tenta descobrir padrões no conjunto de campos de entrada. Os registros são agrupados de modo que os registros em um grupo ou cluster sejam similares entre si, mas que registros em grupos diferentes sejam dissimilares.

O Cluster TwoStep é um método de clusterização em dois passos O primeiro passo faz uma simples passagem pelos dados, quando ele compacta os dados de entrada brutos em um conjunto gerenciável de subclusters. O segundo passo usa um método de armazenamento em cluster hierárquico para mesclar progressivamente os subclusters em clusters cada vez maiores, sem requerer passar novamente pelos dados. O armazenamento em cluster hierárquico tem a vantagem de não requerer selecionar o número de clusters antes do tempo. Muitos métodos de armazenamento em cluster hierárquicos iniciam com registros individuais como clusters individuais e mesclam esses registros recursivamente para produzir

clusters ainda maiores. Embora essas abordagens normalmente manipulam enormes quantias de dados, o pré-armazenamento em cluster inicial do TwoStep torna o armazenamento em cluster hierárquico rápido até mesmo para conjuntos de dados grandes.

Nota: O modelo resultante depende, até certo ponto, da ordem dos dados de treinamento. Reordenar os dados e reconstruir o modelo pode levar a um modelo de cluster final diferente.

Requisitos. Para treinar um modelo do Cluster TwoStep, um ou mais campos com o papel configurado para *Entrada* são necessários. Os campos com o papel configurado para *Destino*, *Ambos* ou *Nenhum* são ignorados. O algoritmo de Cluster TwoStep não manipula valores omissos. Os registros com espaços em branco para qualquer um dos campos de entrada serão ignorados durante a construção do modelo.

Intensidades. O Cluster TwoStep pode manipular tipos de campo combinados e manipular grandes conjuntos de dados de maneira eficiente. Ele também tem a capacidade de testar várias soluções de cluster e escolher a melhor, de modo que você não precisa saber quantos clusters solicitar no início. O Cluster TwoStep pode ser configurado para excluir automaticamente **valores discrepantes**, ou casos extremamente incomuns que possam afetar seus resultados.

Importante:

O IBM SPSS Modeler possui duas versões diferentes do nó Cluster TwoStep:

- O **Cluster TwoStep** é o nó tradicional que é executado no IBM SPSS Modeler Server.
- O **Cluster TwoStep-AS** é executado apenas quando conectado ao IBM SPSS Analytic Server.

Opções de Modelo do Nó Cluster TwoStep

Nome do modelo É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Utilizar dados particionados. Se um campo de partição for definido, essa opção assegurará que apenas os dados da partição de treinamento sejam utilizados para construir o modelo.

Padronizar campos numéricos. Por padrão, o TwoStep padronizará todos os campos de entrada numéricos para a mesma escala, com uma média de 0 e uma variância de 1. Para manter a escala original para campos numéricos, desmarque essa opção. Os campos simbólicos não são afetados.

Excluir valores discrepantes. Se essa opção for selecionada, os registros que não parecerem se ajustar a um cluster substantivo serão excluídos automaticamente da análise. Isso evita que casos distorçam os resultados.

A detecção de valor discrepante ocorre durante o passo de pré-armazenamento em cluster. Quando essa opção é selecionada, os subclusters com poucos registros relativos a outros subclusters são considerados possíveis valores discrepantes e a árvore de subclusters é reconstruída excluindo esses registros. O tamanho abaixo do qual os subclusters são considerados conterem possíveis valores discrepantes é controlado pela opção **Porcentagem**. Alguns desses registros de possíveis valores discrepantes poderão ser incluídos nos subclusters reconstruídos se eles forem semelhantes o suficiente a qualquer um dos novos perfis de subcluster. O restante dos possíveis valores discrepantes que não puderem ser mesclados são considerados valores discrepantes e serão incluídos em um cluster "ruído" e excluídos do passo de armazenamento em cluster hierárquico.

Ao *escorar* dados com um modelo TwoStep que utiliza tratamento de valores discrepantes, os novos casos que tiverem uma distância maior que a especificada para o limite (com base no log da verossimilhança) a partir do cluster substantivo mais próximo serão considerados valores discrepantes e designados ao cluster "ruído" com o nome -1.

Rótulo do cluster. Especifica o formato para o campo de associação de cluster gerado. A associação de cluster pode ser indicada como uma **Sequência de caracteres** com o **Prefixo de rótulo** especificado (por exemplo, `"Cluster 1"`, `"Cluster 2"`, e assim por diante) ou como um **Número**.

Calcular automaticamente o número de clusters. O cluster TwoStep pode analisar muito rapidamente um grande número de soluções de cluster para escolher o número ideal de clusters para os dados de treinamento. Especifique uma variedade de soluções para tentar ao configurar o número **Máximo** e **Mínimo** de clusters. O TwoStep utiliza um processo de dois estágios para determinar o número ideal de clusters. No primeiro estágio, um limite superior no número de clusters no modelo é selecionado com base na mudança no Bayes Information Criterion (BIC) conforme mais clusters são incluídos. No segundo estágio, a mudança na distância mínima entre clusters é localizada para todos os modelos com menos clusters do que a solução BIC mínima. A maior mudança na distância é utilizada para identificar o modelo de cluster final.

Especificar o número de clusters. Se você souber quantos clusters serão incluídos em seu modelo, selecione esta opção e insira o número de clusters.

Medida de distância. Esta seleção determina como a similaridade entre dois clusters é calculada.

- **Log da verossimilhança.** A medida de probabilidade coloca uma distribuição de probabilidade nas variáveis. Variáveis contínuas são consideradas como sendo distribuídas normalmente, ao passo que as variáveis categóricas são consideradas como sendo multinomiais. Todas as variáveis são consideradas independentes.
- **Euclidiana.** A medida Euclidiana é a distância em "linha reta" entre dois clusters. Ela pode ser utilizada somente quando todas as variáveis forem contínuas.

Critério de armazenamento em cluster. Esta seleção determina como o algoritmo de armazenamento em cluster automático determina o número de clusters. O Critério de Informação Bayesiano (BIC) ou o Akaike Information Criterion (AIC) podem ser especificados.

Nuggets do Modelo de Cluster TwoStep

Os nuggets do modelo de cluster TwoStep contêm todas as informações capturadas pelo modelo de armazenamento em cluster, bem como informações sobre os dados de treinamento e o processo de estimação.

Ao executar um fluxo contendo um nugget do modelo de cluster TwoStep, o nó incluirá um novo campo contendo a associação de cluster para esse registro. O novo nome de campo é derivado do nome do modelo, prefixado por *$T-*. Por exemplo, se seu modelo for denominado *TwoStep*, o novo campo será denominado *$T-TwoStep*.

Uma técnica poderosa para obter insight do modelo do TwoStep é utilizar a indução de regra para descobrir as características que distinguem os clusters localizados pelo modelo. Consulte o tópico “Nó C5.0” na página 105 para obter mais informações. Também é possível clicar na guia Modelo do navegador do nugget do modelo para exibir o Visualizador de Cluster, que fornece uma representação gráfica de clusters, campos e níveis de importância. Consulte o tópico “Visualizador de Cluster – Guia Modelo” na página 238 para obter mais informações.

Para obter informações gerais sobre como utilizar o navegador do modelo, consulte “Procurando Nuggets do Modelo” na página 42

Sumarização do Modelo TwoStep

A guia Sumarização para uma nugget do modelo de cluster TwoStep exibe o número de clusters localizados, com informações sobre os dados de treinamento, o processo de estimação e as configurações de construção utilizadas.

Consulte o tópico “Procurando Nuggets do Modelo” na página 42 para obter mais informações.

Nó do Cluster TwoStep-AS

O IBM SPSS Modeler possui duas versões diferentes do nó Cluster TwoStep:

- O **Cluster TwoStep** é o nó tradicional que é executado no IBM SPSS Modeler Server.
- O **Cluster TwoStep-AS** é executado apenas quando conectado ao IBM SPSS Analytic Server.

Análise de Cluster Twostep-AS

O Cluster TwoStep é uma ferramenta exploratória projetada para revelar agrupamentos naturais (ou clusters) dentro de um conjunto de dados que, de outra forma, não seriam aparentes. O algoritmo que é utilizado por este procedimento possui vários recursos desejáveis que o diferenciam das técnicas tradicionais de armazenamento em cluster:

- **Manipulação de variáveis categóricas e contínuas.** Ao considerar variáveis a serem independentes, uma distribuição multinomial normal conjunta pode ser colocada em variáveis categóricas e contínuas.
- **Seleção automática do número de clusters.** Ao comparar os valores de um critério de escolha de modelo em diferentes soluções de armazenamento em cluster, o procedimento poderá determinar automaticamente o número ideal de clusters.
- **Escalabilidade.** Ao construir uma árvore de variáveis de cluster (CF) que sumarizam os registros, o algoritmo TwoStep poderá analisar grandes arquivos de dados.

Por exemplo, empresas de produto de varejo e de consumidor aplicam regularmente técnicas de armazenamento em cluster às informações que descrevem atributos como hábitos de compra, sexo, idade, nível de renda, etc., dos clientes. Essas empresas customizam suas estratégias de desenvolvimento de marketing e de produtos para aumentar as vendas e construir fidelidade à marca.

Nota: Este nó requer o IBM SPSS Analytic Server.

Guia Campos

A guia Campos especifica quais campos são utilizados na análise.

Utilizar papéis predefinidos. Todos os campos com um papel definido de Entrada são selecionados.

Usar designações de campo customizadas. Incluir e remover campos, independentemente de suas designações de papel definidas. É possível selecionar campos com qualquer papel e movê-los para a, ou a partir da lista **Preditores (Entradas)**.

Básicos

Número de clusters

Determinar automaticamente
O procedimento determina o melhor número de clusters, dentro do intervalo especificado. O **Mínimo** deve ser maior que 1. Esta é a opção padrão.

Especificar fixo
O procedimento gera o número especificado de clusters. O **Número** deve ser maior que 1.

Critério de armazenamento em cluster

Essa opção controla a forma com que o algoritmo de armazenamento em cluster automático determina o número de clusters.

Critério de informações Bayesiano (BIC)
Uma medida para selecionar e comparar modelos com base no log da verossimilhança -2. Valores

menores indicam melhores modelos. O BIC também "penaliza" modelos sobreparametrizados (por exemplo, modelos complexos com um número grande de entradas), mas é mais rígido do que o AIC.

Akaike Information Criterion (AIC)
Uma medida para selecionar e comparar modelos com base no log da verossimilhança -2. Valores menores indicam melhores modelos. O AIC "penaliza" modelos sobreparametrizados (modelos complexos com um número grande de entradas, por exemplo).

Método de armazenamento em cluster automático

Se você selecionar **Determinar automaticamente**, escolha entre os métodos de armazenamento em cluster a seguir utilizados para determinar automaticamente o número de clusters:

Utilizar configuração de Critério de Armazenamento em Cluster
A convergência de critérios de informações é a razão de critérios de informações correspondentes entre duas soluções de cluster atuais e a primeira solução de cluster. O critério utilizado é aquele selecionado no grupo Critério de Armazenamento em Cluster.

Salto de distância
O salto de distância é a razão das distâncias correspondentes com duas soluções de cluster consecutivas.

Máximo
Combina os resultados do método de convergência de critérios de informações e do método de salto de distância para produzir o número de clusters correspondente ao segundo salto.

Mínimo
Combina os resultados do método de convergência de critérios de informações e do método de salto de distância para produzir o número de clusters correspondente ao primeiro salto.

Método de Importância de Variável

O **Método de Importância de Variável** determina a importância que as variáveis (campos) têm na solução de cluster. A saída inclui informações sobre a importância geral da variável e a importância de cada campo de variável em cada cluster. As variáveis que não atenderem a um limite mínimo são excluídas.

Utilizar configuração de Critério de Armazenamento em Cluster.
Este é o método padrão, com base no critério que é selecionado no grupo de Critério de Armazenamento em Cluster.

Tamanho do efeito
A importância de variável baseia-se no tamanho do efeito e não nos valores de significância.

Critérios de Árvore de Variáveis

Essas configurações determinam como a árvore de variável do cluster é construída. Ao construir uma árvore de variáveis de cluster e sumarizar os registros, o algoritmo TwoStep poderá analisar grandes arquivos de dados. Em outras palavras, o cluster TwoStep utiliza uma árvore de variável de cluster para construir clusters, permitindo que ele processe muitos casos.

Medida de distância

Esta seleção determina como a similaridade entre dois clusters é calculada.

Verossimilhança de log

A medida de probabilidade coloca uma distribuição de probabilidade nos campos. Campos contínuos são considerados como sendo distribuídos normalmente, ao passo que os campos categóricos são considerados como sendo multinomiais. Supõe-se que todos os campos sejam independentes.

Euclidiano

A medida Euclidiana é a distância em "linha reta" entre dois clusters. A medida euclidiana quadrada e o método Ward são utilizados para calcular a similaridade entre os clusters. Eles podem ser utilizados apenas quando todos os campos forem contínuos.

Clusters de valor discrepante

Incluir clusters de valor discrepante
Inclui clusters para os casos que forem valores discrepantes dos clusters regulares. Se essa opção não estiver selecionada, todos os casos serão incluídos em clusters regulares.

Número de casos em que a folha da árvore de variável é menor que.
Se o número de casos na folha da árvore de variável for menor que o valor especificado, a folha será considerada um valor discrepante. O valor deve ser um número inteiro maior que 1. Se você alterar esse valor, valores maiores poderão resultar em mais clusters de valores discrepantes.

Porcentagem superior de valores discrepantes.
Quando o modelo de cluster é construído, os valores discrepantes são classificados por intensidade de valores. A intensidade do valor discrepante que é necessária para alcançar a porcentagem superior de valores discrepantes é utilizada como o limite para determinar se um caso é classificado como um valor discrepante. Valores maiores significam que mais casos são classificados como valores discrepantes. O valor deve estar entre 1 a 100.

Configurações adicionais

Limite de mudança de distância inicial

O limite inicial que é utilizado para crescer a árvore de variável de cluster. Se a inserção de uma folha em uma folha da árvore gerar uma tensão menor que este limite, a folha não será dividida. Se a tensão exceder este limite, a folha será dividida.

Máximo de ramificações do nó folha

O número máximo de nós filhos que um nó folha pode ter.

Máximo de ramificações do nó não folha

O número máximo de nós filhos que um nó não folha pode ter.

Profundidade máxima da árvore

O número máximo de níveis que a árvore do cluster pode ter.

Ponderação de ajustamento no nível de medição
Reduz a influência de campos categóricos ao aumentar a ponderação para campos contínuos. Esse valor representa um denominador para reduzir a ponderação para campos categóricos. Portanto, um padrão de 6, por exemplo, fornece aos campos categóricos uma ponderação de 1/6.

Alocação de memória

A quantia máxima de memória em megabytes (MB) que o algoritmo do cluster utiliza. Se o procedimento exceder esse máximo, ele utilizará o disco para armazenar as informações que não couberem na memória.

Divisão atrasada
Atraso de reconstrução da árvore de variável de cluster. O algoritmo de armazenamento em cluster reconstrói a árvore de variável de cluster várias vezes à medida que ele avalia novos casos. Essa opção pode melhorar o desempenho ao atrasar essa operação e reduzir o número de vezes em que a árvore é reconstruída.

Padronizar

O algoritmo de clusterização trabalha com campos contínuos padronizados. Por padrão, todos os campos contínuos são padronizados. Para economizar tempo e esforço computacional, é possível mover campos contínuos que já estiverem padronizados para a lista **Não padronizar**.

Seleção de Variáveis

Na tela Seleção de Variável, é possível configurar regras que determinam quando os campos são excluídos. Por exemplo, é possível excluir campos que tiverem muitos valores omissos.

Regras para Excluir Campos

Porcentagem de valores omissos é maior que.
Os campos com uma porcentagem de valores omissos maior que o valor especificado são excluídos da análise. O valor deve ser um número positivo maior que zero e menor que 100.

Número de categorias para campos categóricos é maior que.
Os campos categóricos que tiverem mais que o número especificado de categorias são excluídos da análise. O valor deve ser um número inteiro positivo maior que 1.

Campos com uma Tendência A Um Valor Único

Coeficiente de variação para campos contínuos é menor que.
Os campos contínuos com um coeficiente de variação menor que o valor especificado são excluídos da análise. O coeficiente da variação é a razão entre o desvio padrão com a média. Valores mais baixos tendem a indicar menor variação nos valores. O valor deve estar entre 0 e 1.

Porcentagem de casos em uma única categoria para campos categóricos é maior que.
Os campos categóricos com uma porcentagem de casos em uma categoria única maior que o valor especificado são excluídos da análise. O valor deve ser maior que 0 e menor que 100.

Seleção de variável adaptativa

Esta opção executa uma passagem de dados extra para localizar e remover os campos menos importantes.

Saída do Modelo

Sumarização de Construção de Modelo

Especificações de modelo
Sumarização das especificações de modelo, número de clusters no modelo final e entradas (campos) incluídos no modelo final.

Sumarização de registro
Número e a porcentagem de registros (casos) incluídos e excluídos do modelo.

Entradas excluídas
Para todos os campos não incluídos no modelo final, o motivo do campo foi excluído.

Avaliação

Qualidade do Modelo
Tabela de qualidade e de importância para cada cluster e a qualidade do ajuste geral do modelo.

Gráfico de barras de importância de variável
Gráfico de barras de importância de variável (campo) em todos os clusters. As variáveis (campos) com barras mais longas no gráfico são mais importantes do que os campos com barras menores. Elas também são classificadas em ordem decrescente de importância (a barra na parte superior é a mais importante).

Nuvem de palavras de importância de variável
Nuvem de palavras de importância de variável (campo) em todos os clusters. As variáveis (campos) com texto maior são mais importantes do que aquelas com texto menor.

Clusters de valor discrepante
Essas opções serão desativadas se você optou por não incluir valores discrepantes.

Tabela e gráfico interativos
Tabela e o gráfico de uma intensidade de valor discrepante e a similaridade relativa de clusters de valores discrepantes com clusters regulares. Selecionar diferentes linhas na tabela exibe informações para diferentes clusters de valores diferentes no gráfico.

Tabela dinâmica
Tabela de uma intensidade de valor discrepante e a similaridade relativa de clusters de valores discrepantes com clusters regulares. Esta tabela contém as mesmas informações que a exibição interativa. Esta tabela suporta todas as variáveis padrão para girar e definir tabelas como dinâmicas.

Número máximo
O número máximo de valores discrepantes para exibir na saída. Se houver mais de vinte clusters de valores discrepantes, uma tabela dinâmica será exibida.

Interpretação

Entre perfis de importância de variável de cluster

Tabela e gráfico interativos
Tabela e gráficos de importância de variável e de centros de cluster para cada entrada (campo) utilizada na solução de cluster. Selecionar diferentes linhas na tabela exibe um gráfico diferente. Para campos categóricos, um gráfico de barras é exibido. Para campos contínuos, um gráfico de médias e de desvios padrão é exibido.

Tabela Dinâmica.
Tabela de importância de variável e de centros de cluster para cada entrada (campo). Esta tabela contém as mesmas informações que a exibição interativa. Esta tabela suporta todas as variáveis padrão para girar e definir tabelas como dinâmicas.

Na importância de variável de cluster
Para cada cluster, o centro do cluster e a importância de variável para cada entrada (campo). Há uma tabela separada para cada cluster.

Distâncias de cluster
Um gráfico de painel que exibe as distâncias entre os clusters. Há um painel separado para cada cluster.

Rótulo de cluster

Texto O rótulo para cada cluster é o valor que é especificado para **Prefixo**, seguido por um número sequencial.

Número
O rótulo para cada cluster é um número sequencial.

Opções do Modelo

Nome do modelo. É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Nuggets do Modelo de Cluster TwoStep-AS

O nugget do modelo TwoStep-AS exibe detalhes do modelo na guia Modelo do Visualizador de Saída. Para obter mais informações sobre como utilizar o visualizador, consulte a seção com o título "Trabalhando com a Saída" no Guia do Usuário do Modelador (ModelerUsersGuide.pdf).

Os nuggets do modelo de cluster TwoStep-AS contêm todas as informações capturadas pelo modelo de armazenamento em cluster, bem como informações sobre os dados de treinamento e o processo de estimação.

Ao executar um fluxo contendo um nugget do modelo de cluster TwoStep-AS, o nó incluirá um novo campo contendo a associação de cluster para esse registro. O novo nome de campo é derivado do nome do modelo, prefixado por ***$AS-***. Por exemplo, se o seu modelo for denominado TwoStep, o novo campo será denominado ***$AS-TwoStep***.

Uma técnica poderosa para obter insight do modelo do TwoStep-AS é utilizar a indução de regra para descobrir as características que distinguem os clusters localizados pelo modelo. Consulte o tópico “Nó C5.0” na página 105 para obter informações adicionais.

Para obter informações gerais sobre como utilizar o navegador do modelo, consulte “Procurando Nuggets do Modelo” na página 42

Configurações do Nugget do Modelo de Cluster TwoStep-AS

A guia Configurações fornece opções adicionais para o nugget do modelo TwoStep-AS.

Gerar SQL para este modelo Ao usar dados de um banco de dados, código SQL pode ser enviado por push de volta para o banco de dados para execução, fornecendo desempenho superior para muitas operações.

Selecione uma das opções a seguir para especificar como a geração de SQL é executada.

- **Padrão: Escorar usando o Server Scoring Adapter (se instalado) no processo** Se conectado a um banco de dados com um adaptador de escoragem instalado, gera a SQL utilizando o adaptador de escoragem e funções definidas pelo usuário (UDF) associadas e escora seu modelo no banco de dados. Quando nenhum adaptador de escoragem estiver disponível, essa opção busca seus dados novamente a partir do banco de dados e os escora no SPSS Modeler.
- **Escorar ao converter para SQL nativo** Se selecionada, gera SQL nativo para escorar o modelo no banco de dados.

  Nota: Embora essa opção possa fornecer resultados mais rápidos, o tamanho e a complexidade do SQL nativo aumentam conforme a complexidade do modelo aumenta.
- **Escorar fora do banco de dados** Se selecionada, esta opção busca seus dados novamente a partir de seu banco de dados e os escora no SPSS Modeler.

O Visualizador de Cluster

Os modelos de cluster são geralmente utilizados para localizar grupos (ou clusters) de registros semelhantes com base nas variáveis analisadas, em que a semelhança entre os membros do mesmo grupo é alta e a semelhança entre os membros de grupos diferentes é baixa. Os resultados podem ser utilizados para identificar as associações que de outra forma não seriam aparentes. Por exemplo, a análise de cluster das preferências, do nível de renda e dos hábitos de compra do cliente permite identificar os tipos de clientes que mais poderão responder a uma campanha de marketing específica.

Há duas abordagens para interpretar os resultados em uma exibição de cluster:

- Examinar os clusters para determinar as características exclusivas para esse cluster. *Um determinado cluster contém todos os solicitantes de crédito de alta renda? Esse cluster contém mais registros do que os outros?*
- Examinar os campos entre os clusters para determinar como os valores são distribuídos entre os clusters. *Um nível específico de educação determina a associação em um cluster? Um escore de risco de crédito alto diferencia a associação entre um cluster e outro?*

Utilizando as visualizações principais e as diversas visualizações vinculadas no Visualizador de Cluster, é possível obter insight para ajudá-lo a responder essas questões.

Os nuggets do modelo de cluster a seguir podem ser gerados no IBM SPSS Modeler:

- Nugget do modelo de rede Kohonen
- Nugget do modelo K-Médias
- Nugget do modelo de cluster TwoStep

Para ver informações sobre os nuggets do modelo de cluster, clique com o botão direito no nó de modelo e escolha **Procurar** no menu de contexto (ou **Editar** para nós em um fluxo). Como alternativa, se você estiver utilizando o nó de modelagem do Cluster Automático, clique duas vezes no nugget do cluster necessário dentro do nugget do modelo Cluster Automático. Consulte o tópico "Nó Cluster Automático" na página 74 para obter mais informações.

Visualizador de Cluster – Guia Modelo

A guia Modelo para modelos de cluster mostra uma exibição gráfica das estatísticas de sumarização e das distribuições dos campos entre os clusters, o que é conhecido como o **Visualizador de Cluster**.

Nota: a guia Modelo não está disponível para modelos construídos em versões do IBM SPSS Modeler anteriores a 13.

O Visualizador de Cluster é formado por dois painéis, a visualização principal à esquerda e a visualização vinculada, ou auxiliar, à direita. Há duas visualizações principais:

- Sumarização do Modelo (o padrão). Consulte o tópico "Visualização de Sumarização do Modelo" para obter mais informações.
- Clusters. Consulte o tópico "Visualização de Clusters" na página 239 para obter mais informações.

Há quatro visualizações vinculadas/auxiliares:

- Importância do Preditor. Consulte o tópico "Visualização de Importância do Preditor de Cluster" na página 240 para obter mais informações.
- Tamanhos do Cluster (o padrão). Consulte o tópico "Visualização de Tamanhos de Cluster" na página 240 para obter mais informações.
- Distribuição de Célula. Consulte o tópico "Visualização da Distribuição de Célula" na página 241 para obter mais informações.
- Comparação de Cluster. Consulte o tópico "Visualização Comparação de Cluster" na página 241 para obter mais informações.

Visualização de Sumarização do Modelo

A visualização Sumarização do Modelo mostra uma captura instantânea ou uma sumarização do modelo de cluster, incluindo uma medida de Silhueta de coesão e separação de cluster que é tonalizada para indicar resultados ruins, regulares ou bons. Essa captura instantânea permite verificar rapidamente se a qualidade é "pobre", o que, neste caso, é possível decidir retornar para o nó de modelagem e corrigir as configurações do modelo de cluster para produzir um resultado melhor.

Os resultados de pobre, regular e bom baseiam-se no trabalho de Kaufman e Rousseeuw (1990) com relação à interpretação das estruturas do cluster. Na visualização Sumarização de Modelo, um resultado "bom" equivale aos dados que refletem a classificação de Kaufman e Rousseeuw como uma evidência

razoável ou forte da estrutura de cluster, "regular" reflete sua classificação de evidência fraca e "pobre" reflete a classificação de nenhuma evidência significativa.

A silhueta calcula a média em todos os registros como (B−A) / max(A,B), em que A é a distância do registro ao centro de seu cluster e B é a distância do registro ao centro do cluster mais próximo ao qual ele não pertence. Um coeficiente de silhueta de 1 significa que todos os casos estão localizados diretamente em seus centros do cluster. Um valor de -1 significa que todos os casos estão localizados nos centros de algum outro cluster. Um valor de 0 significa que, em média, os casos estão equidistantes entre o centro de seu próprio cluster e o cluster seguinte mais próximo.

A sumarização inclui uma tabela que contém as seguintes informações:

- **Algoritmo.** O algoritmo de armazenamento em cluster utilizado, por exemplo, "TwoStep".
- **Variáveis de Entrada.** O número de campos, também conhecido como **Entradas** ou **Preditores**.
- **Clusters.** O número de clusters na solução.

Visualização de Clusters

A visualização Clusters contém uma grade de cluster por variáveis que inclui nomes, tamanhos e perfis de cada cluster.

As colunas na grade contêm as informações a seguir:

- **Cluster.** Os números de cluster criados pelo algoritmo.
- **Rótulo.** Quaisquer rótulos aplicados a cada cluster (em branco por padrão). Clique duas vezes na célula para inserir um rótulo que descreva o conteúdo do cluster, por exemplo, "Compradores de carro de luxo".
- **Descrição.** Qualquer descrição do conteúdo do cluster (em branco por padrão). Clique duas vezes na célula para inserir uma descrição do cluster, por exemplo, "Profissionais com 55 anos de idade ou mais que ganham cima de $100.000".
- **Tamanho.** O tamanho de cada cluster como uma porcentagem da amostra de cluster geral. Cada tamanho de célula dentro da grade exibe uma barra vertical que mostra a porcentagem do tamanho dentro do cluster, uma porcentagem do tamanho em formato numérico e as contagens de caso de cluster.
- **Variáveis.** As entradas ou preditores individuais, ordenadas por importância geral por padrão. Se alguma coluna possuir tamanhos iguais, elas serão mostradas em ordem crescente dos números de cluster.

  A importância de variável geral é indicada pela cor do sombreamento do plano de fundo da célula; a variável mais importante é a mais escura e a variável menos importante não é sombreada. Um guia acima da tabela indica a importância atribuída a cada cor da célula de variável.

Ao passar o mouse sobre uma célula, o nome completo/rótulo da variável e o valor de importância para a célula são exibidos. Informações adicionais podem ser exibidas, dependendo da visualização e do tipo de variável. Na visualização Centros de Cluster, isto inclui a estatística da célula e o valor da célula, por exemplo: “Média: 4,32”. Para variáveis categóricas, a célula mostra o nome da categoria (modal) mais frequente e sua porcentagem.

Na visualização Clusters, é possível selecionar várias maneiras para exibir as informações do cluster:

- Transpor clusters e variáveis Consulte o tópico “Transpor Clusters e Variáveis” na página 240 para obter mais informações.
- Ordenar variáveis Consulte o tópico “Ordenar Variáveis” na página 240 para obter mais informações.
- Ordenar clusters. Consulte o tópico “Ordenar Clusters” na página 240 para obter mais informações.
- Ordenar conteúdo da célula. Consulte o tópico “Conteúdos das células” na página 240 para obter mais informações.

Transpor Clusters e Variáveis: Por padrão, os clusters são exibidos como colunas e as variáveis são exibidas como linhas. Para reverter essa exibição, clique no botão **Transpor Clusters e Variáveis** à esquerda dos botões **Classificar Variáveis Por**. Você pode querer fazer isso, por exemplo, para reduzir a quantia de rolagem horizontal necessária para ver os dados quando tiver muitos clusters exibidos.

Ordenar Variáveis: Os botões **Ordenar Variáveis Por** permitem selecionar como as células de variáveis são exibidas:

- **Importância Geral.** Esse é o padrão de ordenação. As variáveis são classificadas em ordem decrescente de importância geral e a ordenação é a mesma entre os clusters. Se quaisquer variáveis tiverem valores de importância empatados, as variáveis empatadas serão listadas em ordem crescente dos nomes de variável.
- **Importância Dentro do Cluster.** As variáveis são ordenadas com relação à sua importância para cada cluster. Se quaisquer variáveis tiverem valores de importância empatados, as variáveis empatadas serão listadas em ordem crescente dos nomes de variável. Quando essa opção é escolhida, a ordenação geralmente varia entre os clusters.
- **Nome.** As variáveis são classificadas por nome em ordem alfabética.
- **Ordem de dados.** As variáveis são ordenadas pela ordem do conjunto de dados.

Ordenar Clusters: Por padrão, os clusters são classificados em ordem decrescente de tamanho. Os botões **Ordenar Clusters Por** permitem classificá-los por nome em ordem alfabética ou, se você tiver criado rótulos exclusivos, em ordem alfanumérica de rótulo.

As variáveis que possuírem o mesmo rótulo são ordenadas por nome do cluster. Se os clusters forem ordenados por rótulo e você editar o rótulo de um cluster, a ordenação será atualizada automaticamente.

Conteúdos das células: Os botões **Células** permitem alterar a exibição do conteúdo da célula para campos de variáveis e de avaliação.

- **Centros do Cluster.** Por padrão, as células exibem os nomes/rótulos de variáveis e a tendência central para cada combinação de cluster/variável. A média é mostrada para campos contínuos e o modo (categoria que ocorre com mais frequência) com a porcentagem de categorias dos campos categóricos.
- **Distribuições Absolutas.** Mostra nomes/rótulos de variáveis e distribuições absolutas das variáveis em cada cluster. Para variáveis categóricas, a exibição mostra gráficos de barras sobrepostos com categorias classificadas em ordem crescente dos valores de dados. Para variáveis contínuas, a exibição mostra um gráfico de densidade leve que utiliza os mesmos terminais e os intervalos de cada cluster.

  A exibição em cor vermelha sólida mostra a distribuição do cluster, ao passo que a exibição opaca representa os dados gerais.
- **Distribuições Relativas.** Mostra nomes/rótulos de variáveis e distribuições relativas nas células. Em geral, as exibições são similares àquelas mostradas para distribuições absolutas, com a exceção de que as distribuições relativas são exibidas.

  A exibição em cor vermelha sólida mostra a distribuição do cluster, ao passo que a exibição opaca representa os dados gerais.
- **Visualização Básica.** Onde houver muitos clusters, poderá ser difícil ver todos os detalhes sem rolagem. Para reduzir a quantidade de rolagem, selecione esta visualização para alterar a exibição para uma versão mais compacta da tabela.

Visualização de Importância do Preditor de Cluster

A visualização Importância do Preditor mostra a importância relativa de cada campo na estimativa do modelo.

Visualização de Tamanhos de Cluster

A visualização Tamanhos de Cluster mostra um gráfico de pizza contendo cada cluster. O tamanho em porcentagem de cada cluster é mostrado em cada fatia, passe o mouse sobre cada fatia para exibir a contagem nessa fatia.

Abaixo do gráfico, uma tabela lista as informações de tamanho a seguir:

- O tamanho do menor cluster (uma contagem e uma porcentagem do todo).
- O tamanho do maior cluster (uma contagem e uma porcentagem do todo).
- A razão do tamanho do maior cluster com o menor cluster.

Visualização da Distribuição de Célula

A visualização de Distribuição de Célula mostra um gráfico expandido e mais detalhado da distribuição dos dados para qualquer célula de variável que você seleciona na tabela no painel principal de Clusters.

Visualização Comparação de Cluster

A visualização Comparação de Cluster consiste em um layout em estilo de grade, com variáveis nas linhas e clusters selecionados nas colunas. Essa visualização ajuda a entender melhor os fatores que compõem os clusters e também permite ver diferenças entre os clusters não apenas comparado com dados gerais, mas também entre si.

Para selecionar clusters para exibição, clique na parte superior da coluna de cluster no painel principal Clusters. Use Ctrl-clique ou Shift-clique para selecionar ou cancelar a seleção de mais de um cluster para comparação.

Nota: é possível selecionar até cinco clusters para exibição.

Os clusters são mostrados na ordem em que eles foram selecionados, ao passo que a ordem dos campos é determinada pela opção **Classificar Variáveis Por**. Ao selecionar **Importância Dentro do Cluster**, os campos são sempre ordenados por importância geral.

Os gráficos em segundo plano mostram as distribuições gerais de cada variável:

- As variáveis categóricas são mostradas como gráficos de ponto, em que o tamanho do ponto indica a categoria mais frequente/modal para cada cluster (por variável).
- As variáveis contínuas são exibidas como boxplots, que mostram as medianas e as amplitudes interquartis gerais.

Sobrepostos nestas visualizações em segundo plano estão os boxplots para clusters selecionados:

- Para variáveis contínuas, os marcadores de ponto quadrados e as linhas horizontais indicam a mediana e a amplitude interquartil para cada cluster.
- Cada cluster é representado por uma cor diferente, mostrada na parte superior da visualização.

Navegando no Visualizador de Cluster

O Visualizador de Cluster é uma exibição interativa. Nela, é possível:

- Selecionar um campo ou um cluster para visualizar mais detalhes.
- Comparar os clusters para selecionar itens de interesse.
- Alterar a exibição.
- Transpor eixos.
- Gerar os nós Derivar, Filtro e Seleção utilizando o menu Gerar.

Utilizando as Barras de Ferramentas

Controle as informações mostradas nos painéis esquerdo e direito usando as opções da barra de ferramentas. É possível alterar a orientação da exibição (de cima para baixo, da esquerda para a direita ou da direita para a esquerda) utilizando os controles da barra de ferramentas. Além disso, também é possível redefinir o visualizador para as configurações padrão e abrir uma caixa de diálogo para especificar o conteúdo da visualização Clusters no painel principal.

As opções **Ordenar Variáveis Por**, **Ordenar Clusters Por**, **Células** e **Exibição** estão disponíveis apenas ao selecionar a visualização **Clusters** no painel principal. Consulte o tópico "Visualização de Clusters" na página 239 para obter mais informações.

Tabela 12. Ícones da barra de ferramentas.

Ícone	Tópico
	Consulte Transpor Clusters e Variáveis
	Consulte Ordenar Variáveis Por
	Consulte Ordenar Clusters Por
	Consulte Células

Gerando Nós a partir de Modelos de Cluster

O menu Gerar permite criar novos nós com base no modelo de cluster. Esta opção está disponível na guia Modelo do modelo gerado e permite gerar nós com base na exibição ou na seleção atual (ou seja, todos os clusters visíveis ou todos selecionados). Por exemplo, é possível selecionar uma única variável e, em seguida, gerar um nó Filtro para descartar todas as outras variáveis (não-visíveis). Os nós gerados são colocados desconectados na tela. Além disso, é possível gerar uma cópia do nugget do modelo na paleta de modelos. Lembre-se de conectar os nós e de fazer as edições desejadas antes da execução.

- **Gerar Nó de Modelagem.** Cria um nó de modelagem na tela de fluxo. Isso será útil, por exemplo, se você tiver um fluxo no qual deseja utilizar estas configurações de modelo, mas não tiver mais o nó de modelagem utilizado para gerá-los.
- **Modelo para Paleta.** Cria um nugget na panela Modelos. Isso é útil em situações em que um colega pode ter enviado um fluxo contendo o modelo e não o próprio modelo.
- **Nó Filtro.** Cria um novo nó Filtro para filtrar campos que não forem utilizados pelo modelo de cluster e/ou que não estiverem visíveis na exibição do Visualizador de Cluster atual. Se houver um nó Tipo antes deste nó Cluster, quaisquer campos com o papel *Destino* serão descartados pelo nó Filtro gerado.
- **Nó Filtro (da seleção).** Cria um novo nó Filtro para filtrar campos com base nas seleções no Visualizador de Cluster. Selecione diversos campos utilizando o método Ctrl-clique. Os campos selecionados no Visualizador de Cluster são descartados posteriormente, mas este comportamento pode ser alterado ao editar o nó Filtro antes da execução.
- **Nó Seleção.** Cria um novo nó Seleção para selecionar registros com base na associação deles em qualquer um dos clusters visíveis na exibição atual do Visualizador de Cluster. Uma condição de seleção é gerada automaticamente.
- **Nó Seleção (da seleção).** Cria um novo nó Seleção para selecionar registros com base na associação nos clusters selecionados no Visualizador de Cluster. Selecione diversos clusters utilizando o método Ctrl-clique.
- **Nó Derivar.** Cria um novo nó Derivar, que deriva um campo flag que designa aos registros um valor de *True* ou *False* com base na associação em todos os clusters visíveis no Visualizador de Cluster. Uma condição de derivação é gerada automaticamente.
- **Nó Derivar (da seleção).** Cria um novo nó Derivar, que deriva um campo flag com base na associação nos clusters selecionados no Visualizador de Cluster. Selecione diversos clusters utilizando o método Ctrl-clique.

Além de gerar os nós, também é possível criar gráficos a partir do menu Gerar. Consulte o tópico "Gerando Gráficos a partir de Modelos de Cluster" na página 243 para obter mais informações.

Controlar Exibição da Visualização de Cluster

Para controlar o que é mostrado na visualização Clusters no painel principal, clique no botão **Exibir** que abre o diálogo Exibição.

Variáveis. Selecionada por padrão. Para ocultar todas as variáveis de entrada, desmarque a caixa de seleção.

Campos de Avaliação. Escolha os campos de avaliação (campos não utilizados para criar o modelo de cluster, mas enviados para o visualizador de modelos para avaliar os clusters) para exibição; nenhum é mostrado por padrão. *Nota* o campo de avaliação deve ser uma sequência com mais de um valor. Esta caixa de seleção estará indisponível se nenhum campo de avaliação estiver disponível.

Descrições de Cluster. Selecionada por padrão. Para ocultar todas as células de descrição de cluster, desmarque a caixa de seleção.

Tamanhos de Cluster. Selecionada por padrão. Para ocultar todas as células de tamanho de cluster, desmarque a caixa de seleção.

Número Máximo de Categorias. Especifique o número máximo de categorias a serem exibidas nos gráficos de variáveis categóricas; o padrão é 20.

Gerando Gráficos a partir de Modelos de Cluster

Os modelos de Cluster fornecem muitas informações; no entanto, essas informações nem sempre poderão estar em um formato facilmente acessível para usuários de negócios. Para fornecer os dados de modo que possam ser facilmente incorporados em relatórios de negócios, apresentações, e assim por diante, é possível produzir gráficos de dados selecionados. Na guia Visualizador de Cluster, é possível gerar um gráfico para um cluster selecionado, criando, assim, apenas um gráfico para os casos nesse cluster.

Nota: é possível gerar um gráfico no Visualizador do Cluster apenas quando o nugget do modelo estiver anexado a outros nós em um fluxo.

Gerar um gráfico

1. Abra o nugget do modelo contendo o Visualizador de Cluster.
2. Na guia Modelo, selecione *Clusters* na lista suspensa **Visualizar**.
3. Na visualização principal, selecione um ou mais clusters para os quais deseja gerar um gráfico.
4. No menu Gerar, selecione **Gráfico (da seleção)**; a guia Gráfico Básico é exibida.
   Nota: somente as guias Básico e Detalhado estão disponíveis quando exibir o Gráfico desta maneira.
5. Utilizando as configurações da guia Básico ou Detalhado, especifique os detalhes a serem exibidos no gráfico.
6. Clique em OK para gerar o gráfico.

O título do gráfico identifica o tipo de modelo e um ou mais clusters que foram escolhidos para inclusão.

Capítulo 12. Regras de Associação

As **Regras de associação** associam uma determinada conclusão (a compra de um certo produto, por exemplo) a um conjunto de condições (a compra de vários outros produtos, por exemplo). Por exemplo, a regra

```
beer <= cannedveg & frozenmeal (173, 17.0%, 0.84)
```

estabelece que geralmente *beer* ocorre quando *cannedveg* e *frozenmeal* ocorrem juntos. A regra é 84% confiável e se aplica a 17% dos dados, ou 173 registros. Os algoritmos de regra de associação localizam automaticamente as associações que puderam ser localizadas manualmente usando técnicas de visualização, como o nó da web .

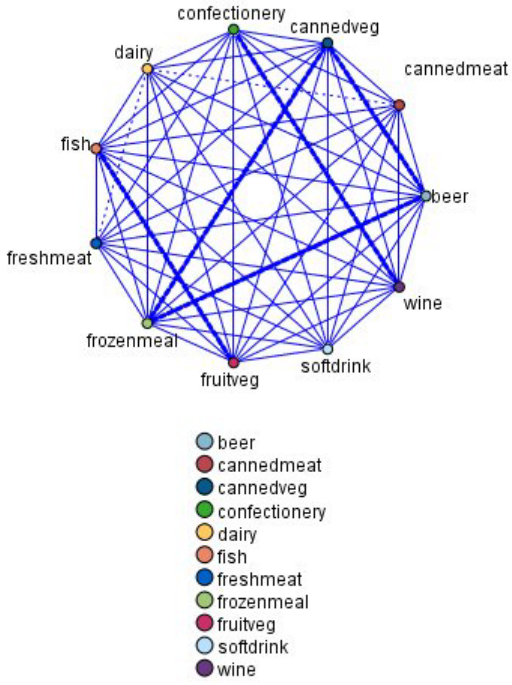


Figura 45. Nó da web mostrando associações entre os itens de cesta de mercado

A vantagem dos algoritmos de regra de associação sobre os algoritmos de árvore de decisão mais padrão (C5.0 e Árvores C&R) é que as associações podem existir entre *qualquer* um dos atributos. Um algoritmo de árvore de decisão construirá regras com apenas uma única conclusão, ao passo que os algoritmos de associação tentam localizar muitas regras, cada qual podendo ter uma conclusão diferente.

A desvantagem dos algoritmos de associação é que eles tentam localizar padrões dentro de um espaço de procura potencialmente muito grande e, como consequência, podem requerer muito mais tempo para executar do que um algoritmo de árvore de decisão. Os algoritmos usam um método **gerar e testar** para localizar regras -- regras simples são geradas inicialmente e validadas com relação ao conjunto de dados. As boas regras são armazenadas e todas as regras, sujeitas a várias restrições, são especializadas. **Especialização** é o processo de incluir condições em uma regra. Essas novas regras são, então, validadas com relação aos dados e o processo armazena iterativamente as melhores e mais interessantes regras localizadas. O usuário normalmente fornece algum limite para o número possível de antecedentes para permitir em uma regra e várias técnicas baseadas em teoria de informações ou esquemas de indexação eficientes são usadas para reduzir o espaço potencialmente grande de procura.

No término do processamento, uma tabela das melhores regras é apresentada. Ao contrário de uma árvore de decisão, esse conjunto de regras de associação não pode ser usado diretamente para fazer predições da forma com que um modelo padrão (como uma árvore de decisão ou uma rede neural) pode. Isso é devido a muitas conclusões possíveis diferentes para as regras. Outro nível de transformação é necessário para transformar as regras de associação em um conjunto de regras de classificação. Portanto, as regras de associação produzidas pelos algoritmos de associação são conhecidas como **modelos não refinados**. Embora o usuário possa procurar esses modelos não refinados, eles não podem ser usados explicitamente como modelos de classificação, a menos que o usuário diga ao sistema para gerar um modelo de classificação a partir do modelo não refinado. Isso é feito no navegador por meio de uma opção do menu Gerar.

Dois algoritmos de regra de associação são suportados:

O nó a priori extrai um conjunto de regras a partir dos dados, retirando as regras com o conteúdo mais alto de informações. O a priori oferece cinco métodos diferentes de selecionar regras e utiliza um esquema de indexação sofisticado para processar grandes conjuntos de dados de maneira eficiente. Para grandes problemas, o a priori geralmente é mais rápido para treinar e, além disso, ele não possui um limite arbitrário quanto ao número de regras que podem ser retidas e pode manipular regras com até 32 condições prévias. O a priori requer que todos os campos de entrada e de saída sejam categóricos e oferece melhor desempenho porque é otimizado para este tipo de dados.

O nó Sequência descobre as regras de associação nos dados sequenciais ou orientados a tempo. Uma sequência é uma lista de conjuntos de itens que tendem a ocorrer em uma ordem previsível. Por exemplo, um cliente que compra uma lâmina de barbear e uma loção pós-barba poderá comprar um creme de barbear na próxima compra. O nó Sequência baseia-se no algoritmo de regras de associação do CARMA que utiliza um método de duas passagens eficiente para localizar sequências.

Dados Tabulares versus Transacionais

Os dados utilizados pelos modelos de regras de associação podem estar em formato transacional ou tabular, conforme descrito abaixo. Estas são as descrições gerais; os requisitos específicos podem variar, conforme discutido na documentação para cada tipo de modelo. Observe que durante a escoragem de modelos, os dados a serem escorados devem espelhar o formato dos dados usados para a construção do modelo. Os modelos construídos usando dados tabulares podem ser usados para escorar apenas dados tabulares, e os modelos construídos usando dados transacionais podem escorar apenas dados transacionais.

Formato Transacional

Os dados transacionais possuem um registro separado para cada transação ou item. Se um cliente fizer diversas compras, por exemplo, cada uma seria um registro separado, com itens associados vinculados a um ID do cliente. Às vezes isso é conhecido também como formato **bobina de papel**.

Cliente	Compra
1	jam
2	milk
3	jam
3	bread
4	jam
4	bread
4	milk

Os nós a priori, CARMA e Sequência podem todos utilizar dados transacionais.

Dados Tabulares

Os dados tabulares (também conhecidos como dados de **cesta** ou **tabela da verdade**) têm os itens representados por flags separados, em que cada campo de flag representa a presença ou ausência de um item específico. Cada registro representa um conjunto completo de itens associados. Os campos de flag podem ser categóricos ou numéricos, embora determinados modelos possam ter requisitos mais específicos.

Cliente	Jam	Bread	Milk
1	X	F	F
2	F	F	X
3	X	X	F
4	X	X	X

Os nós a priori, CARMA, GSAR e Sequência podem todos utilizar dados tabulares.

Nó a priori

O nó a priori também descobre regras de associação nos dados. O a priori oferece cinco métodos diferentes de seleção de regras e usa um esquema de indexação sofisticado para processar conjuntos de dados grandes com eficiência.

Requisitos. Para criar um conjunto de regras a priori, um ou mais campos de *Entrada* e um ou mais campos de *Destino* são necessários. Os campos de entrada e de saída (aqueles com o papel *Entrada*, *Destino* ou *Ambos*) devem ser simbólicos. Campos com o papel *Nenhum* são ignorados. Os tipos de campos devem ser totalmente instanciados antes de executar o nó. Os dados podem estar em formato tabular ou transacional. Consulte o tópico “Dados Tabulares versus Transacionais” na página 246 para obter mais informações.

Intensidades. Para grandes problemas, o a priori é geralmente mais rápido para treinar. . Além disso, ele não tem um limite arbitrário no número de regras que podem ser retidas e pode manipular regras com até 32 condições prévias. O a priori oferece cinco métodos de treinamento diferentes, permitindo maior flexibilidade ao corresponder o método de mineração de dados com o problema em questão.

Opções de Modelo do Nó a priori

Nome do modelo É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Suporte mínimo a antecedente. É possível especificar um critério de suporte para manter as regras no conjunto de regras. **Suporte** refere-se à porcentagem de registros nos dados de treinamento para os quais os antecedentes (a parte "if" da regra) forem verdadeiros. (Observe que esta definição de suporte é diferente daquela utilizada nos nós CARMA e Sequência. Consulte o tópico “Opções do Modelo do Nó Sequência” na página 263 para obter mais informações). Se estiver obtendo regras que se aplicam a subconjuntos muito pequenos de dados, tente aumentar essa configuração.

Nota: A definição de suporte para a priori tem como base o número de registros com os antecedentes. Isso está em contraste com os algoritmos do CARMA e Sequência para os quais a definição de suporte tem como base o número de registros com todos os itens em uma regra (ou seja, ambos antecedentes e subsequentes). Os resultados para modelos de associação mostram as medidas de suporte (antecedente) e de suporte de regra.

Confiança mínima de regra. Também é possível especificar um critério de confiança. A **Confiança** baseia-se nos registros para os quais os antecedentes da regra são verdadeiros e é a porcentagem dos registros para os quais um ou mais subsequentes também são verdadeiros. Em outras palavras, é a porcentagem de preditores com base na regra que estão corretos. As regras com confiança menor que o critério especificado são descartadas. Se você estiver obtendo muitas regras, tente aumentar essa configuração. Se você estiver obtendo pouquíssimas regras (ou nenhuma regra), tente diminuir esta configuração.

Nota: Se necessário, é possível destacar o valor e digitar seu próprio valor. Lembre-se de que se reduzir o valor de confiança abaixo de 1,0, além do processo que requer muita memória livre, você poderá achar que as regras estão demorando um tempo extremamente longo para construir.

Número máximo de antecedentes. É possível especificar o número máximo de condições prévias para qualquer regra. Esta é uma maneira de limitar a complexidade das regras. Se as regras forem muito complexas ou muito específicas, tente diminuir esta configuração. Essa configuração também tem uma grande influência sobre o tempo de treinamento. Se o seu conjunto de regras estiver demorando muito tempo para treinar, tente reduzir essa configuração.

Somente valores reais para flags. Se essa opção for selecionada para dados em formato tabular (tabela da verdade), então apenas os valores reais serão incluídos nas regras resultantes. Isso pode ajudar a tornar as regras mais fáceis de entender. A opção não se aplica a dados em formato transacional. Consulte o tópico “Dados Tabulares versus Transacionais” na página 246 para obter mais informações.

Nota: O nó de construção de modelo CARMA ignorará registros vazios ao construir um modelo se o tipo de campo for um flag, ao passo que o nó de construção de modelo a priori inclui registros vazios. Registros vazios são registros em que todos os campos utilizados na construção de modelo possuem um valor false.

Otimizar. Selecione as opções projetadas para aumentar o desempenho durante a construção de modelo com base em suas necessidades específicas.

- Selecione **Velocidade** para instruir o algoritmo a nunca utilizar spilling de disco para melhorar o desempenho.
- Selecione **Memória** para instruir o algoritmo a utilizar spilling de disco quando apropriado em algum sacrificar de velocidade. Essa opção é selecionada por padrão.

  Nota: Ao executar no modo distribuído, essa configuração pode ser substituída pelas opções do administrador especificadas no arquivo *options.cfg*. Para obter mais informações, consulte o *Guia do Administrador do IBM SPSS Modeler Server* .

Opções Avançadas do Nó a priori

Para aqueles que possuem um conhecimento detalhado da operação a priori, as opções avançadas a seguir permitem fazer um ajuste preciso do processo de indução. Para acessar as opções avançadas, configure o Modo para **Especialista** na guia Especialista.

Medida de avaliação. O a priori suporta cinco métodos de avaliar possíveis regras.

- **Confiança de Regra.** O método padrão utiliza a confiança (ou a precisão) da regra para avaliar regras. Para essa medida, o **Limite inferior de medida de avaliação** é desativado, já que ele é redundante com a opção **Confiança mínima de regra** na guia Modelo. Consulte o tópico “Opções de Modelo do Nó a priori” na página 247 para obter mais informações.
- **Diferença de Confiança.** (Também chamada de **Diferença de confiança absoluta com a anterior**). Essa medida avaliação é a diferença absoluta entre a confiança da regra e sua confiança anterior. Esta opção evita um viés em que os resultados não são igualmente distribuídos. Isso ajuda a evitar que regras "óbvias" sejam mantidas. Por exemplo, pode ser o caso quando 80% dos clientes compram seus produtos mais populares. Uma regra que prediz a compra desse produto popular com 85% de precisão

não tem muito a acrescentar ao seu conhecimento, embora 85% possam parecer uma precisão muito boa em uma escala absoluta. Configure o limite inferior de medida de avaliação para medir a diferença mínima de confiança para a qual deseja que as regras sejam mantidas.

- **Razão de Confiança.** (Também chamada de **diferença de quociente de confiança para 1**) Esta medida de avaliação é a razão da confiança da regra com a confiança anterior (ou, se a razão for maior que um, com seu recíproco) menos 1. Assim como a Diferença de Confiança, esse método leva em consideração distribuições desiguais. Ela é ideal especialmente para localizar regras que preveem eventos raros. Por exemplo, suponha que haja uma condição médica rara que ocorre em apenas 1% dos pacientes. Uma regra que for capaz de prever essa condição 10% do tempo representa uma grande melhoria para a suposição aleatória, embora uma precisão de 10% possa não parecer muito expressiva em uma escala absoluta. Configure o limite inferior de medida de avaliação para a diferença para a qual deseja que as regras sejam mantidas.
- **Diferença de informações.** (Também chamada de **Diferença de informações com a anterior**). Esta medida baseia-se na medida de **ganho de informações**. Se a probabilidade de um determinado subsequente for considerada como um valor lógico (um **bit**), então o ganho de informações será a proporção desse bit que pode ser determinada com base nos antecedentes. A diferença de informações é a diferença entre o ganho de informações, dado os antecedentes, e o ganho de antecedentes, dada apenas a confiança anterior do subsequente. Um recurso importante deste método é que ele leva em consideração o suporte para que as regras que cobrem mais registros sejam preferenciais para um determinado nível de confiança. Configure o limite inferior de medida de avaliação para a diferença de informações para a qual deseja que as regras sejam mantidas.

  Nota: como a escala para esta medida é um pouco menos intuitiva do que as outras escalas, poderá ser necessário experimentar limites inferiores diferentes para obter um conjunto de regras satisfatório.
- **Qui-quadrado Normalizado.** (Também chamado de **medida qui-quadrada normalizada**). Essa medida é um índice estatístico da associação entre os antecedentes e os subsequentes. A medida é normalizada para assumir valores entre 0 e 1. Esta medida depende muito mais do suporte do que da medida de diferença de informações. Configure o limite inferior de medida de avaliação para a diferença de informações para a qual deseja que as regras sejam mantidas.

  Nota: assim como a medida de diferença de informações, a escala para esta medida é um pouco menos intuitiva do que as outras escalas, portanto, poderá ser necessário experimentar limites inferiores diferentes para obter um conjunto de regras satisfatório.

Permitir regras sem antecedentes. Selecione para permitir regras que incluam apenas o subsequente (item ou conjunto de itens). Isso será útil quando estiver interessado em determinar um item ou conjuntos de itens comuns. Por exemplo, `cannedveg` é uma regra de um único item sem um antecedente que indica que comprar *cannedveg* é uma ocorrência comum nos dados. Em alguns casos, você poderá querer incluir essas regras se estiver interessado apenas nas predições mais confiantes. Essa opção está desativada por padrão. Por convenção, o suporte de antecedente para regras sem antecedentes é expresso como 100%, e o suporte de regra será o mesmo que a confiança.

Nó CARMA

O nó CARMA usa um algoritmo de descoberta de regras de associação para descobrir regras de associação nos dados. As regras de associação são instruções no formato

if *antecedent(s)* **then** *consequent(s)*

Por exemplo, se um cliente da web comprar uma placa wireless e um roteador wireless moderno, o cliente também poderá comprar um servidor de música wireless, se oferecido. O modelo do CARMA extrai um conjunto de regras de dados sem a necessidade de especificar campos de entrada ou de destino. Isso significa que as regras geradas podem ser utilizadas para uma variedade maior de aplicações. Por exemplo, é possível utilizar regras geradas por esse nó para localizar uma lista de produtos ou serviços (antecedentes) cujo subsequente é o item que você deseja promover neste período de férias. Usando o IBM SPSS Modeler, é possível determinar quais clientes compraram os produtos antecedentes e construir uma campanha de marketing designada para promover o produto subsequente.

Requisitos. Em contraste com o a priori, o nó CARMA não requer campos de *Entrada* ou de *Destino*. Isso é essencialmente o modo com que o algoritmo funciona e é equivalente a construir um modelo a priori com todos os campos configurados para *Ambos*. É possível restringir quais itens são listados apenas como antecedentes ou subsequentes ao filtrar o modelo após ele ter sido construído. Por exemplo, é possível utilizar o navegador do modelo para localizar uma lista de produtos ou serviços (antecedentes) cujo subsequente é o item que você deseja promover neste período de férias.

Para criar um conjunto de regras do CARMA, é necessário especificar um campo de ID e um ou mais campos de conteúdo. O campo de ID pode ter qualquer nível de papel ou medição. Campos com o papel *Nenhum* são ignorados. Os tipos de campo devem ser totalmente instanciados antes de executar o nó. Assim como o a priori, os dados podem estar em formato tabular ou transacional. Consulte o tópico "Dados Tabulares versus Transacionais" na página 246 para obter mais informações.

Intensidades. O nó CARMA baseia-se no algoritmo de regras de associação CARMA. Em contraste com o a priori, o nó CARMA oferece configurações de construção para suporte de regra (suporte para antecedente e subsequente) ao invés de para suporte de antecedente. O CARMA também permite regras com diversos subsequentes. Assim como o a priori, os modelos gerados por um nó do CARMA podem ser inseridos em um fluxo de dados para criar predições. Consulte o tópico "Nuggets do Modelo" na página 38 para obter mais informações.

Opções de Campos do Nó CARMA

Antes de executar um nó CARMA, deve-se especificar os campos de entrada na guia Campos do nó CARMA. Ao passo que a maioria dos nós de modelagem compartilha opções idênticas da guia Campos, o nó CARMA contém várias opções exclusivas. Todas as opções são discutidas abaixo.

Usar configurações do nó Tipo. Essa opção instrui o nó a utilizar as informações de campo a partir de um nó Tipo de envio de dados. Esse é o padrão.

Usar configurações customizadas. Essa opção instrui o nó a utilizar as informações de campo especificadas aqui ao invés das informações fornecidas em qualquer nó ou nós Tipo de envio de dados. Após selecionar esta opção, especifique os campos abaixo de acordo com se você estiver lendo dados em formato transacional ou tabular.

Usar formato transacional. Esta opção altera os controles do campo no restante desta caixa de diálogo, dependendo se seus dados estiverem em formato tabular ou transacional. Se você utilizar diversos campos com dados transacionais, os itens especificados nestes campos para um registro específico serão considerados para representar os itens localizados em uma única transação com um único registro de data e hora. Consulte o tópico "Dados Tabulares versus Transacionais" na página 246 para obter mais informações.

Dados tabulares

Se **Usar formato transacional** não for selecionada, os campos a seguir serão exibidos.

- **Entradas.** Selecione um ou mais campos de entrada. Isso é semelhante a configurar o papel do campo para *Entrada* em um nó Tipo.
- **Partição.** Este campo permite especificar um campo utilizado para particionar os dados em amostras separadas para os estágios de treinamento, de teste e de validação de construção de modelo. Ao utilizar uma amostra para gerar o modelo e uma amostra diferente para testá-lo, é possível obter uma boa indicação do quão bem o modelo será generalizado para conjuntos de dados maiores que forem semelhantes aos dados atuais. Se diversos campos de partição tiverem sido definidos usando os nós Tipo ou Partição, um campo de partição único deverá ser selecionado na guia Campos em cada nó de modelagem que utiliza particionamento. (Se apenas uma partição estiver presente, ela será utilizada automaticamente sempre que o particionamento estiver ativado). Além disso, observe que para aplicar

a partição selecionada à sua análise, o particionamento também deverá ser ativado na guia Opções de Modelo para o nó. (Desmarcar esta opção permite desativar o particionamento sem alterar as configurações do campo).

Dados transacionais

Se selecionar **Usar formato transacional**, os campos a seguir serão exibidos.

- **ID.** Para dados transacionais, selecione um campo de ID na lista. Campos numéricos ou simbólicos podem ser utilizados como o campo de ID. Cada valor exclusivo deste campo deve indicar uma unidade específica de análise. Por exemplo, em um aplicativo de cesta de mercado, cada ID pode representar um cliente único. Para um aplicativo de análise de log da web, cada ID pode representar um computador (pelo endereço IP) ou um usuário (pelos dados de login).
- **IDs são contíguos.** (apenas nós a priori e CARMA) Se seus dados estiverem pré-ordenados de forma que todos os registros com o mesmo ID sejam agrupados no fluxo de dados, selecione esta opção para acelerar o processamento. Se seus dados não estiverem pré-ordenados (ou se você não tiver certeza), deixe essa opção desmarcada e o nó ordenará os dados automaticamente.

  Nota: se seus dados não estiverem classificados e você selecionar essa opção, resultados inválidos poderão ser obtidos no seu modelo.
- **Conteúdo.** Especifique um ou mais campos de conteúdo para o modelo. Esses campos contêm os itens de interesse na modelagem de associação. É possível especificar diversos campos de sinalização (se os dados estiverem em formato tabular) ou um campo nominal único (se os dados estiverem em formato transacional).

Opções de Modelo do Nó CARMA

Nome do modelo É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Suporte mínimo de regra (%). É possível especificar um critério de suporte. **Suporte de regra** refere-se à proporção de IDs nos dados de treinamento que contêm a regra inteira. (Observe que esta definição de suporte é diferente do suporte de antecedente utilizado nos nós a priori). Se desejar focar em regras mais comuns, aumente essa configuração.

Confiança mínima de regra (%). É possível especificar um critério de confiança para manter as regras no conjunto de regras. **Confiança** refere-se à porcentagem de IDs na qual uma predição correta é feita (dentre todos os IDs para os quais a regra faz uma predição). Ela é calculada como o número de IDs para os quais a regra inteira é localizada dividido pelo número de IDs para os quais os antecedentes são localizados, com base nos dados de treinamento. As regras com confiança menor que o critério especificado são descartadas. Se estiver obtendo regras desinteressantes ou muitas regras, tente aumentar essa configuração. Se estiver obtendo pouquíssimas regras, tente diminuir esta configuração.

Nota: Se necessário, é possível destacar o valor e digitar seu próprio valor. Lembre-se de que se reduzir o valor de confiança abaixo de 1,0, além do processo que requer muita memória livre, você poderá achar que as regras estão demorando um tempo extremamente longo para construir.

Tamanho máximo da regra. É possível configurar o número máximo de `conjuntos de itens` distintos (ao contrário de `itens`) em uma regra. Se as regras de interesse forem relativamente curtas, será possível diminuir esta configuração para acelerar a construção do conjunto de regras.

Nota: O nó de construção de modelo CARMA ignorará registros vazios ao construir um modelo se o tipo de campo for um flag, ao passo que o nó de construção de modelo a priori inclui registros vazios. Registros vazios são registros em que todos os campos utilizados na construção de modelo possuem um valor false.

Opções Avançadas do Nó CARMA

Para aqueles que possuem um conhecimento detalhado da operação do nó CARMA, as opções avançadas a seguir permitem fazer um ajuste preciso do processo de construção de modelo. Para acessar as opções avançadas, configure o modo para **Especialista** na guia Especialista.

Excluir regras com diversos subsequentes. Selecione para excluir subsequentes de "duas cabeças", ou seja, subsequentes que contêm dois itens. Por exemplo, a regra `bread & cheese & fish -> wine&fruit` contém um subsequente de duas cabeças, `wine&fruit`. Por padrão, tais regras são incluídas.

Configurar o valor de poda. Para conservar memória, o algoritmo CARMA utilizado periodicamente remove (**poda**) conjuntos de itens frequentes de sua lista de conjuntos de itens possíveis durante o processamento. Selecione esta opção para ajustar a frequência de poda e o número que você especificar determina a frequência da poda. Insira um valor menor para diminuir os requisitos de memória do algoritmo (e potencialmente aumentar o tempo de treinamento necessário), ou insira um valor maior para acelerar o treinamento (e potencialmente aumentar os requisitos de memória). O valor padrão é 500.

Variar suporte. Selecione para aumentar a eficiência ao excluir conjuntos de itens incomuns que parecem ser frequentes quando forem incluídos de maneira desigual. Isso é feito ao iniciar com um nível de suporte maior e reduzi-lo até o nível especificado na guia Modelo. Insira um valor para **Número estimado de transações** para especificar a velocidade com que o nível de suporte deve ser reduzido.

Permitir regras sem antecedentes. Selecione para permitir regras que incluam apenas o subsequente (item ou conjunto de itens). Isso será útil quando estiver interessado em determinar um item ou conjuntos de itens comuns. Por exemplo, `cannedveg` é uma regra de um único item sem um antecedente que indica que comprar *cannedveg* é uma ocorrência comum nos dados. Em alguns casos, você poderá querer incluir essas regras se estiver interessado apenas nas predições mais confiantes. Esta opção está desmarcada por padrão.

Nuggets do Modelo de Regra de Associação

Os nuggets do modelo de regra de associação representam as regras descobertas por um dos nós de modelagem de regra de associação a seguir:

- a priori
- CARMA

Os nuggets do modelo contêm informações sobre as regras extraídas de seus dados durante a construção de modelo.

Nota: A escoragem do nugget de regra de associação poderá ser incorreta se você não ordenar os dados transacionais por ID.

Visualizando Resultados

É possível procurar as regras geradas pelos modelos de associação (a priori, CARMA) e modelos de sequência utilizando a guia Modelo na caixa de diálogo. Procurar um nugget do modelo mostra as informações sobre as regras e fornece opções para filtrar e ordenar os resultados antes de gerar novos nós ou escorar o modelo.

Escoragem do Modelo

Nuggets do modelo refinados (a priori, CARMA e Sequência) podem ser incluídos em um fluxo e usados para escoragem. Consulte o tópico "Utilizando Nuggets do Modelo em Fluxos" na página 49 para obter mais informações. Os nuggets do modelo utilizados para escoragem incluem uma guia Configurações extra em suas respectivas caixas de diálogo. Consulte o tópico "Configurações do Nugget de Modelo de Regra de Associação" na página 256 para obter mais informações.

Um nugget do modelo não refinado não pode ser usado para escoragem em seu formato bruto. Ao invés disso, é possível gerar um conjunto de regras e utilizá-lo para escoragem. Consulte o tópico “Gerando um Conjunto de Regras a partir de um Nugget do Modelo de Associação” na página 257 para obter mais informações.

Detalhes do Nugget de Modelo de Regra de Associação

Na guia Modelo de um nugget do modelo de Regra de Associação, é possível ver uma tabela contendo as regras extraídas pelo algoritmo. Cada linha na tabela representa uma regra. A primeira coluna representa os subsequentes (a parte "then" da regra), ao passo que a próxima coluna representa os antecedentes (a parte "if" da regra). As colunas subsequentes contêm informações da regra, como confiança, suporte e elevação.

As regras de associação são frequentemente mostradas no formato na tabela a seguir.

Tabela 13. Exemplo de uma regra de associação

Subsequente	Antecedente
`Drug = drugY`	`Sex = F` `BP = HIGH`

A regra de exemplo é interpretada como *if Sex = "F" and BP = "HIGH," then Drug is likely to be drugY*, ou pode ser escrita de outra forma, como *for records where Sex = "F" and BP = "HIGH," Drug is likely to be drugY*. Utilizando a barra de ferramentas da caixa de diálogo, é possível optar por exibir informações adicionais, como confiança, suporte e instâncias.

Menu Ordenar. O botão do menu Ordenar na barra de ferramentas controla a ordenação das regras. A direção da ordenação (crescente ou decrescente) pode ser alterada utilizando o botão de direção de ordenação (seta para cima ou para baixo).

As regras podem ser ordenadas por:

- Suporte
- Confiança
- Suporte de Regra
- Subsequente
- Elevar
- Implementabilidade

Menu Mostrar/Ocultar. O menu Mostrar/Ocultar (botão da barra de ferramentas de critérios) controla as opções para a exibição de regras.

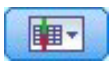


Figura 46. Botão Mostrar/Ocultar

As opções de exibição a seguir estão disponíveis:

- **ID de Regra** exibe o ID da regra designada durante a construção de modelo. Um ID de regra permite identificar quais regras estão sendo aplicadas para uma determinada predição. Os IDs de regra também permitem mesclar posteriormente informações de regras adicionais, como implementabilidade e informações do produto ou de antecedentes.
- **Instâncias** exibe informações sobre o número de IDs exclusivos aos quais a regra se aplica -- ou seja, para os quais os antecedentes são verdadeiros. Por exemplo, dada a regra `bread -> cheese`, o número de registros nos dados de treinamento que incluem o antecedente *bread* é referido como **instâncias**.

- **Suporte** exibe o suporte de antecedente - ou seja, a proporção de IDs para os quais os antecedentes são verdadeiros, com base nos dados de treinamento. Por exemplo, se 50% dos dados de treinamento incluírem a compra de pão, então a regra `bread -> cheese` terá um suporte de antecedente de 50%. *Nota*: o suporte conforme definido aqui é o mesmos que as instâncias, mas é representado como uma porcentagem.
- **Confiança** exibe a razão do suporte de regra para o suporte de antecedente. Isso indica a proporção de IDs com um ou mais antecedentes especificados para os quais um ou mais subsequentes são também verdadeiros. Por exemplo, se 50% dos dados de treinamento contiverem pão (indicando o suporte de antecedente), mas apenas 20% contiverem ambos pão e queijo (indicando o suporte de regra), então a confiança para a regra `bread -> cheese` será `Rule Support / Antecedent Support`, ou seja, 40%.
- **Suporte de Regra** exibe a proporção de IDs para os quais a regra inteira, antecedentes e um ou mais subsequentes são verdadeiros. Por exemplo, se 20% dos dados de treinamento contiverem pão e queijo, então o suporte de regra para a regra `bread -> cheese` será 20%.
- **Elevação** exibe a razão de confiança para a regra para a probabilidade anterior de ter o subsequente. Por exemplo, se 10% de toda a população compra pão, então a regra que prediz se as pessoas comprarão ou não pão com 20% de confiança terá um aumento de 20/10 = 2. Se outra regra informar que as pessoas comprarão pão com 11% de confiança, então a regra possuirá uma elevação de aproximadamente 1, o que significa que ter um ou mais antecedentes não faz muita diferença na probabilidade de ter o subsequente. Em geral, regras com uma elevação diferente de 1 serão mais interessantes do que as regras com elevação próxima de 1.
- **Implementabilidade** é uma medida de qual porcentagem dos dados de treinamento satisfaz as condições do antecedente, mas não satisfaz o subsequente. Em termos de compra de produto, isso significa basicamente qual porcentagem da base total de clientes possui (ou comprou) o(s) antecedente(s), mas ainda não comprou o subsequente. A estatística de implementabilidade é definida como `((Antecedent Support in # of Records - Rule Support in # of Records) / Number of Records) * 100`, em que *Antecedent Support* significa o número de registros para os quais os antecedentes são verdadeiros e *Rule Support* significa o número de registros para os quais ambos antecedentes e subsequentes são verdadeiros.

Botão Filtrar. O botão Filtrar (ícone de funil) no menu expande a parte inferior da caixa de diálogo para mostrar um painel no qual os filtros de regra ativos são exibidos. Os filtros são utilizados para limitar o número de regras exibidas na guia Modelos.

Figura 47. Botão Filtrar

Para criar um filtro, clique no ícone Filtro à direita do painel expandido. Isso abre uma caixa de diálogo separada na qual é possível especificar restrições para exibir regras. Observe que o botão Filtrar é geralmente utilizado em conjunto com o menu Gerar para primeiro filtrar regras e, em seguida, gerar um modelo contendo esse subconjunto de regras. Para obter mais informações, consulte “Especificando Filtros para Regras” na página 255 abaixo.

Botão Localizar Regra. O botão Localizar Regra (ícone de binóculo) permite procurar as regras mostradas para um ID de regra especificado. A caixa de exibição adjacente indica o número de regras exibidas atualmente fora do número disponível. Os IDs de regra são designados pelo modelo na ordem de descoberta no tempo e são incluídos nos dados durante a escoragem.

Figura 48. Botão Localizar Regra

Para reordenar os IDs de regras:

1. É possível reorganizar os IDs de regras no IBM SPSS Modeler ao primeiro ordenar a tabela de exibição de regra de acordo com a medição desejada, como confiança ou elevação.
2. Em seguida, crie um modelo filtrado utilizando as opções do menu Gerar.
3. Na caixa de diálogo Modelo Filtrado, selecione **Renumerar regras consecutivamente, começando com** e especifique um número inicial.

Consulte o tópico "Gerando um Modelo Filtrado" na página 258 para obter mais informações.

Especificando Filtros para Regras

Por padrão, os algoritmos de regra, como a priori, CARMA e Sequência, podem gerar um grande e inconveniente número de regras. Para aprimorar a clareza ao procurar ou aperfeiçoar a escoragem de regra, deve-se considerar as regras de filtragem para que os subsequentes e antecedentes de interesse sejam exibidos de modo mais proeminente. Utilizando as opções de filtragem na guia Modelo de um navegador de regra, é possível abrir uma caixa de diálogo para especificar qualificações de filtro.

Subsequentes. Selecione **Ativar Filtro** para ativar as opções para filtragem de regras com base na inclusão ou exclusão de subsequentes especificados. Selecione **Inclui quaisquer** para criar um filtro no qual as regras contêm pelo menos um dos subsequentes especificados. Como alternativa, selecione **Exclui** para criar um filtro que exclui subsequentes especificados. É possível selecionar subsequentes utilizando o ícone do selecionador à direita da caixa de listagem. Isso abre uma caixa de diálogo listando todos os subsequentes presentes nas regras geradas.

Nota: os subsequentes podem conter mais de um item. Os filtros apenas verificam se um subsequente contém um dos itens especificados.

Antecedentes. Selecione **Ativar Filtro** para ativar as opções para filtragem de regras com base na inclusão ou exclusão dos antecedentes especificados. É possível selecionar itens utilizando o ícone do selecionador à direita da caixa de listagem. Isso abre uma caixa de diálogo listando todos os antecedentes presentes nas regras geradas.

- Selecione **Inclui todos** para configurar o filtro como um inclusivo no qual todos os antecedentes especificados devem ser incluídos em uma regra.
- Selecione **Inclui quaisquer** para criar um filtro no qual as regras contêm pelo menos um dos antecedentes especificados.
- Selecione **Exclui** para criar um filtro que exclui regras que contêm um antecedente especificado.

Confiança. Selecione **Ativar Filtro** para ativar as opções para filtragem de regras com base no nível de confiança de uma regra. É possível utilizar os controles **Mín.** e **Máx.** para especificar um intervalo de confiança. Quando estiver procurando modelos gerados, a confiança é listada como uma porcentagem. Quando escorar a saída, a confiança é expressa como um número entre 0 e 1.

Suporte de Antecedente. Selecione **Ativar Filtro** para ativar as opções para filtragem de regras com base no nível de suporte de antecedente de uma regra. O suporte de antecedente indica a proporção de dados de treinamento que contém os mesmos antecedentes que a regra atual, tornando-o semelhante a um índice de popularidade. É possível utilizar os controles **Mín.** e **Máx.** para especificar um intervalo utilizado para filtrar regras com base no nível de suporte.

Elevação. Selecione **Ativar Filtro** para ativar as opções de filtragem de regras com base na medição de elevação de uma regra. *Nota*: a filtragem de elevação está disponível apenas para modelos de associação construídos após a liberação 8.5 ou para modelos anteriores que contêm uma medição de elevação. Os modelos de sequência não contêm essa opção.

Clique em **OK** para aplicar todos os filtros que foram ativados nesta caixa de diálogo.

Gerando Gráficos de Regras

Os nós Associação fornecem muitas informações; no entanto, essas informações nem sempre poderão estar em um formato facilmente acessível para usuários de negócios. Para fornecer os dados de modo que possam ser facilmente incorporados em relatórios de negócios, apresentações, e assim por diante, é possível produzir gráficos de dados selecionados. Na guia Modelo, é possível gerar um gráfico para uma regra selecionada, criando, assim, apenas um gráfico para os casos nessa regra.

1. Na guia Modelo, selecione a regra na qual você está interessado.
2. No menu Gerar, selecione **Gráfico (da seleção)**. A guia Gráfico Básico é exibida.

   Nota: somente as guias Básico e Detalhado estão disponíveis quando exibir o Gráfico desta maneira.
3. Utilizando as configurações da guia Básico ou Detalhado, especifique os detalhes a serem exibidos no gráfico.
4. Clique em OK para gerar o gráfico.

O título do gráfico identifica a regra e os detalhes do antecedente que foram escolhidos para inclusão.

Configurações do Nugget de Modelo de Regra de Associação

Esta guia Configurações é utilizada para especificar as opções de escoragem para modelos de associação (a priori e CARMA). Esta guia estará disponível somente após o nugget do modelo ter sido incluído em um fluxo para propósitos de escoragem.

Nota: A caixa de diálogo para procurar um modelo não refinado não inclui a guia Configurações, uma vez que ele não pode ser escorado. Para escorar o modelo "não refinado", deve-se primeiro gerar um conjunto de regras. Consulte o tópico “Gerando um Conjunto de Regras a partir de um Nugget do Modelo de Associação” na página 257 para obter mais informações.

Número máximo de predições Especifique o número máximo de predições que são incluídas para cada conjunto de itens de cesta. Esta opção é utilizada em conjunto com o Critério de Regra abaixo para produzir as “principais” predições, em que *principais* indica o nível mais alto de suporte de confiança, elevação, e assim por diante, conforme especificado abaixo.

Critério de Regra Selecione a medida usada para determinar a intensidade das regras. As regras são classificadas pela intensidade dos critérios selecionados aqui para retornar as principais predições para um conjunto de itens. Os critérios disponíveis são mostrados na lista a seguir.

- Confiança
- Suporte
- Suporte de regra (Suporte * Confiança)
- Elevar
- Implementabilidade

Permitir predições de repetição Selecione para incluir diversas regras com o mesmo subsequente quando escorar. Por exemplo, selecionar essa opção permite que as regras a seguir sejam escoradas:

```
bread & cheese -> wine
cheese & fruit -> wine
```

Desative esta opção para excluir predições repetidas durante a escoragem.

Nota: Regras com diversos subsequentes (`bread & cheese & fruit -> wine & pate`) serão consideradas predições repetidas apenas se todos os subsequentes (`wine & pate`) tiverem sido preditos antes.

Ignorar itens de cesta não correspondidos Selecione para ignorar a presença de itens adicionais no conjunto de itens. Por exemplo, quando essa opção é selecionada para uma cesta contendo `[tent & sleeping bag & kettle]`, a regra `tent & sleeping bag-> gas_stove` será aplicada, apesar do item extra (`kettle`) presente na cesta.

Pode haver algumas circunstâncias em que itens extras devem ser excluídos. Por exemplo, é provável que alguém que compra uma barraca, um saco de dormir e uma chaleira já possa ter um fogão a gás, indicado pela presença da chaleira. Em outras palavras, um fogão a gás pode não ser a melhor predição. Nesses casos, deve-se desmarcar **Ignorar itens de cesta não correspondidos** para assegurar que os antecedentes da regra correspondam exatamente ao conteúdo de uma cesta. Por padrão, itens não correspondidos são ignorados.

Verificar se as predições não estão na cesta. Selecione para assegurar que os antecedentes também não estejam presentes na cesta. Por exemplo, se o propósito da escoragem é fazer uma recomendação de móveis domésticos, então é improvável que uma cesta que já contenha uma mesa de jantar compre outra mesa. Nesse caso, deve-se selecionar esta opção. Por outro lado, se os produtos forem perecíveis ou descartáveis (como queijo, alimento para bebês ou papel higiênico), então as regras nas quais o subsequente já estiver presente na cesta poderão ser úteis. Nesse último caso, a opção mais útil pode ser **Não verificar predições na cesta** abaixo.

Verificar se as predições estão na cesta Selecione esta opção para assegurar que os subsequentes também estejam presentes na cesta. Essa abordagem é útil quando você estiver tentando ganhar insight sobre clientes ou transações existentes. Por exemplo, você pode querer identificar regras com a elevação mais alta e, em seguida, explorar quais clientes se enquadram nessas regras.

Não verificar predições na cesta Selecione para incluir todas as regras durante a escoragem, independentemente da presença ou ausência de subsequentes na cesta.

Gerar SQL para este modelo Ao usar dados de um banco de dados, código SQL pode ser enviado por push de volta para o banco de dados para execução, fornecendo desempenho superior para muitas operações.

Selecione uma das opções a seguir para especificar como a geração de SQL é executada.

- **Padrão: Escorar usando o Server Scoring Adapter (se instalado) no processo** Se conectado a um banco de dados com um adaptador de escoragem instalado, gera a SQL utilizando o adaptador de escoragem e funções definidas pelo usuário (UDF) associadas e escora seu modelo no banco de dados. Quando nenhum adaptador de escoragem estiver disponível, essa opção busca seus dados novamente a partir do banco de dados e os escora no SPSS Modeler.
- **Escorar fora do Banco de dados** Se selecionada, esta opção busca seus dados novamente a partir do banco de dados e os escora no SPSS Modeler.

Sumarização do Nugget de Modelo de Regra de Associação

A guia Sumarização de um nugget do modelo de regra de associação exibe o número de regras descobertas e o mínimo e máximo de suporte, elevação, confiança e implementabilidade de regras no conjunto de regras.

Gerando um Conjunto de Regras a partir de um Nugget do Modelo de Associação

Os nuggets do modelo de associação, como a priori e CARMA, podem ser utilizados para escorar dados diretamente, ou é possível primeiro gerar um subconjunto de regras, conhecido como um **conjunto de regras**. Os conjuntos de regras são particularmente úteis quando estiver trabalhando com um modelo não refinado, que não pode ser utilizado diretamente para escoragem. Consulte o tópico “Modelos Não Refinados” na página 52 para obter mais informações.

Para gerar um conjunto de regras, escolha **Conjunto de regras** no menu Gerar no navegador de nugget do modelo. É possível especificar as opções a seguir para converter as regras em um conjunto de regras:

Nome do conjunto de regras. Permite especificar o nome do novo nó Conjunto de Regras.

Criar nó em. Controla o local do novo nó Conjunto de Regras gerado. Selecione **Tela**, **Paleta GM** ou **Ambos**.

Campo de destino. Determina qual campo de saída será utilizado para o nó Conjunto de Regras gerado. Selecione um campo de saída único na lista.

Suporte mínimo. Especifique o suporte mínimo de regras a serem preservadas no conjunto de regras gerado. As regras com suporte menor que o valor especificado não serão incluídas no novo conjunto de regras.

Confiança mínima. Especifique a confiança mínima para regras a serem preservadas no conjunto de regras gerado. As regras com confiança menor que o valor especificado não serão incluídas no novo conjunto de regras.

Valor padrão. Permite especificar um valor padrão para o campo de destino que é designado a registros escorados para os quais nenhuma regra é disparada.

Gerando um Modelo Filtrado

Para gerar um modelo filtrado a partir de um nugget do modelo de associação, como um nó a priori, CARMA ou Conjunto de Regras de Sequência, escolha **Modelo Filtrado** a partir do menu Gerar no navegador de nugget do modelo. Isso cria um modelo de subconjunto que inclui somente as regras exibidas atualmente no navegador. *Nota*: não é possível gerar modelos filtrados para modelos não refinados.

É possível especificar as seguintes opções para filtragem de regras:

Nome do Novo Modelo. Permite especificar o nome do novo nó Modelo Filtrado.

Criar nó em. Controla o local do novo nó Modelo Filtrado. Selecione **Tela**, **Paleta GM** ou **Ambos**.

Numeração de regra. Especifique como os IDs de regra serão numerados no subconjunto de regras incluído no modelo filtrado.

- **Reter números de ID de regra original.** Selecione para manter a numeração original de regras. Por padrão, as regras recebem um ID correspondente à sua ordem de descoberta pelo algoritmo. Essa ordem pode variar dependendo do algoritmo utilizado.
- **Numerando novamente regras de modo consecutivo iniciando com.** Selecione para designar novos IDs de regra para as regras filtradas. Novos IDs são designados com base na ordenação exibida na tabela do navegador de regras na guia Modelo, iniciando com o número que você especificar aqui. É possível especificar o número de início para os IDs usando as setas à direita.

Escorando Regras de Associação

Os escores produzidos pela execução de novos dados por meio de um nugget do modelo de regra de associação são retornados em campos separados. Três novos campos são incluídos para cada predição, com *P* representando a predição, *C* representando a confiança e *I* representando o ID de regra. A organização desses campos de saída depende se os dados de entrada estão em formato tabular ou transacional. Consulte “Dados Tabulares versus Transacionais” na página 246 para obter uma visão geral desses formatos.

Por exemplo, suponha que você esteja escorando dados de cesta utilizando um modelo que gera predições com base nas três regras a seguir:

```
Rule_15  bread&wine -> meat (confidence 54%)
Rule_22  cheese -> fruit (confidence 43%)
Rule_5  bread&cheese -> frozveg (confidence 24%)
```

Dados tabulares. Para dados tabulares, três predições (3 é o padrão) são retornadas em um único registro.

Tabela 14. Escores em formato tabular.

ID	Bread	Wine	Cheese	P1	C1	I1	P2	C2	I2	P3	C3	I3
Fred	1	1	1	meat	0.54	15	fruit	0.43	22	frozveg	.24	5

Dados transacionais. Para dados transacionais, um registro separado é gerado para cada predição. As predições ainda são incluídas em colunas separadas, mas os escores são retornados conforme eles são calculados. Isso resulta em registros com predições incompletas, conforme mostrado na saída de amostra a seguir. As segunda e terceira predições (P2 e P3) estão em branco no primeiro registro, com as confianças e os IDs de regras associados. No entanto, conforme os escores são retornados, o registro final conterá todas as três predições.

Tabela 15. Escores em formato transacional.

ID	Item	P1	C1	I1	P2	C2	I2	P3	C3	I3
Fred	bread	meat	0.54	14	$null$	$null$	$null$	$null$	$null$	$null$
Fred	cheese	meat	0.54	14	fruit	0.43	22	$null$	$null$	$null$
Fred	wine	meat	0.54	14	fruit	0.43	22	frozveg	0.24	5

Para incluir apenas predições completas para fins de relatório ou de implementação, utilize um nó Seleção para selecionar registros completos.

Nota: os nomes dos campos utilizados nestes exemplos são abreviados para maior clareza. Durante o uso real, os campos de resultados para modelos de associação são denominados conforme mostrado na tabela a seguir.

Tabela 16. Nomes de campos de resultados para modelos de associação.

Novo campo	Exemplo de nome de campo
Predição	*$A-TRANSACTION_NUMBER-1*
Confiança (ou outro critério)	*$AC-TRANSACTION_NUMBER-1*
ID da regra	*$A-Rule_ID-1*

Regras com Diversos Subsequentes

O algoritmo CARMA permite regras com diversos subsequentes, por exemplo:

```
bread -> wine&cheese
```

Quando estiver escorando essas regras de "duas cabeças", as predições serão retornadas no formato exibido na tabela a seguir.

Tabela 17. Resultados de escoragem, incluindo uma predição com diversos subsequentes.

ID	Bread	Wine	Cheese	P1	C1	I1	P2	C2	I2	P3	C3	I3
Fred	1	1	1	meat&veg	0.54	16	fruit	0.43	22	frozveg	.24	5

Em alguns casos, poderá ser necessário dividir tais escores antes da implementação. Para dividir uma predição com diversos subsequentes, será necessário analisar o campo utilizando as funções de sequência de caracteres do CLEM.

Implementando Modelos de Associação

Ao escorar modelos de associação, as predições e as confianças são geradas em colunas separadas (em que *P* representa a predição, *C* representa a confiança e *I* representa o ID da regra). Este será o caso, independentemente se os dados de entrada forem tabulares ou transacionais. Consulte o tópico "Escorando Regras de Associação" na página 258 para obter mais informações.

Ao preparar os escores para implementação, você poderá achar que seu aplicativo está exigindo que você transponha seus dados de saída para um formato com predições em linhas ao invés de colunas (uma predição por linha, às vezes conhecida como formato "rolo de papel").

Transpondo Escores Tabulares

É possível transpor escores tabulares de colunas para linhas utilizando uma combinação de passos no IBM SPSS Modeler, conforme descrito nos passos a seguir.

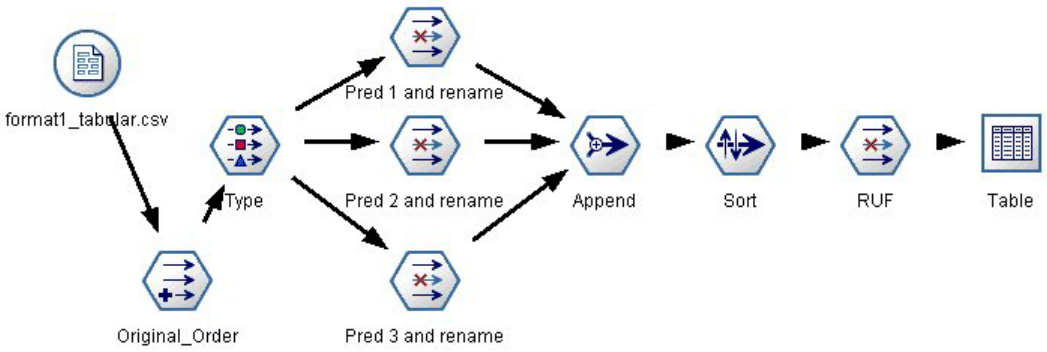


Figura 49. Fluxo de exemplo utilizado para transpor dados tabulares para o formato de rolo de papel

1. Utilize a função `@INDEX` em um nó Derivar para determinar a ordem atual das predições e salvar este indicador em um novo campo, como *Original_order*.
2. Inclua um nó Tipo para assegurar que todos os campos sejam instanciados.
3. Utilize um nó Filtro para renomear os campos de predição, de confiança e de ID (*P1, C1, I1*) padrão para campos comuns, como *Pred*, *Crit* e *Rule_ID*, que serão utilizados para anexar registros mais tarde. É necessário um nó filtro para cada predição gerada.

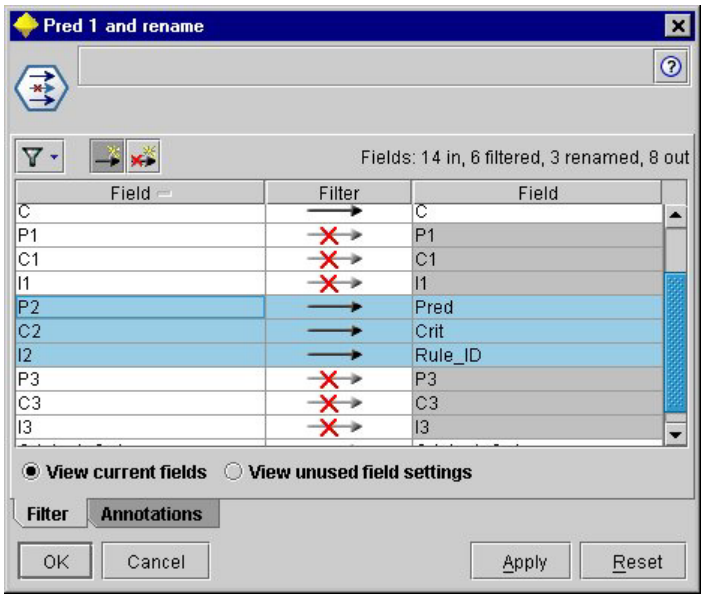


Figura 50. Filtrando campos para as predições 1 e 3 enquanto renomeia campos para a predição 2.

4. Utilize um nó Anexar para anexar valores para os campos *Pred*, *Crit*, e *Rule_ID* compartilhados.
5. Anexe um nó Ordenar para classificar os registros em ordem crescente para o campo *Original_order* e em ordem decrescente para *Crit*, que é o campo utilizado para ordenar as predições por critérios como confiança, elevação e suporte.
6. Utilize outro nó Filtro para filtrar o campo *Original_order* da saída.

Neste ponto, os dados estão prontos para implementação.

Transpondo Escores Transacionais

O processo é semelhante à transposição de escores transacionais. Por exemplo, o fluxo mostrado abaixo transpõe escores para um formato com uma predição única em cada linha, conforme necessário para implementação.

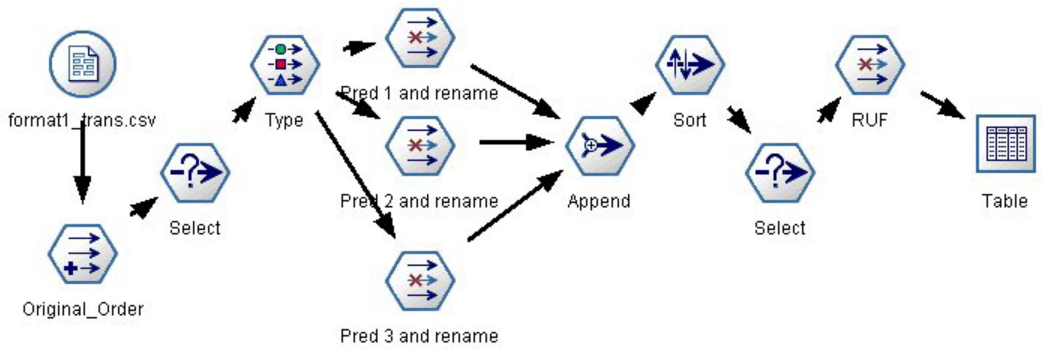


Figura 51. Fluxo de exemplo utilizado para transpor dados transacionais para o formato de rolo de papel

Com a inclusão de dois nós Seleção, o processo é idêntico àquele explicado anteriormente para dados tabulares.

- O primeiro nó Seleção é utilizado para comparar IDs de regras entre os registros adjacentes e incluir apenas registros exclusivos ou indefinidos. Esse nó Seleção utiliza a expressão do CLEM para selecionar registros: `ID /= @OFFSET(ID,-1) ou @OFFSET(ID,-1) = undef`.

- O segundo nó Seleção é utilizado para descartar regras estranhas, ou regras em que `Rule_ID` possui um valor nulo. Esse nó Seleção utiliza a seguinte expressão do CLEM para descartar registros: `not(@NULL(Rule_ID))`.

Para obter mais informações sobre como transpor escores para implementação, entre em contato com o Suporte Técnico.

Nó Sequência

O nó Sequência descobre padrões em dados sequenciais ou orientados por tempo, no formato `bread -> cheese`. Os elementos de uma sequência são **conjuntos de itens** que constituem uma transação única. Por exemplo, se uma pessoa vai a uma loja e compra pão e leite e alguns dias depois ela retorna para a loja e compra queijo, a atividade de compra dessa pessoa poderá ser representada como dois conjuntos de itens. O primeiro conjunto de itens contém pão e leite e o segundo contém queijo. Uma **sequência** é uma lista de conjuntos de itens que tendem a ocorrer em uma ordem previsível. O nó Sequência detecta as sequências frequentes e cria um nó de modelo gerado que pode ser usado para fazer predições.

Requisitos. Para criar um conjunto de regras de Sequência, é necessário especificar um campo de ID, um campo de tempo opcional e um ou mais campos de conteúdo. Observe que essas configurações devem ser feitas na guia Campos do nó de modelagem e não podem ser lidas a partir de um nó Tipo de envio de dados. O campo de ID pode ter qualquer nível de papel ou medição. Se você especificar um campo de tempo, ele poderá ter qualquer papel, mas seu armazenamento deverá ser numérico, de data, de hora ou de registro de data e hora. Se você não especificar um campo de tempo, o nó Sequência usará um registro de data e hora implícito, em vigor usando números de linhas como valores de tempo. Os campos de conteúdo podem ter qualquer nível de medição e papel, mas todos os campos de conteúdo devem ser do mesmo tipo. Se forem numéricos, eles deverão ser intervalos de números inteiros (não intervalos reais).

Intensidades. O nó Sequência baseia-se no algoritmo de regras de associação do CARMA que utiliza um método de duas passagens eficiente para localizar sequências. Além disso, o nó do modelo gerado criado por um nó Sequência pode ser inserido em um fluxo de dados para criar predições. O nó do modelo gerado também pode gerar SuperNodes para detectar e contar sequências específicas e fazer predições com base em sequências específicas.

Opções de Campos do Nó Sequência

Antes de executar um nó Sequência, deve-se especificar os campos de ID e de conteúdo na guia Campos do nó Sequência. Se desejar usar um campo de tempo, ele também deverá ser especificado aqui.

Campo de ID. Selecione um campo de ID na lista. Campos numéricos ou simbólicos podem ser utilizados como o campo de ID. Cada valor exclusivo deste campo deve indicar uma unidade específica de análise. Por exemplo, em um aplicativo de cesta de mercado, cada ID pode representar um cliente único. Para um aplicativo de análise de log da web, cada ID pode representar um computador (pelo endereço IP) ou um usuário (pelos dados de login).

- **IDs são contíguos.** Se seus dados estiverem pré-ordenados de modo que todos os registros com o mesmo ID sejam agrupados no fluxo de dados, selecione esta opção para acelerar o processamento. Se seus dados não estiverem pré-ordenados (ou se você não tiver certeza), deixe essa opção desmarcada e o nó Sequência ordenará os dados automaticamente.

Nota: se seus dados não estiverem ordenados e você selecionar essa opção, resultados inválidos poderão ser obtidos no seu modelo Sequência.

Campo de tempo. Se desejar utilizar um campo nos dados para indicar tempos de evento, selecione **Usar campo de tempo** e especifique o campo a ser utilizado. O campo de tempo deve ser numérico, de data, de hora ou de registro de data e hora. Se nenhum campo de tempo for especificado, supõe-se que os

registros chegarão da origem de dados em ordem sequencial, e os números de registro são utilizados como valores de tempo (o primeiro registro ocorre no tempo "1", o segundo no tempo "2", e assim por diante).

Campos de conteúdo. Especifique um ou mais campos de conteúdo para o modelo. Esses campos contêm os eventos de interesse na modelagem de sequência.

O nó Sequência pode manipular dados em formato tabular ou transacional. Se você utilizar diversos campos com dados transacionais, os itens especificados nestes campos para um registro específico serão considerados para representar os itens localizados em uma única transação com um único registro de data e hora. Consulte o tópico "Dados Tabulares versus Transacionais" na página 246 para obter mais informações.

Partição. Este campo permite especificar um campo utilizado para particionar os dados em amostras separadas para os estágios de treinamento, de teste e de validação de construção de modelo. Ao utilizar uma amostra para gerar o modelo e uma amostra diferente para testá-lo, é possível obter uma boa indicação do quão bem o modelo será generalizado para conjuntos de dados maiores que forem semelhantes aos dados atuais. Se diversos campos de partição tiverem sido definidos usando os nós Tipo ou Partição, um campo de partição único deverá ser selecionado na guia Campos em cada nó de modelagem que utiliza particionamento. (Se apenas uma partição estiver presente, ela será utilizada automaticamente sempre que o particionamento estiver ativado). Além disso, observe que para aplicar a partição selecionada à sua análise, o particionamento também deverá ser ativado na guia Opções de Modelo para o nó. (Desmarcar esta opção permite desativar o particionamento sem alterar as configurações do campo).

Opções do Modelo do Nó Sequência

Nome do modelo É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Utilizar dados particionados. Se um campo de partição for definido, essa opção assegurará que apenas os dados da partição de treinamento sejam utilizados para construir o modelo.

Suporte mínimo de regra (%) É possível especificar um critério de suporte. *Suporte de regra* refere-se à proporção de IDs nos dados de treinamento que contêm a sequência inteira. Se desejar focar em sequências mais comuns, aumente essa configuração.

Confiança mínima de regra (%) É possível especificar um critério de confiança para manter as sequências no conjunto de sequências. *Confiança* refere-se à porcentagem de IDs na qual uma predição correta é feita, dentre todos os IDs para os quais a regra faz uma predição. Ela é calculada como o número de IDs para os quais a sequência inteira é localizada dividido pelo número de IDs para os quais os antecedentes são localizados, com base nos dados de treinamento. As sequências com confiança menor que o critério especificado são descartadas. Se você estiver obtendo muitas sequências ou sequências desinteressantes, tente aumentar essa configuração. Se estiver obtendo pouquíssimas sequências, tente diminuir esta configuração.

Nota: Se necessário, é possível destacar o valor e digitar seu próprio valor. Lembre-se de que se reduzir o valor de confiança abaixo de 1,0, além do processo que requer muita memória livre, você poderá achar que as regras estão demorando um tempo extremamente longo para construir.

Tamanho máximo de sequência É possível configurar o número máximo de itens distintos em uma sequência. Se as sequências de interesse forem relativamente curtas, será possível diminuir esta configuração para acelerar a construção do conjunto de sequências.

Predições para incluir no fluxo Especifique o número de predições a serem incluídas no fluxo pelo nó Modelo gerado resultante. Para obter informações adicionais, consulte "Nuggets do Modelo de Sequência" na página 265.

Opções Avançadas do Nó Sequência

Para aqueles que possuem um conhecimento detalhado da operação do nó Sequência, as opções avançadas a seguir permitem fazer um ajuste preciso do processo de construção de modelo. Para acessar as opções avançadas, configure o Modo para **Especialista** na guia Especialista.

Configurar duração máxima. Se essa opção for selecionada, as sequências serão limitadas àqueles com uma duração (o tempo entre a configuração do primeiro e do último item) menor ou igual ao valor especificado. Se você não tiver especificado um campo de tempo, a duração será expressa em termos de linhas (registros) nos dados brutos. Se o campo de tempo utilizado for um campo de hora, data ou de registro de data e hora, a duração será expressa em segundos. Para campos numéricos, a duração é expressa nas mesmas unidades que o próprio campo.

Configurar o valor de poda. O algoritmo CARMA utilizado no nó Sequência remove (**poda**) periodicamente conjuntos de itens a partir de sua lista de possíveis conjuntos de itens durante o processamento para conservar memória. Selecione esta opção para ajustar a frequência da poda. O número especificado determina a frequência da poda. Insira um valor menor para diminuir os requisitos de memória do algoritmo (e potencialmente aumentar o tempo de treinamento necessário), ou insira um valor maior para acelerar o treinamento (e potencialmente aumentar os requisitos de memória).

Configurar o máximo de sequências na memória. Se esta opção for selecionada, o algoritmo do CARMA limitará seu armazenamento de memória de sequências de candidatos durante a construção de modelo para o número de sequências especificado. Selecione esta opção se o IBM SPSS Modeler estiver utilizando memória excessiva durante a construção de modelos de Sequência. Observe que o valor máximo de sequências que você especificar aqui será o número de sequências candidatas rastreadas internamente conforme o modelo é construído. Esse número deve ser muito maior que o número de sequências que você espera no modelo final.

Restringir diferenças entre conjuntos de itens. Esta opção permite especificar restrições nos intervalos de tempo que separam conjuntos de itens. Se selecionada, os conjuntos de itens com diferenças de tempo menores que a **Diferença mínima** ou maiores que a **Diferença máxima** que você especificar não serão considerados para fazer parte de uma sequência. Utilize esta opção para evitar continuar sequências que incluam intervalos de tempo longos ou aqueles que ocorrem em um período de tempo muito curto.

Nota: se o campo de tempo utilizado for uma hora, data ou registro de data e hora, o intervalo de tempo será expresso em segundos. Para campos numéricos, a diferença de tempo é expressa nas mesmas unidades que o campo de tempo.

Por exemplo, considere a lista de transações a seguir.

Tabela 18. Exemplo de lista de transações.

ID	Hora	Conteúdo
1001	1	apples
1001	2	bread
1001	5	cheese
1001	6	dressing

Se construir um modelo desses dados com a diferença mínima configurada para 2, você deverá obter as seguintes sequências:

```
apples -> cheese

apples -> dressing

bread -> cheese

bread -> dressing
```

As sequências como `apples -> bread` não deverão ser exibidas porque a diferença entre `apples` e `bread` é menor que a diferença mínima. De modo semelhante, considere os seguintes dados alternativos.

Tabela 19. Exemplo de lista de transações.

ID	Hora	Conteúdo
1001	1	apples
1001	2	bread
1001	5	cheese
1001	20	dressing

Se a diferença máxima for configurada como 10, não deverão ser exibidas sequências com `dressing` porque a diferença entre `cheese` e `dressing` é muito grande para que eles sejam considerados parte da mesma sequência.

Nuggets do Modelo de Sequência

Os nuggets do modelo de sequência representam as sequências localizadas para um campo de saída específico descoberto pelo nó Sequência e podem ser incluídos nos fluxos para gerar predições.

Ao executar um fluxo contendo um nó Sequência, o nó inclui um par de campos contendo predições e valores de confiança associados a cada predição no modelo de sequência nos dados. Por padrão, três pares de campos contendo as três principais predições (e seus valores de confiança associados) são incluídos. É possível alterar o número de predições geradas ao construir o modelo configurando as opções de modelo do nó Sequência no momento da construção e também na guia Configurações após incluir o nugget do modelo em um fluxo. Consulte o tópico "Configurações do Nugget do Modelo de Sequência" na página 268 para obter mais informações.

Os novos nomes de campo são derivados do nome do modelo. Os nomes de campo são *$S-sequence-n* para o campo de predição (em que *n* indica a *n*ésima predição) e *$SC-sequence-n* para o campo de confiança. Em um fluxo com diversos nós Regras de Sequência em uma série, os novos nomes de campo incluirão os números no prefixo para diferenciá-los uns dos outros. O primeiro nó Conjunto de Sequências no fluxo utilizará os nomes comuns, o segundo nó utilizará nomes que iniciam com *$S1-* e *$SC1-*, o terceiro nó utilizará nomes que iniciam com *$S2-* e *$SC2-*, e assim por diante. As predições são exibidas em ordem de confiança, de modo que *$S-sequence-1* contém a predição com a confiança mais alta, *$S-sequence-2* contém a predição com a próxima confiança mais alta, e assim por diante. Para registros nos quais o número de predições disponíveis é menor que o número de predições solicitadas, as predições restantes contêm o valor `$null$`. Por exemplo, se apenas duas predições puderem ser feitas para um registro específico, os valores de *$S-sequence-3* e *$SC-sequence-3* serão `$null$`.

Para cada registro, as regras no modelo são comparadas com o conjunto de transações processadas para o ID atual até o momento, incluindo o registro atual e quaisquer registros anteriores com o mesmo ID e registro e data e hora anterior. As *k* regras com os valores mais altos de confiança que se aplicam a este conjunto de transações são utilizadas para gerar as *k* predições para o registro, em que *k* é o número de predições especificadas na guia Configurações após incluir o modelo no fluxo. (Se diversas regras

preverem o mesmo resultado para o conjunto de transações, apenas a regra com a confiança mais alta será utilizada). Consulte o tópico "Configurações do Nugget do Modelo de Sequência" na página 268 para obter mais informações.

Assim como acontece com outros tipos de modelos de regras de associação, o formato de dados deverá corresponder ao formato usado na construção do modelo de sequência. Por exemplo, os modelos construídos usando dados tabulares podem ser usados para escorar somente dados tabulares. Consulte o tópico "Escorando Regras de Associação" na página 258 para obter mais informações.

Nota: ao escorar dados utilizando um nó Conjunto de Sequências gerado em um fluxo, quaisquer configurações de tolerância ou de diferença que forem selecionadas na construção do modelo serão ignoradas para propósitos de escoragem.

Predições a partir de Regras de Sequência

O nó manipula os registros em uma maneira dependente de tempo (ou dependente de ordem, se nenhum campo de registro data e hora foi utilizado para construir o modelo). Os registros devem ser ordenados pelo campo de ID e pelo campo de registro de data e hora (se presente). No entanto, as predições não estão empatadas ao registro de data e hora do registro ao qual elas são incluídas. Elas simplesmente fazem referência a itens que mais provavelmente irão ocorrer *em algum ponto no futuro*, dado o histórico das transações para o ID atual até o registro atual.

Observe que as predições para cada registro não dependem necessariamente das transações desse registro. Se as transações do registro atual não acionarem uma regra específica, as regras serão selecionadas com base nas transações anteriores para o ID atual. Em outras palavras, se o registro atual não incluir nenhuma informação preditiva útil na sequência, a predição da última transação útil para esse ID será conduzida para o registro atual.

Por exemplo, suponha que você tenha um modelo Sequência com a única regra

```
Jam -> Bread (0.66)
```

e você transmite os registros a seguir para ele.

Tabela 20. Exemplo de registros.

ID	Compra	Predição
001	jam	bread
001	milk	bread

Observe que o primeiro registro gera uma predição de *bread*, como se espera. O segundo registro também contém uma predição de *bread*, porque não há nenhuma regra para *jam* seguida por *milk*, portanto, a transação *milk* não inclui nenhuma informação útil e a regra `Jam -> Bread` ainda se aplica.

Gerar Novos Nós

O menu Gerar permite criar novos SuperNodes com base no modelo de sequência.

- **SuperNode de Regra.** Cria um SuperNode que pode detectar e contar ocorrências de sequências de dados escorados. Esta opção estará desativada se nenhuma regra estiver selecionada. Consulte o tópico "Gerando um SuperNode de Regra a partir de um Nugget do Modelo de Sequência" na página 268 para obter mais informações.
- **Modelo para Paleta.** Retorna o modelo para a paleta Modelos. Isso é útil em situações em que um colega pode ter enviado um fluxo contendo o modelo e não o próprio modelo.

Detalhes do Nugget do Modelo de Sequência

A guia Modelo para um nugget do modelo de Sequência exibe as regras extraídas pelo algoritmo. Cada linha na tabela representa uma regra, com o antecedente (a parte "if" da regra) na primeira coluna seguido pelo subsequente (a parte "then" da regra) na segunda coluna.

Cada regra é mostrada no formato a seguir.

Tabela 21. Formato de Regra

Antecedente	Subsequente
beer e cannedveg	beer
fish fish	fish

A primeira regra de exemplo é interpretada como *para IDs que tinham "beer" e "cannedveg" na mesma transação, provavelmente haverá uma ocorrência subsequente de "beer"*. A segunda regra de exemplo pode ser interpretada como *para IDs que tinham "fish" em uma transação e, em seguida, "fish" em outra, há uma probabilidade de ocorrência subsequente de "fish"*. Observe que na primeira regra, *beer* e *cannedveg* são comprados ao mesmo tempo e, na segunda regra, *fish* é comprado em duas transações separadas.

Menu Ordenar. O botão do menu Ordenar na barra de ferramentas controla a ordenação das regras. A direção da ordenação (crescente ou decrescente) pode ser alterada utilizando o botão de direção de ordenação (seta para cima ou para baixo).

As regras podem ser ordenadas por:

- % de Suporte
- % de Confiança
- % de Suporte de regra
- Subsequente
- Primeiro Antecedente
- Último Antecedente
- Número de Itens (antecedentes)

Por exemplo, a tabela a seguir é classificada em ordem decrescente por número de itens. As regras com diversos itens no conjunto antecedente precedem àquelas com menos itens.

Tabela 22. Regras classificada por número de itens

Antecedente	Subsequente
beer e cannedveg e frozenmeal	frozenmeal
beer e cannedveg	beer
fish fish	fish
softdrink	softdrink

Mostrar/ocultar menu de critérios. O botão Mostrar/ocultar menu de critérios (ícone de grade) controla as opções para a exibição de regras. As opções de exibição a seguir estão disponíveis:

- **Instâncias** exibe informações sobre o número de IDs exclusivos para os quais a *sequência completa* – ambos antecedentes e subsequentes - ocorre. (Observe que isso difere dos modelos de Associação, em que o número de instâncias refere-se ao número de IDs para os quais *apenas* os antecedentes se aplicam). Por exemplo, dada a regra `bread -> cheese`, o número de IDs nos dados de treinamento que incluem ambos *bread* e *cheese* é referido como **instâncias**.

- **Suporte** exibe a proporção de IDs nos dados de treinamento para os quais os antecedentes são verdadeiros. Por exemplo, se 50% dos dados de treinamento incluírem o antecedente *bread*, o suporte para a regra `bread -> cheese` será de 50%. (Ao contrário dos modelos de Associação, o suporte *não* baseia-se no número de instâncias, conforme observado anteriormente).
- **Confiança** exibe a porcentagem de IDs na qual uma predição correta é feita, dentre todos os IDs para os quais a regra faz uma predição. Ela é calculada como o número de IDs para os quais a sequência inteira é localizada dividido pelo número de IDs para os quais os antecedentes são localizados, com base nos dados de treinamento. Por exemplo, se 50% dos dados de treinamento contiverem `cannedveg` (indicando suporte de antecedente), mas apenas 20% contiverem ambos `cannedveg` e `frozenmeal`, então a confiança para a regra `cannedveg -> frozenmeal` será de `Rule Support / Antecedent Support`, ou seja, 40%.
- O **Suporte de Regra** para modelos de Sequência baseia-se em instâncias e exibe a proporção de registros de treinamento para os quais a regra inteira, os antecedentes e um ou mais subsequentes são verdadeiros. Por exemplo, se 20% dos dados de treinamento contiverem ambos *bread* e *cheese*, então o suporte de regra para a regra `bread -> cheese` será de 20%.

Observe que as proporções baseiam-se em transações válidas (transações com pelo menos um item ou valor real observado) e não no total de transações. Transações inválidas – sem itens ou valores reais - são descartadas para estes cálculos.

Botão Filtrar. O botão Filtrar (ícone de funil) no menu expande a parte inferior da caixa de diálogo para mostrar um painel no qual os filtros de regra ativos são exibidos. Os filtros são utilizados para limitar o número de regras exibidas na guia Modelos.

Figura 52. Botão Filtrar

Para criar um filtro, clique no ícone Filtro à direita do painel expandido. Isso abre uma caixa de diálogo separada na qual é possível especificar restrições para exibir regras. Observe que o botão Filtrar é geralmente utilizado em conjunto com o menu Gerar para primeiro filtrar regras e, em seguida, gerar um modelo contendo esse subconjunto de regras. Para obter mais informações, consulte “Especificando Filtros para Regras” na página 255 abaixo.

Configurações do Nugget do Modelo de Sequência

A guia Configurações de um nugget do modelo Sequência exibe opções de escoragem para o modelo. Esta guia estará disponível somente após o modelo ter sido incluído nas telas de fluxo para escoragem.

Número máximo de predições. Especifique o número máximo de predições que são incluídas para cada conjunto de itens de cesta. As regras com os valores de confiança mais altos que se aplicam a este conjunto de transações são utilizadas para gerar predições para o registro até o limite especificado.

Sumarização do Nugget do Modelo de Sequência

A guia Sumarização de um nugget do modelo de regra de sequência exibe o número de regras descobertas e o mínimo e o máximo de suporte e confiança nas regras. Se você tiver executado um nó Análise anexado a este nó de modelagem, as informações dessa análise também serão exibidas nesta seção.

Consulte o tópico “Procurando Nuggets do Modelo” na página 42 para obter mais informações.

Gerando um SuperNode de Regra a partir de um Nugget do Modelo de Sequência

Para gerar um SuperNode de regra com base em uma regra de sequência:

1. Na guia Modelo do nugget do modelo de regra de sequência, clique em uma linha na tabela para selecionar a regra desejada.
2. Nos menus do navegador de regras, escolha:

   Gerar > **SuperNode de Regra**

Importante: Para utilizar o SuperNode gerado, deve-se ordenar os dados por campo de ID (e por campo de Tempo, se houver), antes de transmiti-los para o SuperNode. O SuperNode não detectará adequadamente as sequências em dados não ordenados.

É possível especificar as opções a seguir para gerar um SuperNode de regra:

Detectar. Especifica como as correspondências são definidas para dados transmitidos para o SuperNode.

- **Antecedentes apenas.** O SuperNode identificará uma correspondência a qualquer momento que ele localizar os antecedentes para a regra selecionada na ordem correta dentro de um conjunto de registros que tiverem o mesmo ID, independentemente se o subsequente também for localizado. Observe que isso não leva em consideração as configurações de restrição de tolerância de registro de data e hora ou de diferença de item a partir do nó de modelagem Sequência original. Quando o último conjunto de itens antecedentes for detectado no fluxo (e todos os outros antecedentes tiverem sido localizados na ordem apropriada), todos os registros subsequentes com o ID atual conterão a sumarização selecionada abaixo.
- **Sequência inteira.** O SuperNode identificará uma correspondência a qualquer momento que ele localizar os antecedentes e o subsequente para a regra selecionada na ordem correta dentro de um conjunto de registros que tiverem o mesmo ID. Isso não leva em consideração as configurações de restrição de tolerância de registro de data e hora ou de diferença de item a partir do nó de modelagem Sequência original. Quando o subsequente for detectado no fluxo (e todos os antecedentes também tiverem sido localizados na ordem correta), o registro atual e todos os registros subsequentes com o ID atual conterão a sumarização selecionada abaixo.

Exibição. Controla como as sumarizações de correspondência são incluídas nos dados na saída do SuperNode de Regra.

- **Valor subsequente para a primeira ocorrência.** O valor incluído nos dados é o valor subsequente predito com base na primeira ocorrência da correspondência. Os valores são incluídos como um novo campo denominado *rule_n_consequent*, em que *n* é o número da regra (com base na ordem de criação de SuperNodes de Regra no fluxo).
- **Valor real para a primeira ocorrência.** O valor incluído nos dados será verdadeiro se houver pelo menos uma correspondência para o ID e falso se não houver nenhuma correspondência. Os valores são incluídos como um novo campo denominado *rule_n_flag*.
- **Contagem de ocorrências.** O valor incluído nos dados é o número de correspondências para o ID. Os valores são incluídos como um novo campo denominado *rule_n_count*.
- **Número de regra.** O valor incluído é o número da regra para a regra selecionada. Os **Números de regra** são designados com base na ordem na qual o SuperNode foi incluído no fluxo. Por exemplo, o primeiro SuperNode de Regra é considerado *regra 1*, o segundo SuperNode de Regra é considerado *regra 2*, e assim por diante. Essa opção é mais útil quando estiver incluindo diversos SuperNodes de Regra em seu fluxo. Os valores são incluídos como um novo campo denominado *rule_n_number*.
- **Incluir figuras de confiança.** Se selecionada, esta opção incluirá a confiança da regra no fluxo de dados, bem como a sumarização selecionada. Os valores são incluídos como um novo campo denominado *rule_n_confidence*.

Nó Regras de Associação

As regras de associação são instruções no formato a seguir.

Por exemplo, "Se um cliente comprar uma lâmina de barbear e uma loção pós-banho, então esse cliente comprará um creme de barbear com 80% de confiança". O nó Regras de Associação extrai um conjunto de regras a partir dos dados, retirando as regras com o maior conteúdo de informações. O nó de Regras de Associação é muito semelhante ao nó a priori, no entanto, há algumas diferenças importantes:

- O nó Regras de Associação não pode processar dados transacionais.
- O nó Regras de Associação pode processar dados que tiverem o tipo de armazenamento Lista e o nível de medição Coleção.
- O nó Regras de Associação poderá ser utilizado com o IBM SPSS Analytic Server. Isso fornece escalabilidade e permite processar Big Data e aproveitar processamento paralelo mais rápido.
- O nó Regras de Associação fornece configurações adicionais, como a capacidade de restringir o número de regras que são geradas, aumentando, assim, a velocidade de processamento.
- A saída do nugget do modelo é mostrada no Visualizador de Saída.

Nota: O nó Regras de Associação não suporta os passos Avaliação de Modelo ou Desafiante Campeão no IBM SPSS Collaboration and Deployment Services.

Nota: O nó Regras de Associação ignorará registros vazios quando construir um modelo se o tipo de campo for flag. Registros vazios são registros em que todos os campos utilizados na construção de modelo possuem um valor false.

Um fluxo que mostra um exemplo de trabalho utilizando Regras de Associação, denominado `geospatial_association.str`, e que faz referência aos arquivos de dados `InsuranceData.sav`, `CountyData.sav` e `ChicagoAreaCounties.shp` está disponível a partir do diretório `Demos` da sua instalação do IBM SPSS Modeler. É possível acessar o diretório `Demos` do grupo de programa IBM SPSS Modeler no menu Iniciar do Windows. O arquivo `geospatial_association.str` está no diretório `streams`.

Regras de Associação - Opções de Campo

Na guia **Campos**, escolha se deseja usar as configurações de papel de campo que já estiverem definidas nos nós de envio de dados, como um nó Tipo de envio de dados, ou fazer as designações de campo manualmente.

Usar papéis predefinidos
: Esta opção utiliza as configurações de papel (como destinos, ou preditores) a partir de um nó Tipo de envio de dados (ou na guia Tipos de um nó de origem de envio de dados). Os campos com um papel de entrada são considerados como Condições, os campos com um papel de destino são considerados como Predições e os campos que são utilizados como entradas e destinos são considerados como tendo ambos os papéis.

Usar designações de campo customizadas
: Escolha esta opção se desejar designar destinos, preditores e outros papéis manualmente nessa tela.

Campos
: Se você selecionou **Usar designações de campo customizadas**, utilize os botões de seta para designar itens manualmente a partir desta lista para as caixas à direita da tela. Os ícones indicam os níveis de medição válidos para cada campo.

Ambos (condição ou predição)
: Os campos incluídos nessa lista podem assumir o papel de condição ou de predição nas regras que forem geradas pelo modelo. Isso ocorre basicamente regra por regra, portanto, um campo poderá ser uma condição em uma regra e uma predição em outra regra.

Apenas predição
: Os campos incluídos nessa lista podem aparecer apenas como uma predição (também conhecidos como "subsequentes") de uma regra. A presença de um campo nessa lista não significa que o campo é utilizado em qualquer regra, apenas que, se for utilizado, poderá ser apenas uma predição.

Apenas condição
Os campos incluídos nessa lista podem aparecer apenas como uma condição (também conhecidos como "antecedentes") de uma regra. A presença de um campo nessa lista não significa que o campo é utilizado em qualquer regra, apenas que, se for utilizado, poderá ser apenas uma condição.

Regras de Associação - Construção de Regra

Itens por regra

Utilize essas opções para especificar quantos itens, ou valores, podem ser utilizados em cada regra.

Nota: O total combinado desses dois campos não pode exceder 10.

Condições máximas
Selecione o número máximo de condições que podem ser incluídas em uma regra única.

Predições máximas
Selecione o número máximo de predições que podem ser incluídas em uma regra única.

Construção de regra

Utilize essas opções para especificar o número e o tipo de regras para construir.

Número máximo de regras
Especifique o número máximo de regras que podem ser consideradas para uso na construção de regras para seu modelo.

Critério de regra para N principais
Selecione o critério que é utilizado para estabelecer quais são as N principais regras, em que N é o valor que é inserido no campo **Número máximo de regras**. É possível escolher entre os critérios a seguir:

- **Confiança**
- **Suporte de Regra**
- **Suporte de condição**
- **Elevar**
- **Implementabilidade**

Somente valores reais para flags
Quando seus dados estiverem em formato tabular, selecione esta opção para incluir apenas valores reais para os campos de flag nas regras resultantes. Selecionar valores reais pode ajudar a tornar as regras mais fáceis de entender. A opção não se aplica a dados em formato transacional. Para obter informações adicionais, consulte “Dados Tabulares versus Transacionais” na página 246.

Critério de regra

Se você selecionar **Ativar critério de regra**, será possível utilizar essas opções para selecionar a intensidade mínima que as regras devem atender para serem consideradas para uso em seu modelo.

- **Confiança** Especifique o valor mínimo de porcentagem para o nível de Confiança para uma regra que é produzida pelo modelo. Se o modelo produzir uma regra com um nível menor que essa quantia, a regra será descartada.
- **Suporte de Regra** Especifique o valor mínimo de porcentagem para o nível de Suporte de Regra para uma regra que é produzida pelo modelo. Se o modelo produzir uma regra com um nível menor que essa quantia, a regra será descartada.

- **Suporte de Condição** Especifique o valor mínimo de porcentagem para o nível de Suporte de Condição para uma regra que é produzida pelo modelo. Se o modelo produzir uma regra com um nível menor que a quantia especificada, a regra será descartada.
- **Elevação** Especifique o valor de Elevação mínimo permitido para uma regra que é produzida pelo modelo. Se o modelo produzir uma regra com um valor menor que a quantia especificada, a regra será descartada.

Excluir regras

Em alguns casos, a associação entre dois ou mais campos é conhecida ou é autoevidente, em que, nesse caso, é possível excluir regras nas quais os campos preveem uns aos outros. Ao excluir regras que contenham ambos os valores, você reduz entrada irrelevante e aumenta as chances de localizar resultados úteis.

Campos
Selecione os campos associados que você não desejar utilizar juntos na construção da regra. Por exemplo, os campos associados podem ser Fabricantes de Carro e Modelo de Carro, ou Ano Letivo e Idade do Aluno. Quando o modelo cria as regras, se a regra contiver pelo menos um dos campos selecionados em um dos lados da regra (condição ou predição), a regra será descartada.

Regras de Associação - Transformações

Categorização

Utilize estas opções para especificar como os campos contínuos (intervalo numérico) são categorizados.

Número de categorias
Quaisquer campos contínuos configurados para serem categorizados automaticamente são divididos pelo número de categorias igualmente espaçadas que você especificar. É possível selecionar qualquer número no intervalo de 2 a 10.

Campos de lista

Comprimento máximo da lista
Para restringir o número de itens a serem incluídos no modelo se o comprimento de um campo de lista for desconhecido, insira o comprimento máximo da lista. É possível selecionar qualquer número no intervalo de 1 a 100. Se uma lista for maior que o número inserido, o modelo ainda utilizará o campo, mas incluirá valores somente até este número, e quaisquer valores extras no campo são ignorados.

Regras de Associação - Saída

Utilize as opções nessa área de janela para controlar qual saída é gerada quando o modelo é construído.

Tabelas de regras

Utilize estas opções para criar um ou mais tipos de tabelas que exibem o melhor número de regras (com base em um número que você especificar) para cada critério selecionado.

Confiança
A confiança é a razão do suporte de regra com o suporte de condição. Dentre os itens com os valores da condição listados, a porcentagem que possui os valores subsequentes preditos. Cria uma tabela que contém as N melhores regras de associação que se baseiam em confiança a serem incluídas na saída (em que N é o valor de **Regras para exibir**).

Suporte de Regra
A proporção de itens para os quais a regra, as condições e predições inteiras são verdadeiras. Para todos os itens no conjunto de dados, a porcentagem que é considerada e predita corretamente pela regra. Essa medida dá uma importância geral da regra. Cria uma tabela que

contém as N melhores regras de associação que se baseiam em suporte de regra a serem incluídas na saída (em que N é o valor de **Regras para exibir**).

Elevar A razão da confiança de regra e a probabilidade anterior de ter a predição. A razão do valor de Confiança de uma regra versus a porcentagem de valores de Subsequentes que ocorrem na população geral. Essa razão fornece uma medição do quão bem a regra melhora sobre a chance. Cria uma tabela que contém as N melhores regras de associação que se baseiam na elevação a serem incluídas na saída (em que N é o valor de **Regras para exibir**).

Suporte de condição
A proporção de itens para os quais as condições forem verdadeiras. Cria uma tabela que contém as N melhores regras de associação que se baseiam em suporte de antecedente a serem incluídas na saída (em que N é o valor de **Regras para exibir**).

Implementabilidade
Uma medida de qual porcentagem dos dados de treinamento satisfaz a condição, mas não a predição. Esta medida mostra a frequência de perda da regra. Ela é efetivamente o oposto de Confiança. Cria uma tabela que contém as N melhores regras de associação que se baseiam em implementabilidade a serem incluídas na saída (N é o valor de **Regras para exibir**).

Regras para exibir
Configure o número máximo de regras a serem exibidas nas tabelas.

Tabelas de informações de modelo

Utilize uma ou mais dessas opções para selecionar quais tabelas modelo incluir na saída.

- **Transformações de Campo**
- **Sumarização de registros**
- **Estatísticas de Regra**
- **Valores Mais Frequentes**
- **Campos Mais Frequentes**

Nuvem de palavras ordenável de regras.

Utilize essas opções para criar uma nuvem de palavras que exibe as saídas de regras. As palavras são exibidas em tamanhos de texto maiores para indicar sua importância.

Criar uma nuvem de palavras ordenável.
Selecione esta caixa para criar uma nuvem palavras ordenável em sua saída.

Ordem padrão
Selecione o tipo de ordenação a ser utilizado quando criar inicialmente a nuvem de palavras. A nuvem de palavras é interativa e é possível alterar o critério no Visualizador do Modelo para ver diferentes regras e ordenações. É possível escolher entre as opções de ordenação a seguir:

- Confiança.
- Suporte de Regra
- Elevar
- Suporte de Condição.
- Implementabilidade

Máximo de regras a serem exibidas
Configure o número de regras a serem exibidas na nuvem de palavras; é possível escolher no máximo 20.

Regras de Associação - Opções de Modelo

Utilize as configurações nesta guia para especificar as opções de escoragem para os modelos de Regras de Associação.

Nome do modelo É possível gerar o nome do modelo que se baseia automaticamente no campo de destino (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Número máximo de predições Especifique o número máximo de predições que são incluídas no resultado da escoragem. Esta opção é utilizada com as entradas **Critério de Regra** para produzir as "principais" predições, em que "principais" indica o nível mais alto de confiança, de suporte, de elevação, e assim por diante.

Critério de Regra Selecione a medida que é usada para determinar a intensidade das regras. As regras são ordenadas pela força dos critérios que são selecionados aqui para retornar as principais predições de um conjunto de itens. É possível escolher a partir de 5 critérios diferentes.

- **Confiança** A confiança é a razão do suporte de regra para o suporte de condição. Dentre os itens com os valores da condição listados, a porcentagem que possui os valores subsequentes preditos.
- **Suporte de Condição** A razão de itens para os quais as condições forem verdadeiras.
- **Suporte de Regra** A razão de itens para os quais a regra inteira, as condições e as predições forem verdadeiras. Ela é calculada multiplicando o valor de **Suporte de Condição** pelo valor de **Confiança**.
- **Elevação** A razão da confiança de regra e a probabilidade anterior de ter a predição.
- **Implementabilidade** Uma medida de qual porcentagem dos dados de treinamento satisfaz a condição, mas não a predição.

Permitir predições repetidas Para incluir diversas regras com a mesma predição durante a escoragem, marque essa caixa de opção. Por exemplo, selecionar isso permite que as regras a seguir sejam escoradas.

```
bread & cheese -> wine
cheese & fruit -> wine
```

Nota: Regras com diversas predições (`bread & cheese & fruit -> wine & pate`) serão consideradas predições repetidas apenas se todas as predições (`wine & pate`) tiverem sido preditas antes.

Escorar regras apenas quando as predições não estiverem presentes na entrada Para assegurar que as predições também não estejam presentes na entrada, selecione esta opção. Por exemplo, se o propósito da escoragem é fazer uma recomendação de móveis domésticos, então é improvável que uma entrada que já contenha uma mesa de jantar compre outra mesa. Nesse caso, selecione esta opção. Por outro lado, se os produtos forem perecíveis ou descartáveis (como queijo, alimento para bebês ou papel higiênico), então as regras em que o subsequente já estiver presente na entrada poderão ser úteis. Nesse último caso, a opção mais útil pode ser **Escorar todas as regras**.

Escorar regras apenas quando as predições estiverem presentes na entrada Para assegurar que as predições também estejam presentes na entrada, selecione esta opção. Essa abordagem é útil quando você estiver tentando ganhar insight sobre clientes ou transações existentes. Por exemplo, você pode querer identificar regras com a elevação mais alta e, em seguida, explorar quais clientes se enquadram nessas regras.

Escorar todas as regras Para incluir todas as regras durante a escoragem, independentemente da presença ou da ausência das predições, selecione essa opção.

Nuggets do Modelo de Regras de Associação

O nugget do modelo contém informações sobre as regras extraídas de seus dados durante a construção de modelo.

Visualizando Resultados

É possível procurar as regras geradas pelos modelos de Regras de Associação utilizando a guia Modelo na caixa de diálogo. Procurar um nugget do modelo mostra as informações sobre as regras antes de gerar novos nós ou escorar o modelo.

Escoragem do Modelo

Os nuggets do modelo refinados podem ser incluídos em um fluxo e usados para escoragem. Consulte o tópico “Utilizando Nuggets do Modelo em Fluxos” na página 49 para obter mais informações. Os nuggets do modelo utilizados para escoragem incluem uma guia Configurações extra em suas respectivas caixas de diálogo. Consulte o tópico “Configurações do Nugget de Modelo de Regras de Associação” para obter mais informações.

Detalhes do Nugget de Modelo de Regras de Associação

O nugget do modelo Regras de Associação exibe detalhes do modelo na guia Modelo do Visualizador de Saída. Para obter mais informações sobre como utilizar o visualizador, consulte a seção com o título "Trabalhando com a Saída" no Guia do Usuário do Modelador (ModelerUsersGuide.pdf).

Uma operação de modelagem GSAR cria diversos novos campos com o prefixo $A, conforme mostrado na tabela a seguir.

Tabela 23. Novos campos criados pela operação de modelagem Regras de Associação

Nome do campo	Descrição
$A-<prediction>#	Este campo contém a predição do modelo para os registros escorados. O <prediction> é o nome do campo incluído no papel Predições no modelo e # é uma sequência de números para as regras de saída (por exemplo, se o escore for configurado para incluir 3 regras, a sequência de números será de 1 a 3).
$AC-<prediction>#	Este campo contém a confiança na predição. O <prediction> é o nome do campo incluído no papel Predições no modelo e # é uma sequência de números para as regras de saída (por exemplo, se o escore for configurado para incluir 3 regras, a sequência de números será de 1 a 3).
$A-Rule_ID#	Esta coluna contém o ID da regra predita para cada registro no conjunto de dados escorados. O # é uma sequência de números para as regras de saída (por exemplo, se o escore for configurado para incluir 3 regras, a sequência de números será de 1 a 3).

Configurações do Nugget de Modelo de Regras de Associação

A guia Configurações de um nugget do modelo Regras de Associação exibe opções de escoragem para o modelo. Esta guia estará disponível somente após o modelo ter sido incluído nas telas de fluxo para escoragem.

Número máximo de predições Especifique o número máximo de predições que são incluídas para cada conjunto de itens. As regras com os valores de confiança mais altos que se aplicam a este conjunto de transações são utilizadas para gerar predições para o registro até o limite especificado. Utilize esta opção com a opção **Critério de Regra** para produzir as “principais” predições, em que *principais* indica o nível mais alto de confiança, suporte, elevação, e assim por diante.

Critério de Regra Selecione a medida que é usada para determinar a intensidade das regras. As regras são ordenadas pela força dos critérios que são selecionados aqui para retornar as principais predições de um conjunto de itens. É possível escolher entre os critérios a seguir:

- **Confiança**
- **Suporte de Regra**
- **Elevar**
- **Suporte de condição**
- **Implementabilidade**

Permitir predições repetidas Marque essa caixa de seleção para incluir diversas regras com o mesmo subsequente durante a escoragem. Por exemplo, selecionar essa opção significa que a regra a seguir pode ser escorada:

```
bread & cheese -> wine
cheese & fruit -> wine
```

Desmarque essa caixa de seleção para excluir predições repetidas durante a escoragem.

Nota: Regras com diversos subsequentes (`bread & cheese & fruit -> wine & pate`) serão consideradas predições repetidas apenas se todos os subsequentes (`wine & pate`) tiverem sido preditos antes.

Escorar regras apenas quando as predições não estiverem presentes na entrada Selecione para assegurar que os subsequentes não estejam presentes também na entrada. Por exemplo, se o propósito da escoragem é fazer uma recomendação de móveis domésticos, então é improvável que uma entrada que já contenha uma mesa de jantar compre outra mesa. Nesse caso, selecione esta opção. Por outro lado, se os produtos forem perecíveis ou descartáveis (como queijo, alimento para bebês ou papel higiênico), então as regras nas quais o subsequente já estiver presente na entrada poderão ser úteis. Nesse último caso, a opção mais útil pode ser **Escorar todas as regras**.

Escorar regras apenas quando as predições estiverem presentes na entrada Selecione essa opção para assegurar que os subsequentes estejam presentes também na entrada. Essa abordagem é útil quando você estiver tentando ganhar insight sobre clientes ou transações existentes. Por exemplo, você pode querer identificar regras com a elevação mais alta e, em seguida, explorar quais clientes se enquadram nessas regras.

Escorar todas as regras Selecione essa opção para incluir todas as regras durante a escoragem, independentemente da presença ou da ausência dos subsequentes na entrada.

Capítulo 13. Modelos de Série Temporal

Por que Prever?

A previsão significa prever os valores de uma ou mais séries ao longo do tempo. Por exemplo, você pode querer prever a demanda esperada de uma linha de produtos ou serviços para alocar recursos para manufatura e distribuição. Como o planejamento de decisões leva tempo para implementar, as previsões são uma ferramenta essencial para muitos processos de planejamento.

Os métodos de modelagem de séries temporais supõem que o histórico se repete - se não exatamente, próximo o bastante para tomar as melhores decisões no futuro estudando o passado. Para prever as vendas para o próximo ano, por exemplo, você provavelmente começa analisando as vendas deste ano e trabalha retroativamente para descobrir quais tendências ou padrões, se houver, foram desenvolvidos nos últimos anos. Porém os padrões podem ser difíceis de avaliar. Se as suas vendas aumentarem em várias semanas seguidas, por exemplo, isso faz parte de um ciclo sazonal ou é o início de uma tendência em longo prazo?

Utilizando as técnicas de modelagem estatística, é possível analisar os padrões em seus dados passados e projetar esses padrões para determinar um intervalo dentro do qual os valores futuros da série deverão cair. O resultado é previsões mais precisas nas quais basear suas decisões.

Dados de Séries Temporais

Uma **série temporal** é uma coleção ordenada de medições feitas em intervalos regulares -- por exemplo, preços de ações diários ou dados de vendas semanais. As medições podem ser tudo aquilo que for de seu interesse, e cada série pode ser classificada geralmente da seguinte forma:

- **Dependente.** Uma série que você deseja prever.
- **Preditor.** Uma série que pode ajudá-lo a explicar a resposta -- por exemplo, utilizando um orçamento de publicidade para prever vendas. Os preditores podem ser utilizados apenas com modelos ARIMA.
- **Evento.** Uma série preditora especial utilizada para contabilizar incidentes recorrentes previsíveis – por exemplo, promoções de vendas.
- **Intervenção.** Uma série preditora especial utilizada para contabilizar incidentes passados únicos - por exemplo, uma indisponibilidade de energia ou greve de funcionário.

Os intervalos podem representar qualquer unidade de tempo, no entanto, o intervalo deve ser o mesmo para todas as medições. Além disso, qualquer intervalo para o qual não houver nenhuma medição deverá ser configurado como valor omisso. Assim, o número de intervalos com medições (incluindo aqueles com valores omissos) define o período de tempo do span de histórico dos dados.

Características de Séries Temporais

Estudar o comportamento passado de uma série ajuda a identificar padrões e fazer as melhores predições. Quando representadas, muitas séries temporais demonstram um ou mais dos seguintes recursos:

- Tendências
- Ciclos sazonais e não sazonais
- Pulsos e passos
- Valores discrepantes

Tendências

Uma **tendência** é uma mudança gradual para cima ou para baixo no nível da série ou da tendência dos valores de série para aumentar ou diminuir ao longo do tempo.

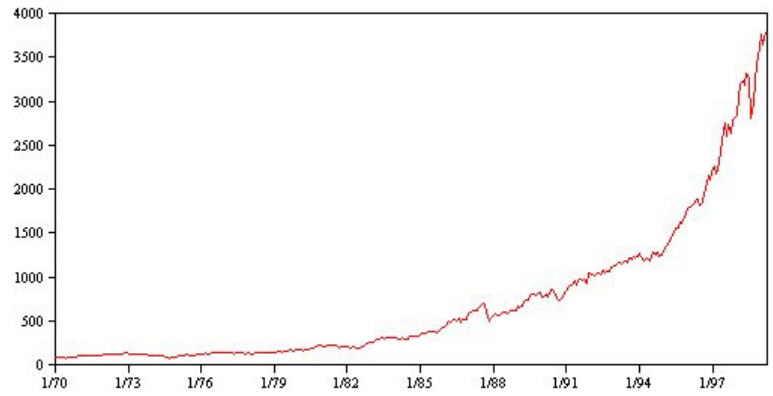


Figura 53. Tendência

As tendências são **locais** ou **globais**, mas uma única série pode exibir ambos os tipos. Historicamente, os gráficos de séries do índice do mercado de ações mostram uma tendência global para cima. As tendências locais para baixo aparecem em tempos de recessão e as tendências locais para cima apareceram em tempos de prosperidade.

As tendências também podem ser **lineares** ou **não lineares**. As tendências lineares são incrementos aditivos positivos ou negativos no nível da série, comparado com o efeito do interesse simples no principal. As tendências não lineares são muitas vezes multiplicativas, com incrementos que são proporcionais a um ou mais valores de série anteriores.

As tendências lineares globais são bem ajustadas e previstas pelos modelos de suavização exponencial ou ARIMA. Na construção de modelos ARIMA, as séries mostram tendências que são geralmente diferenciadas para remover o efeito da tendência.

Ciclos Sazonais

Um **ciclo sazonal** é um padrão repetitivo previsível nos valores da série.

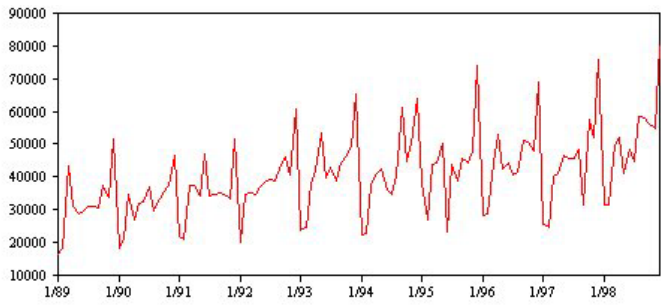


Figura 54. Ciclo sazonal

Os ciclos sazonais estão ligados ao intervalo de sua série. Por exemplo, dados mensais normalmente abrangem trimestres e anos. Uma série mensal pode mostrar um ciclo trimestral significativo com uma caída no primeiro trimestre ou um ciclo anual com um pico em cada Dezembro. Declara-se que as séries que mostram um ciclo sazonal demonstram **sazonalidade**.

Os padrões sazonais são úteis na obtenção de bons ajustes e previsões e há modelos de suavização exponencial e ARIMA que capturam sazonalidade.

Ciclos Não Sazonais

Um **ciclo não sazonal** é um padrão repetitivo e possivelmente imprevisível nos valores de série.

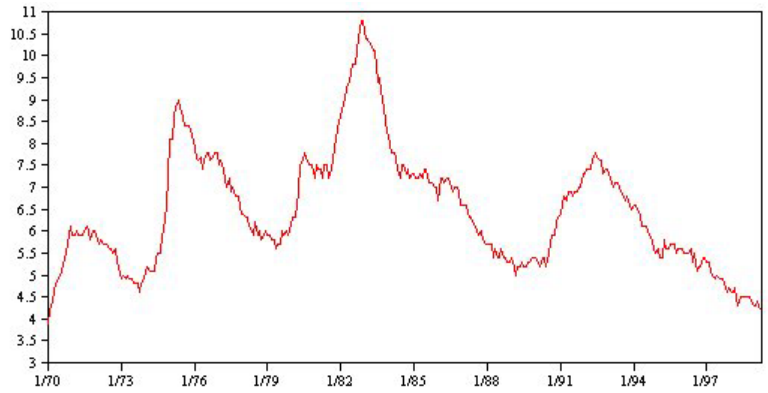


Figura 55. Ciclo não sazonal

Algumas séries, como taxa de desemprego, exibem claramente o comportamento cíclico, no entanto, a periodicidade do ciclo varia ao longo do tempo, tornando difícil prever quando uma alta ou baixa ocorrerá. Outras séries podem ter ciclos previsíveis, mas não se ajustam de modo organizado no calendário gregoriano ou possuem ciclos maiores que um ano. Por exemplo, as marés seguem o calendário lunar, viagens e comércio internacionais relacionados a Jogos Olímpicos aumentam de quatro em quatro anos e há muitos feriados religiosos em que as datas gregorianas alteram de ano a ano.

Os padrões cíclicos não sazonais são difíceis de modelar e geralmente aumentam a incerteza na hora da previsão. O mercado de ações, por exemplo, fornece várias instâncias de séries que vem desafiando os esforços dos previsores. Todos os mesmos padrões não sazonais devem ser considerados quando eles existirem. Em muitos casos, ainda é possível identificar um modelo que ajuste os dados históricos razoavelmente bem, fornecendo a melhor chance de minimizar a incerteza na previsão.

Pulsos e Passos

Muitas séries experimentam mudanças bruscas no nível. Elas geralmente aparecem em dois tipos:

- Uma mudança *temporária* repentina, ou **pulso**, no nível da série
- Uma mudança *permanente* repentina, ou **passo**, no nível da série

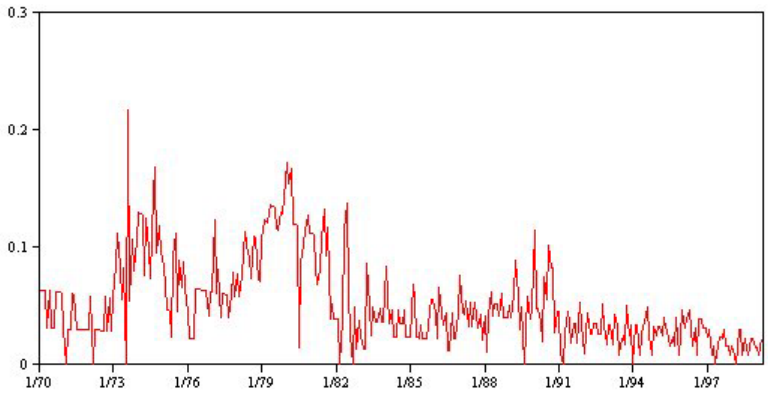


Figura 56. Séries com um pulso

Quando passos ou pulsos são observados, é importante localizar uma explicação plausível. Os modelos de séries temporais são projetados para considerar uma mudança gradual não repentina. Como resultado, eles tendem a subestimar pulsos e serem arruinados pelos passos, levando a ruins ajustes de modelo e

previsões incertas. (Algumas instâncias de sazonalidade podem aparentar exibir mudanças repentinas no nível, mas o nível é constante de um período sazonal para outro).

Se uma perturbação puder ser explicada, ela poderá ser modelada utilizando uma **intervenção** ou um **evento**. Por exemplo, em agosto de 1973, um embargo de petróleo imposto pela Organization of Petroleum Exporting Countries (OPEC) causou uma mudança drástica nas taxas de inflação, que voltou aos níveis normais nos meses seguintes. Ao especificar uma **intervenção de ponto** para o mês do embargo, é possível melhorar o ajuste de seu modelo, melhorando, assim, suas previsões indiretamente. Por exemplo, uma loja varejista pode achar que as vendas foram muito maiores que o normal no dia em que todos os itens foram marcados como 50% de desconto. Ao especificar a promoção de 50% de desconto como um **evento** recorrente, é possível melhorar o ajuste de seu modelo e estimar o efeito de repetir a promoção futuramente.

Valores discrepantes

As mudanças no nível de uma série temporal que não puderem ser explicadas são referidas como **valores discrepantes**. Estas observações estão inconsistentes com o restante da série e podem influenciar significativamente a análise e, consequentemente, afetar a capacidade de previsão do modelo de série temporal.

A figura a seguir exibe vários tipos de valores discrepantes que normalmente ocorrem em séries temporais. As linhas azuis representam uma série sem valores discrepantes. As linhas vermelhas sugerem um padrão que poderá estar presente se a série contiver valores discrepantes. Todos esses valores discrepantes são classificados como **deterministas** porque eles afetam somente o nível médio da série.

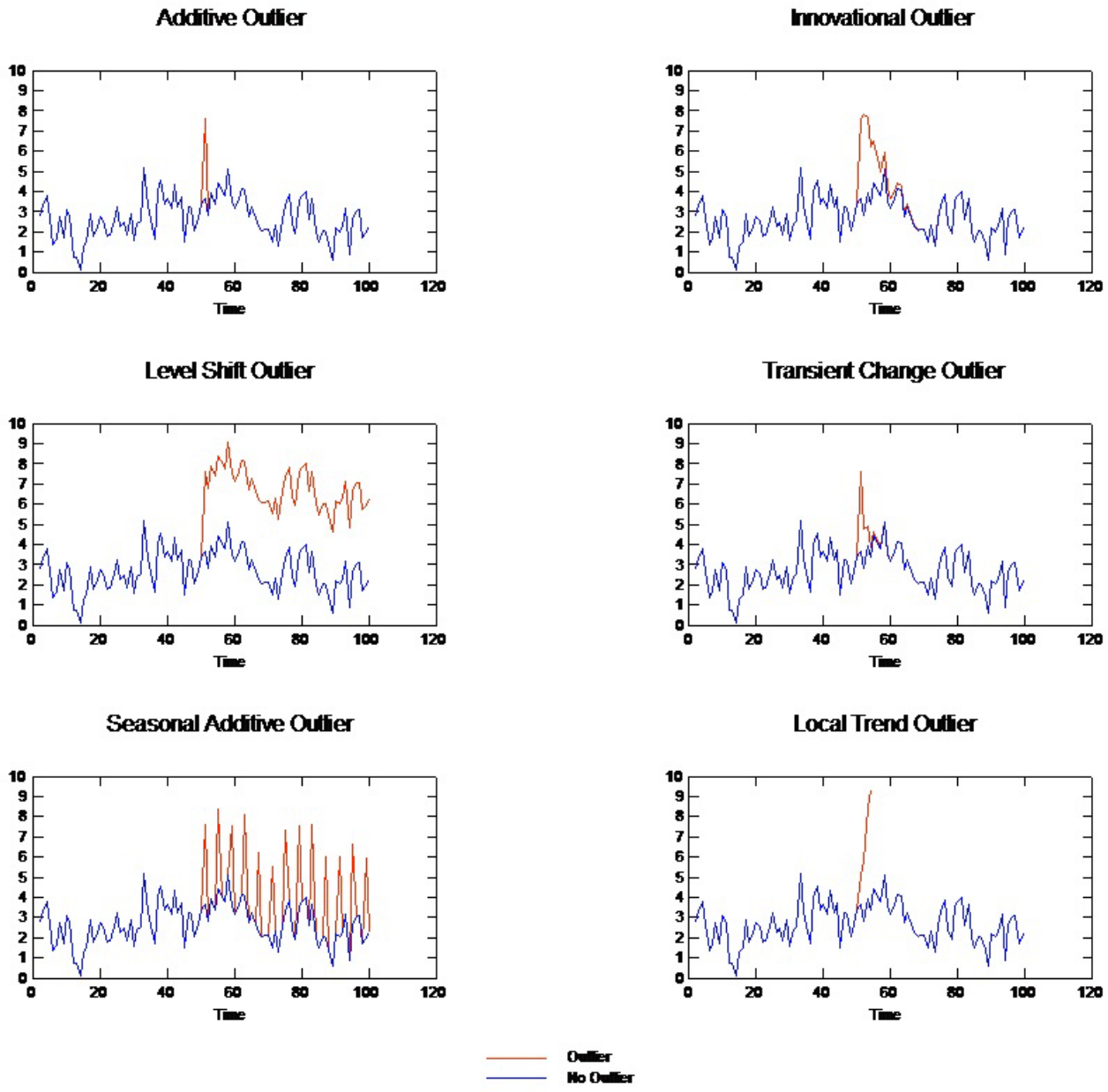


Figura 57. Tipos de valores discrepantes

- **Valor Discrepante Aditivo.** Um valor discrepante aditivo aparece como um valor surpreendentemente grande ou pequeno ocorrendo para uma observação única. Observações subsequentes não são afetadas por um valor discrepante aditivo. Os valores discrepantes aditivos consecutivos normalmente são referidos como **correções de valor discrepante aditivo**.
- **Valor Discrepante Inovador.** Um valor discrepante inovador é caracterizado por um impacto inicial com efeitos que aguardam observações subsequentes. A influência dos valores discrepantes pode aumentar com o decorrer do tempo.
- **Valor Discrepante de Mudança de Nível.** Para uma mudança de nível, todas as observações que aparecerem após o valor discrepante se movem para um novo nível. Em contraste com valores discrepantes aditivos, um valor discrepante de mudança de nível afeta muitas observações e tem um efeito permanente.

- **Valor Discrepante de Mudança Temporária.** Os valores discrepantes de mudança temporária são semelhantes aos valores discrepantes de mudança de nível, mas o efeito do valor discrepante diminui exponencialmente durante as observações subsequentes. Eventualmente, a série retorna para seu nível normal.
- **Valor Discrepante Aditivo Sazonal.** Um valor discrepante aditivo sazonal aparece como um valor surpreendentemente grande ou pequeno ocorrendo repetidamente em intervalos regulares.
- **Valor Discrepante de Tendência Local.** Um valor discrepante de tendência local produz um desvio geral na série causado por um padrão nos valores discrepantes após o início do valor discrepante inicial.

A detecção de valor discrepante em séries temporais envolve determinar o local, o tipo e a magnitude de quaisquer valores discrepantes presentes. Tsay (1988) propôs um processo iterativo para detecção da mudança do nível médio para identificar valores discrepantes determinísticos. Esse processo envolve comparar um modelo de série temporal que se supõe que nenhum valor discrepante esteja presente com outro modelo que incorpora valores discrepantes. As diferenças entre os modelos geram estimativas de efeito de tratar qualquer ponto específico como um valor discrepante.

Funções de Autocorrelação e de Autocorrelação Parcial

A autocorrelação e a autocorrelação parcial são medidas de associação entre valores de séries atuais e anteriores e indicam quais valores de série anteriores são mais úteis para prever valores futuros. Com esse conhecimento, é possível determinar a ordem dos processos no modelo ARIMA. Mais especificamente,

- **Função de autocorrelação (FAC).** No lag *k*, esta é a correlação entre os valores de série que são intervalos *k* separados.
- **Função de autocorrelação parcial (FACP).** No lag *k*, esta é a correlação entre os valores de série que são intervalos *k* separados, considerando os valores entre os intervalos.

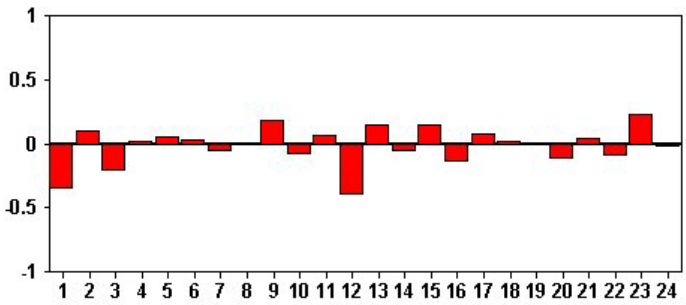


Figura 58. Gráfico FAC para uma série

O eixo *x* do gráfico FAC indica o lag no qual a autocorrelação é calculada e o eixo *y* indica o valor da correlação (entre -1 e 1). Por exemplo, um aumento no lag 1 em um gráfico FAC indica uma forte correlação entre cada valor de série e o valor anterior, um aumento no lag 2 indica uma forte correlação entre cada valor e o valor que ocorre dois pontos antes, e assim por diante.

- Uma correlação positiva indica que grandes valores atuais correspondem aos valores grandes no lag especificado, e uma correlação negativa indica que grandes valores atuais correspondem aos valores pequenos no lag especificado.
- O valor absoluto de uma correlação é uma medida da força da associação, com valores absolutos maiores indicando relacionamentos mais fortes.

Transformações de Série

As transformações são geralmente úteis para estabilizar uma série antes de estimar os modelos. Isso é importante principalmente para modelos ARIMA, que requerem que a série seja **estacionária** antes de os modelos serem estimados. Uma série será estacionária se o nível global (média) e desvio médio do nível (variância) forem constantes em toda a série.

Enquanto as séries de maior interesse não são estacionárias, o ARIMA entrará em vigor somente se essa série puder se tornar estacionária pela aplicação de transformações, como logaritmo natural, diferenciação ou diferenciação sazonal.

Transformações de estabilização de variância. A série na qual a variância se altera ao longo do tempo geralmente pode ser estabilizada usando uma transformação logarítmica natural ou uma transformação raiz quadrada. Essas também são chamadas de transformações funcionais.

- **Logarítmica natural.** O logaritmo natural é aplicado aos valores de série.
- **Raiz quadrada.** A função de raiz quadrada é aplicada aos valores de série.

As transformações logarítmica natural e raiz quadrada não podem ser utilizadas para séries com valores negativos.

Transformações de estabilização de nível. Uma recusa lenta dos valores na FAC indica que cada valor de série está fortemente correlacionado ao valor anterior. Ao analisar a mudança nos valores de série, você obtém um nível estável.

- **Diferenciação simples.** As diferenças entre cada valor e o valor anterior na série são calculadas, exceto o valor mais antigo na série. Isso significa que a série diferenciada terá um valor a menos que a série original.
- **Diferenciação sazonal.** Idêntico à diferenciação simples, exceto que as diferenças entre cada valor e o valor sazonal anterior são calculadas.

Quando uma diferenciação simples ou sazonal estiver simultaneamente em uso com a transformação logarítmica ou raiz quadrada, a transformação de estabilização de variância sempre é aplicada primeiro. Quando as diferenciações simples e sazonais estiverem em uso, os valores de série resultantes serão os mesmos, independentemente se a diferenciação simples ou a diferenciação sazonal for aplicada primeiro.

Série do Preditor

A série do preditor inclui dados relacionados que podem ajudar a explicar o comportamento da série a ser prevista. Por exemplo, um varejista com base na web ou em catálogo pode prever as vendas com base no número de catálogos enviados por email, no número de linhas telefônicas abertas ou no número de ocorrências na página da Web da empresa.

Qualquer série pode ser utilizada como um preditor, desde que a série se estenda no futuro até o ponto em que você deseja prever e que possua dados completos sem valores omissos.

Tenha cuidado ao incluir preditores em um modelo. Incluir grandes quantidades de preditores aumenta o tempo necessário para estimar modelos. Embora incluir preditores melhore a capacidade de um modelo em ajustar os dados históricos, isso não significa necessariamente que o modelo faz um melhor trabalho de previsão, portanto, isso só aumenta a complexidade e não vale a pena o esforço. Idealmente, o objetivo é identificar o modelo mais simples que faça um bom trabalho de previsão.

Como regra geral, recomenda-se que o número de preditores seja menor que o tamanho da amostra dividido por 15 (no máximo, um preditor para 15 casos).

Preditores com dados omissos. Preditores com dados omissos ou incompletos não podem ser utilizados na previsão. Isso se aplica a dados históricos e valores futuros. Em alguns casos, é possível evitar esta limitação ao configurar o span de estimação do modelo para excluir os dados mais antigos quando estimar os modelos.

Nó de Modelagem de Séries Temporais

O nó Séries Temporais estima modelos de suavização exponencial, Média Móvel Integrada AutoRegressiva (ARIMA) univariada e ARIMA multivariada (ou função de transferência) para séries temporais e produz previsões com base nos dados de séries temporais.

A **Suavização exponencial** é um método de previsão que usa valores de ponderação de observações de séries anteriores para prever valores futuros. Assim, a suavização exponencial não se baseia em um entendimento teórico dos dados. Ela prevê um ponto por vez, ajustando suas previsões conforme novos dados aparecem. A técnica é útil para a previsão de séries que exibem tendência, sazonalidade, ou ambos. É possível escolher entre uma variedade de modelos de suavização exponencial que diferem em termos de tratamento de tendência e sazonalidade.

Os modelos **ARIMA** fornecem métodos mais sofisticados para modelar tendência e componentes sazonais do que os modelos de suavização exponencial e, em particular, têm um benefício maior de permitir inclusão de variáveis independentes (preditoras) no modelo. Isso envolve especificar explicitamente as ordens autorregressivas e de média móvel, bem como o grau de diferenciação. É possível incluir variáveis preditoras e definir funções de transferência para qualquer uma ou todas elas, bem como especificar detecção automática de valores discrepantes ou de um conjunto explícito de valores discrepantes.

Nota: na prática, os modelos ARIMA serão mais úteis se desejar incluir preditores que possam ajudar a explicar o comportamento das séries que estiverem sendo previstas, como o número de catálogos enviados por email ou o número de ocorrências em uma página da web da empresa. Os modelos de suavização exponencial descrevem o comportamento da série temporal sem tentar entender por que ela se comporta dessa maneira. Por exemplo, uma série que historicamente atinge um pico a cada 12 meses provavelmente continuará fazendo isso, mesmo se você não souber por quê.

Também está disponível um **Modelador Especialista**, que tenta identificar e estimar automaticamente o modelo ARIMA ou de suavização exponencial de melhor ajuste para uma ou mais variáveis de destino, eliminando, assim, a necessidade de identificar um modelo apropriado por meio de avaliação e erro. Na dúvida, utilize o Modelador Especialista

Se variáveis preditoras forem especificadas, o Modelador Especialista selecionará essas variáveis para inclusão nos modelos ARIMA que possuírem um relacionamento estatisticamente significativo com a série dependente. As variáveis de modelo são transformadas onde apropriado utilizando a diferenciação e/ou a transformação de raiz quadrada ou de logarítmica natural. Por padrão, o Modelador Especialista considera todos os modelos de suavização exponencial e todos os modelos ARIMA e seleciona o melhor modelo dentre eles para cada campo de destino. É possível, no entanto, limitar o Modelador Especialista para selecionar apenas o melhor dos modelos de suavização exponencial ou somente o melhor dos modelos ARIMA. Também é possível especificar detecção automática de valores discrepantes.

Exemplo. Um analista de um provedor de banda larga nacional é necessário para produzir previsões e assinaturas do usuário para prever a utilização da largura de banda. As previsões são necessárias para cada um dos mercados locais que formam a base do assinante nacional. É possível utilizar a modelagem de séries temporais para produzir previsões para os próximos três meses para um número de mercados locais.

Requisitos

O nó Séries Temporais é diferente de outros nós do IBM SPSS Modeler em que não é possível simplesmente inseri-lo em um fluxo e executar o fluxo. O nó Séries Temporais deve ser sempre precedido

por um nó Intervalos de Tempo que especifica informações como o intervalo de tempo a ser utilizado (anos, trimestres, meses, etc.) , os dados a serem utilizados para estimativa e até quando no futuro estender uma previsão, se utilizado.

Os dados de séries temporais devem ser igualmente espaçados. Os métodos para modelar dados de séries temporais requerem um intervalo uniforme entre cada medição, com quaisquer valores omissos indicados por linhas vazias. Se seus dados ainda não atenderem a esse requisito, o nó Intervalos de Tempo poderá transformar valores conforme necessário.

Outros pontos a serem observados com relação aos nós Séries Temporais são:

- Os campos devem ser numéricos
- Os campos de data não podem ser utilizados como entradas
- As partições são ignoradas

Opções do Campo

A guia Campos é onde você especifica os campos a serem utilizados na construção do modelo. Antes de poder construir um modelo, é necessário especificar quais campos você deseja utilizar como destinos e como entradas. Geralmente, o nó Séries Temporais utiliza as informações do campo a partir de um nó Tipo de envio de dados. Se você estiver utilizando um nó Tipo para selecionar campos de entrada e de destino, não será necessário alterar nada nesta guia.

Usar configurações do nó de tipo. Essa opção instrui o nó a utilizar as informações de campo a partir de um nó Tipo de envio de dados. Esse é o padrão.

Usar configurações customizadas. Essa opção instrui o nó a utilizar as informações de campo especificadas aqui ao invés das informações fornecidas em qualquer nó ou nós Tipo de envio de dados. Após selecionar esta opção, especifique os campos abaixo. Observe que os campos armazenados como datas não são aceitos como campos de destino ou de entrada.

- **Destinos.** Selecione um ou mais campos de destino. Isso é semelhante a configurar o papel do campo para *Destino* em um nó Tipo. Os campos de destino para um modelo de série temporal devem ter um nível de medição de *Contínuo*. Um modelo separado é criado para cada campo de destino. Um campo de destino considera todos os campos de *Entrada* especificados, exceto si mesmo, como entradas possíveis. Assim, o mesmo campo poderá ser incluído em ambas as listas; esse campo será utilizado como uma entrada possível para todos os modelos, exceto aquele no qual ele é um destino.
- **Entradas.** Selecione um ou mais campos de entrada. Isso é semelhante a configurar o papel do campo para *Entrada* em um nó Tipo. Os campos de entrada para um modelo de série temporal devem ser numéricos.

Opções do Modelo de Série Temporal

Nome do modelo. Especifica o nome designado ao modelo que é gerado quando o nó é executado.

- **Automático.** Gera o nome do modelo automaticamente com base nos nomes de campo de destino ou de ID ou no nome do tipo de modelo nos casos em que nenhum destino é especificado (como modelos de armazenamento em cluster).
- **Customizado.** Permite especificar um nome customizado para o nugget do modelo.

Continuar a estimação usando modelo(s) existente(s). Se você já gerou um modelo de série temporal, selecione esta opção para reutilizar as configurações de critérios especificadas para esse modelo e gerar um novo nó de modelo na paleta Modelos ao invés de construir um novo modelo desde o início. Dessa forma, é possível economizar tempo ao reestimar e produzir uma nova previsão que seja baseada nas mesmas configurações de modelo como antes, porém utilizando dados mais recentes. Assim, por exemplo, se o modelo original de uma série temporal específica era tendência linear de Holt, o mesmo tipo de modelo será utilizado para reestimar e prever esses dados. O sistema não tenta localizar novamente o melhor tipo de modelo para os novos dados.

Selecionar essa opção desativa os controles de **Método** e **Critérios**. Consulte o tópico "Reestimando e Prevendo" na página 291 para obter informações adicionais.

Método. É possível escolher Modelador Especialista, Suavização Exponencial ou ARIMA. Consulte o tópico "Nó de Modelagem de Séries Temporais" na página 284 para obter mais informações. Selecione **Critérios** para especificar opções para o método selecionado.

- **Modelador Especialista.** Escolha esta opção para utilizar o Modelador Especialista, que localiza automaticamente o modelo de melhor ajuste para cada série dependente.
- **Suavização Exponencial.** Utilize esta opção para especificar um modelo de suavização exponencial customizado.
- **ARIMA.** Utilize esta opção para especificar um modelo ARIMA customizado.

Informações de Intervalo de Tempo

Esta seção da caixa de diálogo contém informações sobre as especificações para estimativas e previsões feitas no nó Intervalos de Tempo. Observe que esta seção não será exibida se você escolher a opção **Continuar a estimação usando modelo(s) existente(s)**.

A primeira linha das informações indica se quaisquer registros são excluídos do modelo ou utilizados como validação.

A segunda linha fornece informações sobre quaisquer períodos de previsão especificados no nó Intervalos de Tempo.

Se a primeira linha lê **Nenhum intervalo de tempo definido**, o que indica que nenhum nó Intervalos de Tempo está conectado. Essa situação causará um erro ao tentar executar o fluxo; deve-se incluir um nó Intervalos de Tempo antes do nó Séries Temporais.

Informações Diversas

Largura do limite de confiança (%). Os intervalos de confiança são calculados para as predições do modelo e para autocorrelações residuais. É possível especificar quaisquer valores positivos menores que 100. Por padrão, um intervalo de confiança de 95% é utilizado.

Número máximo de lags na saída FAC e FACP. É possível configurar o número máximo de lags mostrados em tabelas e gráficos de autocorrelações e autocorrelações parciais.

Apenas modelo de escoragem de construção. Marque esta caixa para reduzir a quantidade de dados que é armazenada no modelo. Isso poderá melhorar o desempenho durante a construção de modelos com números muito grandes de séries temporais (dezenas de milhares). Se essa opção for selecionada, as guias Modelo, Parâmetros e Residuais não serão exibidas no nugget do modelo Séries Temporais, mas ainda será possível escorar os dados como de costume.

Critérios do Modelador Especialista de Séries Temporais

Tipo de modelo. As opções a seguir estão disponíveis:

- **Todos os modelos.** O Modelador Especialista considera os modelos de suavização ARIMA e exponencial.
- **Apenas modelos de suavização exponencial.** O Modelador Especialista considera apenas os modelos de suavização exponencial.
- **Apenas modelos ARIMA.** O Modelador Especialista considera apenas os modelos ARIMA.

Modelador Especialista considera modelos sazonais. Essa opção será ativada apenas se uma periodicidade tiver sido definida para o conjunto de dados ativo. Quando essa opção é selecionada, o

Modelador Especialista considera modelos sazonais e não sazonais. Se essa opção não estiver selecionada, o Modelador Especialista considerará apenas modelos não sazonais.

Eventos e Intervenções. Permite designar determinados campos de entrada como campos de evento ou de intervenção. Fazer isso identifica um campo como contendo dados de séries temporais afetados por eventos (situações recorrentes previsíveis, por exemplo, promoções de vendas) ou intervenções (incidentes únicos, por exemplo, indisponibilidade de energia ou greve de funcionário). O Modelador Especialista considerará somente regressão simples e não funções de transferência arbitrárias de entradas identificadas como campos de evento ou de intervenção.

Os campos de entrada devem ter um nível de medição de *Flag*, *Nominal* ou *Ordinal* e ser numéricos (por exemplo, 1/0, não True/False, para um campo de flag), antes de serem incluídos nessa lista. Consulte o tópico "Pulsos e Passos" na página 279 para obter mais informações.

Valores discrepantes

Detectar valores discrepantes automaticamente. Por padrão, a detecção automática de valores discrepantes não é executada. Selecione esta opção para executar detecção automática de valores discrepantes e, em seguida, selecione os tipos de valores discrepantes desejados. Consulte o tópico "Valores discrepantes" na página 280 para obter mais informações.

Critérios de Suavização Exponencial de Séries Temporais

Tipo de modelo. Os modelos de suavização exponencial são classificados como [1] sazonais ou não sazonais. Os modelos Sazonais estarão disponíveis apenas se a periodicidade definida utilizando o nó Intervalos de Tempo for sazonal. As periodicidades sazonais são: períodos cíclicos, anos, trimestres, meses, dias por semana, horas por dia, minutos por dia, e segundos por dia.

- **Simples.** Esse modelo é apropriado para uma série na qual não houver nenhuma tendência ou sazonalidade. O único parâmetro de suavização relevante é nível. A suavização exponencial simples é mais semelhante a um ARIMA com zero ordem de autorregressão, uma ordem de diferenciação, uma ordem de média de movimentação e nenhuma constante.
- **Tendência linear de Holt.** Esse modelo é apropriado para uma série na qual houver uma tendência linear e nenhuma sazonalidade. Seus parâmetros de suavização relevantes são nível e tendência e, nesse modelo, eles não são restritos pelos valores uns dos outros. O modelo de Holt é mais geral do que o modelo de Brown, mas poderá demorar mais para calcular estimativas para grandes séries. A suavização exponencial de Holt é mais semelhante a um ARIMA com zero ordem de autorregressão, duas ordens de diferenciação e duas ordens de média móvel.
- **Tendência linear de Brown.** Esse modelo é apropriado para uma série na qual houver uma tendência linear e nenhuma sazonalidade. Seus parâmetros de suavização relevantes são nível e tendência, no entanto, nesse modelo, eles são considerados iguais. O modelo de Brown é, portanto, um caso especial do modelo de Holt. A suavização exponencial de Brown é mais semelhante a um ARIMA com zero ordem de autorregressão, duas ordens de diferenciação e duas ordens de média móvel, com o coeficiente para a segunda ordem de média móvel igual à metade do coeficiente da primeira ordem ao quadrado.
- **Tendência amortecida.** Esse modelo é apropriado para uma série na qual houver uma tendência linear que está desaparecendo e nenhuma sazonalidade. Seus parâmetros de suavização relevantes são nível, tendência e tendência de amortecimento. A suavização exponencial amortecida é mais semelhante a um ARIMA com uma ordem de autorregressão, uma ordem da diferenciação e duas ordens de média móvel.
- **Sazonal simples.** Esse modelo é apropriado para uma série sem nenhuma tendência e com um efeito sazonal que seja constante ao longo do tempo. Seus parâmetros de suavização relevantes são nível e season. A suavização exponencial sazonal é mais semelhante a um ARIMA com zero ordem de

1. Gardner, E. S. 1985. Exponential smoothing: The state of the art. *Journal of Forecasting,* 4, 1-28.

autorregressão, uma ordem de diferenciação, uma ordem da diferenciação sazonal e as ordens 1, *p* e *p*+1 de média móvel, em que *p* é o número de períodos em um intervalo sazonal. Para dados mensais, *p*=12.

- **Aditiva de Winter.** Esse modelo é apropriado para uma série na qual houver uma tendência linear e um efeito sazonal que seja constante ao longo do tempo. Seus parâmetros de suavização relevantes são nível, tendência e season. A suavização exponencial aditiva de Winter é muito semelhante a um ARIMA com zero ordem de autorregressão, uma ordem de diferenciação; uma ordem de diferenciação sazonal e *p*+1 ordens de média móvel, em que *p* é o número de períodos em um intervalo sazonal. Para dados mensais, *p*=12.
- **Multiplicativa de Winter.** Esse modelo é apropriado para uma série na qual houver uma tendência linear e um efeito sazonal que altera com a magnitude da série. Seus parâmetros de suavização relevantes são nível, tendência e season. A suavização exponencial multiplicativa de Winter não é semelhante a nenhum modelo ARIMA.

Transformação de Resposta. É possível especificar uma transformação para ser executada em cada variável dependente antes de ser modelada. Consulte o tópico "Transformações de Série" na página 283 para obter mais informações.

- **Nenhum.** Nenhuma transformação é executada.
- **Raiz quadrada.** A transformação raiz quadrada é executada.
- **Logarítmica natural.** A transformação logarítmica natural é executada.

Critérios do ARIMA de Séries Temporais

O nó Séries Temporais permite construir modelos ARIMA sazonais ou não sazonais customizados, também conhecidos como modelos Box-Jenkins, com ou sem um conjunto fixo de variáveis de entrada (preditoras) [2]. É possível definir funções de transferência para qualquer uma ou todas variáveis de entrada e especificar detecção automática de valores discrepantes ou de um conjunto explícito de valores discrepantes.

Todas as variáveis de entrada especificadas são explicitamente incluídas no modelo. Isso está em contraste com o uso do Modelador Especialista, em que as variáveis de entrada serão incluídas apenas se elas possuírem um relacionamento estatisticamente significativo com a variável de destino.

Modelo

A guia Modelo permite especificar a estrutura de um modelo ARIMA customizado.

Ordens ARIMA. Insira valores para os vários componentes ARIMA de seu modelo nas células correspondentes da grade de Estrutura. Todos os valores devem ser números inteiros não negativos. Para componentes autorregressivos e de média móvel, o valor representa a ordem máxima. Todas as ordens inferiores positivas serão incluídas no modelo. Por exemplo, se você especificar 2, o modelo incluirá as ordens 2 e 1. As células na coluna Sazonal serão ativadas apenas se uma periodicidade tiver sido definida para o conjunto de dados ativo.

- **Autorregressiva (p).** O número de ordens autorregressivas no modelo. As ordens autorregressivas especificam quais valores anteriores da série são utilizados para prever valores atuais. Por exemplo, uma ordem autorregressiva de 2 especifica que o valor de dois períodos de tempo da série no passado será utilizado para prever o valor atual.
- **Diferença (d).** Especifica a ordem de diferenciação aplicada à série antes de estimar os modelos. A diferenciação é necessária quando tendências estiverem presentes (as séries com tendências normalmente são não estacionárias e a modelagem ARIMA assume estacionariedade) e é utilizada para remover seus efeitos. A ordem da diferenciação corresponde ao grau de tendência das séries -- a

2. Box, G. E. P., G. M. Jenkins, e G. C. Reinsel. 1994. *Time series analysis: Forecasting and control*, 3ª ed. Englewood Cliffs, N.J.: Prentice Hall.

diferenciação de primeira ordem considera tendências lineares, a diferenciação de segunda ordem considera tendências quadráticas, e assim por diante.

- **Média Móvel (q).** O número de ordens de média móvel no modelo. As ordens de média móvel especificam como os desvios da média de série para valores anteriores são utilizados para prever valores atuais. Por exemplo, as ordens de média móvel de 1 e 2 especificam que os desvios do valor médio das séries de cada um dos dois últimos períodos de tempo são considerados ao prever valores atuais da série.

Ordens Sazonais. Os componentes autorregressivo, média móvel e de diferenciação sazonais desempenham os mesmos papéis que seus correspondentes não sazonais. Para ordens sazonais, no entanto, os valores atuais da série são afetados pelos valores anteriores da série separados por um ou mais períodos sazonais. Por exemplo, para dados mensais (período de sazonal de 12), uma ordem sazonal de 1 significa que o valor de série atual é afetado pelo valor de série 12 períodos anteriores ao período atual. Em seguida, uma ordem sazonal de 1, para os dados mensais, será o mesmo que especificar uma ordem não sazonal de 12.

Transformação de Resposta. É possível especificar uma transformação para ser executada em cada variável de destino antes de ser modelada. Consulte o tópico "Transformações de Série" na página 283 para obter mais informações.

- **Nenhum.** Nenhuma transformação é executada.
- **Raiz quadrada.** A transformação raiz quadrada é executada.
- **Logarítmica natural.** A transformação logarítmica natural é executada.

Incluir constante no modelo. A inclusão de uma constante é padrão, a menos que você tenha certeza de que o valor de série médio geral é 0. Excluir a constante é recomendado quando diferenciação for aplicada.

Transferir Funções

A guia Funções de Transferência permite definir funções de transferência para qualquer um ou todos os campos de entrada. As funções de transferência permitem especificar a maneira pela qual os valores passados destes campos são utilizados para prever valores futuros da série de resposta.

A guia será exibida apenas se os campos de entrada (com o papel configurado como *Entrada*) forem especificados, seja no nó Tipo ou na guia Campos do nó Séries Temporais (selecione **Usar configurações customizadas – Entradas**).

A lista superior mostra todos os campos de entrada. As informações restantes nesta caixa de diálogo são específicas para o campo de entrada selecionado na lista.

Ordens de Função de Transferência. Insira valores para os vários componentes da função de transferência nas células correspondentes da grade Estrutura. Todos os valores devem ser números inteiros não negativos. Para componentes de numerador e denominador, o valor representa a ordem máxima. Todas as ordens inferiores positivas serão incluídas no modelo. Além disso, a ordem 0 é sempre incluída para componentes de numerador. Por exemplo, se você especificar 2 para o numerador, o modelo incluirá as ordens 2, 1 e 0. Se você especificar 3 para o denominador, o modelo incluirá as ordens 3, 2, e 1. As células na coluna Sazonal serão ativadas apenas se uma periodicidade tiver sido definida para o conjunto de dados ativo.

Numerador. A ordem de numerador da função de transferência especifica quais valores anteriores da série independente (preditora) selecionada são utilizados para prever valores atuais da série dependente. Por exemplo, uma ordem de numerador de 1 especifica que o valor de um período de série independente no passado -- assim como o valor atual da série independente -- é utilizado para prever o valor atual de cada série dependente.

Denominador. A ordem de denominador da função de transferência especifica como os desvios da média de série, para os valores anteriores da série independente (preditora) selecionada, são utilizados para prever os valores atuais da série dependente. Por exemplo, uma ordem de denominador de 1 especifica que os desvios do valor médio de um período de série independente no passado são considerados ao prever o valor atual de cada série dependente.

Diferença. Especifica a ordem de diferenciação aplicada à série independente (preditora) selecionada antes de estimar os modelos. A diferenciação é necessária quando tendências estiverem presentes e é utilizada para remover seu efeito.

Ordens Sazonais. Os componentes de numerador, denominador e diferenciação sazonais desempenham os mesmos papéis que seus correspondentes não sazonais. Para ordens sazonais, no entanto, os valores atuais da série são afetados pelos valores anteriores da série separados por um ou mais períodos sazonais. Por exemplo, para dados mensais (período de sazonal de 12), uma ordem sazonal de 1 significa que o valor de série atual é afetado pelo valor de série 12 períodos anteriores ao período atual. Em seguida, uma ordem sazonal de 1, para os dados mensais, será o mesmo que especificar uma ordem não sazonal de 12.

Atraso. Configurar um atraso faz com que a influência do campo de entrada seja atrasada pelo número de intervalos especificados. Por exemplo, se o atraso for configurado como 5, o valor do campo de entrada no tempo t não afetará as previsões até que decorram cinco períodos ($t + 5$).

Transformação. A especificação de uma função de transferência para um conjunto de variáveis independentes também inclui uma transformação opcional a ser executada nessas variáveis.

- **Nenhum.** Nenhuma transformação é executada.
- **Raiz quadrada.** A transformação raiz quadrada é executada.
- **Logarítmica natural.** A transformação logarítmica natural é executada.

Manipulando Valores Discrepantes

A guia Valores Discrepantes fornece um número de opções para manipular valores discrepantes nos dados [3].

Não detectar valores discrepantes ou modelá-los. Por padrão, os valores discrepantes não são detectados nem modelados. Selecione esta opção para desativar qualquer detecção ou modelagem de valores discrepantes.

Detectar valores discrepantes automaticamente. Selecione esta opção para executar detecção automática de valores discrepantes, e selecione um ou mais dos tipos de valores discrepantes mostrados.

Tipo de Valores Discrepantes para Detectar. Selecione um ou mais tipos de valores discrepantes que deseja detectar. Os tipos suportados são:

- Aditivo (padrão)
- Mudança de nível (padrão)
- Inovador
- Transiente
- Aditivo sazonal
- Tendência local
- Curva de nível aditiva

Consulte o tópico “Valores discrepantes” na página 280 para obter mais informações.

3. Pena, D., G. C. Tiao, e R. S. Tsay, eds. 2001. *A course in time series analysis*. Nova York: John Wiley and Sons.

Gerando Modelos de Séries Temporais

Essa seção fornece algumas informações gerais sobre determinados aspectos da geração de modelos de séries temporais:

- Gerando diversos modelos
- Usando modelos de séries temporais na previsão
- Reestimando e prevendo

O nugget do modelo gerado é descrito em um tópico separado. Consulte o tópico “Nugget do Modelo de Série Temporal” na página 292 para obter mais informações.

Gerando Diversos Modelos

A modelagem de séries temporais no IBM SPSS Modeler gera um modelo único (ou ARIMA ou suavização exponencial) para cada campo de destino. Portanto, se você tiver diversos campos de destino, o IBM SPSS Modeler gerará diversos modelos em uma única operação, economizando tempo e permitindo comparar as configurações com cada modelo.

Se desejar comparar um modelo ARIMA e um modelo de suavização exponencial com o mesmo campo de destino, será possível desempenhar execuções separadas do nó Séries Temporais, especificando um modelo diferente a cada vez.

Utilizando Modelos de Série Temporal na Previsão

Uma operação de construção de série temporal utiliza uma série específica de casos ordenados, conhecidos como o span de estimação, para construir um modelo que possa ser utilizado para prever valores futuros da série. Este modelo contém informações sobre o período de tempo utilizado, incluindo o intervalo. Para prever o uso desse modelo, as mesmas informações de período de tempo e de intervalo devem ser utilizadas com a mesma série tanto para a variável de destino quanto para variáveis preditoras.

Por exemplo, suponha que no início de janeiro você deseja prever vendas mensais do Produto 1 para os primeiros três meses do ano. Você constrói um modelo utilizando os dados de vendas mensais reais do Produto 1 de janeiro a dezembro do ano anterior (que chamaremos de Ano 1), configurando o Intervalo de Tempo para "Meses." Em seguida, é possível utilizar o modelo para prever vendas do Produto 1 para os três primeiros meses do Ano 2.

Na realidade, é possível prever qualquer número de meses adiante, mas é claro que quanto mais no futuro você tentar prever, menos eficaz será o modelo. No entanto, não é possível prever as três primeiras semanas do Ano 2 porque o intervalo utilizado para construir o modelo foi "Meses." Também não faz sentido utilizar esse modelo para prever as vendas do Produto 2 -- um modelo de série temporal é relevante apenas para os dados que foram utilizados para defini-lo.

Reestimando e Prevendo

O período de estimação é codificado permanentemente no modelo que é gerado. Isso significa que quaisquer valores fora do período de estimação serão ignorados se você aplicar o modelo atual aos novos dados. Portanto, um modelo de série temporal deverá ser reestimado cada vez que novos dados estiverem disponíveis, em contraste com outros modelos do IBM SPSS Modeler que podem ser reaplicados inalterados para fins de escoragem.

Para continuar o exemplo anterior, suponha que até o início de abril no Ano 2 você tenha os dados de vendas mensais reais de janeiro a março. No entanto, se você reaplicar o modelo gerado no início de janeiro, ele fará novamente uma previsão de janeiro a março e ignorará os dados de vendas conhecidos para esse período.

A solução é gerar um novo modelo com base nos dados reais atualizados. Supondo que você não altere os parâmetros de previsão, o novo modelo poderá ser utilizado para prever os próximos três meses, de abril a junho. Se você ainda tiver acesso ao fluxo que foi utilizado para gerar o modelo original, será

possível simplesmente substituir a referência ao arquivo de origem nesse fluxo por uma referência ao arquivo que contém os dados atualizados e executar novamente o fluxo para gerar o novo modelo. No entanto, se tudo o que você tiver for apenas o modelo original salvo em um arquivo, ele ainda poderá ser usado para gerar um nó Séries Temporais que, em seguida, poderá ser incluído em um novo fluxo contendo uma referência ao arquivo de origem atualizado. Contanto que este novo fluxo preceda o nó Séries Temporais com um nó Intervalos de Tempo no qual o intervalo é configurado para "Meses," executar este novo fluxo gerará então o novo modelo necessário.

Nugget do Modelo de Série Temporal

Uma operação de modelagem de série temporal cria diversos novos campos com o prefixo $TS-, conforme mostrado na tabela a seguir.

Tabela 24. Novos campos criados pela operação de modelagem de séries temporais.

Nome do campo	Descrição
$TS-*colname*	O valor previsto pelo modelo para cada série de resposta.
$TSLCI-*colname*	Os intervalos de confiança inferiores para cada série prevista.*
$TSUCI-*colname*	Os intervalos de confiança superiores para cada série prevista.*
$TSNR-*colname*	O valor residual de ruído para cada coluna dos dados de modelo gerados.*
$TS-Total	O total dos valores $TS-*colname* para essa linha.
$TSLCI-Total	O total dos valores $TSLCI-*colname* para esta linha.*
$TSUCI-Total	O total dos valores $TSUCI-*colname* para esta linha.*
$TSNR-Total	O total dos valores $TSNR-*colname* para esta linha.*

* A visibilidade desses campos (por exemplo, na saída de um nó Tabela anexado) depende das opções na guia Configurações do nugget do modelo Séries Temporais. Consulte o tópico “Configurações de Modelo de Série Temporal” na página 295 para obter mais informações.

O nugget do modelo Séries Temporais exibe detalhes sobre os vários modelos selecionados para cada uma das entradas de série no nó de construção Séries Temporais. Diversas séries (como dados relacionados às linhas de produtos, regiões ou armazenamentos) podem ser inseridas, e um modelo separado é gerado para cada série de resposta. Por exemplo, se a renda na região oriental for considerada adequada para um modelo ARIMA, mas a região ocidental for adequada apenas para uma média móvel simples, cada região será escorada com o modelo apropriado.

A saída padrão mostra, para cada modelo construído, o tipo do modelo, o número de preditores especificado e a medida de qualidade de ajuste (*R*-quadrado estacionário é o padrão). Se você tiver especificado métodos de valores discrepantes, haverá uma coluna mostrando o número de valores discrepantes detectados. A saída padrão também inclui colunas para Ljung-Box Q, graus de liberdade e valores de significância.

Também é possível escolher saída avançada, que exibe as colunas adicionais a seguir:

- R-quadrado
- RMSE (Erro Quadrático Médio Raiz)
- MAPE (Erro Percentual Absoluto Médio)
- MAE (Erro Médio Absoluto)
- MaxAPE (Erro Percentual Absoluto Máximo)
- MaxAE (Erro Absoluto Máximo)
- Norm. BIC (Critério de Informação Bayesiano Normalizado)

Gerar. Permite gerar um nó de modelagem de Séries Temporais de volta para o fluxo ou um nugget do modelo na paleta.

- **Gerar Nó de Modelagem.** Coloca um nó de modelagem de Séries Temporais em um fluxo com as configurações utilizadas para criar esse conjunto de modelos. Fazer isso será útil, por exemplo, se você tiver um fluxo no qual deseja utilizar estas configurações de modelo, mas não tiver mais o nó de modelagem utilizado para gerá-los.
- **Modelo para Paleta.** Coloca um nugget do modelo que contém todas as respostas no gerenciador de Modelos.

Modelo

Figura 59. Botões Selecionar Tudo e Desmarcar Tudo

Caixas de seleção. Escolha quais modelos deseja utilizar na escoragem. Todas as caixas são selecionadas por padrão. Os botões **Selecionar tudo** e **Desmarcar tudo** agem em todas as caixas em uma única operação.

Ordenar por. Permite ordenar as linhas de saída em ordem crescente ou decrescente para uma coluna especificada da exibição. A opção "Selecionado" ordena a saída com base em uma ou mais linhas selecionadas por caixas de seleção. Isso pode ser útil, por exemplo, para fazer com que campos de destino denominados "Market_1" a "Market_9" sejam exibidos antes de "Market_10," como a ordenação padrão exibe "Market_10" imediatamente após "Market_1".

Visualização. A visualização padrão (simples) exibe o conjunto básico de colunas de saída. A opção Avançado exibe colunas adicionais para medidas de qualidade do ajuste.

Número de registros usados na estimação. O número de linhas no arquivo de dados de origem original.

Resposta. O campo ou campos identificados como os campos de destino (aqueles com um papel de *Resposta*) no nó Tipo.

Modelo. O tipo de modelo usado para este campo de destino.

Preditores. O número de preditores (aqueles com um papel de *Entrada*) utilizado para este campo de destino.

Valores Discrepantes. Essa coluna será exibida apenas se você tiver solicitado (nos critérios de Modelador Especialista ou ARIMA) a detecção automática de valores discrepantes. O valor mostrado é o número de valores discrepantes detectados.

R-quadrado estacionário. Uma medida que compara a parte estacionária do modelo com um modelo de média simples. Esta medida é preferível ao R-quadrado ordinário quando houver um padrão de tendência ou sazonal. O R-quadrado estacionário pode ser negativo com um intervalo de infinito negativo para 1. Valores negativos significam que o modelo sob consideração é pior do que o modelo de linha de base. Valores positivos significam que o modelo sob consideração é melhor que o modelo de linha de base.

R-quadrado. Medida de qualidade de ajuste de um modelo linear, às vezes chamada de coeficiente de determinação. É a proporção de variação na variável dependente explicada pelo modelo de regressão. O valor varia de 0 a 1. Valores pequenos indicam que o modelo não ajusta bem os dados.

RMSE. Erro Quadrático Médio Raiz. A raiz quadrada do erro quadrático médio. Uma medida de quanto uma série dependente varia a partir de seu nível predito pelo modelo, expressa nas mesmas unidades que a série dependente.

MAPE. Erro Percentual Absoluto Médio. Uma medida do quanto uma série dependente varia de seu nível predito pelo modelo. Ela é independente das unidades utilizadas e pode, portanto, ser utilizada para comparar séries com unidades diferentes.

MAE. Erro médio absoluto. Medidas do quanto a série varia a partir de seu nível predito pelo modelo. O MAE é relatado nas unidades da série original.

MaxAPE. Erro Percentual Absoluto Máximo.O maior erro previsto, expresso como uma porcentagem. Esta medida é útil para imaginar um cenário do pior caso para suas previsões.

MaxAE. Erro Máximo Absoluto. O maior erro previsto, expresso nas mesmas unidades que a série dependente. Assim como o MaxAPE, é útil para imaginar o cenário do pior caso para suas previsões. O erro absoluto máximo e o erro percentual absoluto máximo poderão ocorrer em diferentes pontos da série - por exemplo, quando um erro absoluto de um valor de série grande for um pouco maior que o erro absoluto de um valor de série pequeno. Nesse caso, o erro absoluto máximo ocorrerá no valor de série maior e o erro percentual absoluto máximo ocorrerá no valor de série menor.

BIC normalizado. Critério de Informação Bayesiano Normalizado. Uma medida geral do ajuste geral de um modelo que tenta considerar a complexidade do modelo. É um escore com base no erro quadrático médio e inclui uma penalidade para o número de parâmetros no modelo e no comprimento da série. A penalidade remove a vantagem de modelos com mais parâmetros, tornando a estatística fácil de comparar entre diferentes modelos da mesma série.

Q. A estatística Ljung-Box Q. Um teste da aleatoriedade dos erros residuais nesse modelo.

df. Graus de liberdade. O número de parâmetros do modelo que variam livremente ao estimar uma resposta específica.

Sig. Valor de significância da estatística Ljung-Box. Um valor de significância menor que 0,05 indica que os erros residuais não são aleatórios.

Estatísticas de Sumarização. Esta seção contém várias estatísticas de sumarização para diferentes colunas, incluindo valores médios, mínimos, máximos e percentis.

Parâmetros do Modelo de Série Temporal

A guia Parâmetros lista detalhes dos vários parâmetros que foram utilizados para construir um modelo selecionado.

Exibir parâmetros para o modelo. Selecione o modelo para o qual deseja exibir os detalhes do parâmetro.

Resposta. O nome do campo de destino (com o papel *Resposta*) previsto por este modelo.

Modelo. O tipo de modelo usado para este campo de destino.

Campo (apenas modelos ARIMA). Contém uma entrada para cada uma das variáveis utilizadas no modelo, com o destino primeiro, seguido pelos preditores, se houver.

Transformação. Indica o tipo de transformação que foi especificado, se houver, para este campo antes de o modelo ser construído.

Parâmetro. O parâmetro de modelo para o qual os detalhes a seguir são exibidos:

- **Lag (apenas modelos ARIMA).** Indica os lags, se houver, considerados para este parâmetro no modelo.

- **Estimativa.** A estimativa de parâmetro. Este valor é utilizado no cálculo do valor de previsão e dos intervalos de confiança para o campo de destino.
- **SE.** O erro padrão da estimativa de parâmetro.
- **t.** O valor da estimativa de parâmetro dividido pelo erro padrão.
- **Sig.** O nível de significância para a estimativa de parâmetro. Os valores acima de 0,05 são considerados como não estatisticamente significativos.

Resíduos de Modelo de Séries Temporais

A guia Resíduos mostra a função de autocorrelação (FAC) e a função de autocorrelação parcial (FACP) dos resíduos (as diferenças entre valores esperados e reais) para cada modelo construído. Consulte o tópico "Funções de Autocorrelação e de Autocorrelação Parcial" na página 282 para obter mais informações.

Exibir gráfico para o modelo. Selecione o modelo para o qual deseja exibir o FAC e o FACP de resíduo.

Sumarização do Modelo de Série Temporal

A guia Sumarização de um nugget do modelo exibe informações sobre o modelo em si (*Análise*), sobre os campos usados no modelo (*Campos*), sobre as configurações utilizadas ao construir o modelo (*Configurações de Construção*) e sobre o treinamento do modelo (*Sumarização do Treinamento*).

Ao procurar o nó pela primeira vez, os resultados da guia Sumarização são reduzidos. Para ver os resultados de interesse, utilize o controle expansor à esquerda de um item para desdobrá-lo ou clique no botão **Expandir Tudo** para mostrar todos os resultados. Para ocultar os resultados após terminar de visualizá-los, use o controle expansor para reduzir os resultados específicos que deseja ocultar ou clique no botão **Reduzir Tudo** para reduzir todos os resultados.

Análise. Exibe informações sobre o modelo específico.

Campos. Lista os campos utilizados como o destino e as entradas na construção do modelo.

Configurações da Construção. Contém informações sobre as configurações utilizadas na construção do modelo.

Sumarização do Treinamento. Mostra o tipo de modelo, o fluxo utilizado para criá-lo, o usuário que o criou, quando ele foi construído e o tempo decorrido para construir o modelo.

Configurações de Modelo de Série Temporal

A guia Definições permite especificar quais campos extras são criados pela operação de modelagem.

Criar novos campos para cada modelo a ser escorado. Permite especificar os novos campos a serem criados para cada modelo a ser escorado.

- **Calcular limites de confiança superiores e inferiores.** Se marcada, cria novos campos (com os prefixos padrão $TSLCI- e $TSUCI-) para os intervalos de confiança inferiores e superiores, respectivamente, para cada campo de destino, com os totais desses valores.
- **Calcular resíduos de ruído.** Se marcada, cria um novo campo (com prefixo padrão $TSNR-) para os resíduos de modelo para cada campo de destino, com um total desses valores.

Nó de modelagem Spatio-Temporal Prediction

O Spatio-Temporal Prediction (STP) possui muitos aplicativos em potencial, como gerenciamento de energia para prédios ou instalações, análise de desempenho e previsão para engenheiros de serviço mecânico ou planejamento de transportes públicos. Nesses aplicativos, medições, como uso de energia, geralmente demandam espaço e tempo. Algumas questões que podem ser relevantes para o registro dessas medições incluem quais fatores afetarão as observações futuras, o que pode ser feito para por em prática uma mudança desejada ou o que melhor gerencia o sistema? Para abordar essas questões, é

possível utilizar técnicas estatísticas que possam prever valores futuros em locais diferentes, bem como modelar explicitamente fatores ajustáveis para executar análise 'what-if'.

A análise do STP utiliza dados que contêm dados do local, campos de entrada para predição (preditores), um campo de tempo e um campo de destino. Cada local possui inúmeras linhas nos dados que representam os valores de cada preditor em cada momento da medição. Após os dados serem analisados, eles podem ser usados para prever valores de destino em qualquer local dentro dos dados de formato que são usados na análise. A análise também pode prever em que momento no futuro os dados de entrada serão conhecidos.

Nota: O nó STP não suporta os passos Avaliação de Modelo ou Desafiante Campeão no IBM SPSS Collaboration and Deployment Services.

Um fluxo que mostra um exemplo de trabalho utilizando STP, denominado `server_demo.str`, e que faz referência aos arquivos de dados `room_data.csv` e `score_data.csv` está disponível a partir do diretório `Demos` da sua instalação do IBM SPSS Modeler. É possível acessar o diretório `Demos` do grupo de programa IBM SPSS Modeler no menu Iniciar do Windows. O arquivo `server_demo.str` está no diretório `streams`.

Spatio-Temporal Prediction – Opções de Campo

Na guia Campos, você escolhe se deseja utilizar as configurações de papel do campo que já estiverem definidas em nós de envio de dados ou fazer as designações de campo manualmente.

Usar papéis predefinidos
Esta opção utiliza as configurações de papel (apenas destinos e preditores) a partir de um nó Tipo de envio de dados (ou na guia Tipos de um nó de origem de envio de dados).

Usar designações de campo customizadas
Para designar destinos, preditores e outros papéis manualmente nessa tela, selecione esta opção.

Campos
Exibe todos os campos nos dados que podem ser selecionados. Utilize os botões de seta para designar itens manualmente a partir desta lista para as várias caixas à direita da tela. Os ícones indicam os níveis de medição válidos para cada campo.

Nota: O STP requer 1 registro por local, por intervalo de tempo para funcionar corretamente, portanto, estes são campos obrigatórios.

Na parte inferior da área de janela **Campos**, clique no botão **Tudo** para selecionar todos os campos, independentemente do nível de medição, ou clique em um botão de nível de medição individual para selecionar todos os campos com esse nível de medição.

Destino
Selecione um campo como o destino para a predição.

Nota: Os campos podem ser selecionados apenas com um nível de medição de contínuo.

Localização
Selecione o tipo de local a ser usado no modelo.

Nota: É possível selecionar apenas os campos com um nível de medição de geoespacial.

Rótulo de localização
Os dados de forma geralmente incluem um campo que mostra os nomes das variáveis na camada, por exemplo, podem ser nomes de estados ou municípios. Utilize esse campo para associar um nome, ou um rótulo, a um local ao selecionar um campo categórico para rotular o campo **Local** escolhido em sua saída.

Campo de tempo
Selecione os campos de tempo para uso em suas predições.

Nota: É possível selecionar apenas os campos com um nível de medição de contínuo e um tipo de armazenamento de tempo, data, registro de data e hora ou número inteiro.

Preditores (Entradas)
Escolha um ou mais campos como entradas para a predição.

Nota: É possível selecionar apenas os campos com um nível de medição de contínuo.

Spatio-Temporal Prediction - Intervalos de Tempo

Na área de janela Intervalos de Tempo, é possível selecionar as opções para configurar o intervalo de tempo e qualquer agregação necessária ao longo do tempo.

A preparação de dados é necessária para converter os campos de tempo em um índice antes de poder construir um modelo do STP; para que conversão seja possível, o campo de tempo deverá ter um intervalo constante entre os registros. Se seus dados ainda não contiverem essas informações, utilize as opções nessa área de janela para configurar este intervalo antes de poder utilizar o nó de modelagem.

Intervalo de Tempo Selecione o intervalo no qual deseja que o conjunto de dados seja convertido. As opções disponíveis dependem do tipo de armazenamento do campo que é escolhido como o **Campo de Tempo** para o modelo na guia Campos.

- **Períodos** Disponível apenas para campos de tempo de número inteiro; esta é uma série de intervalos, com um intervalo uniforme entre cada medição, que não corresponde a nenhum dos outros intervalos disponíveis.
- **Anos** Disponível apenas para os campos de tempo Registro de Data e Hora.
- **Trimestres** Disponível apenas para os campos de tempo Data ou Registro de Data e Hora. Se escolher essa opção, será solicitado a selecionar o **Mês de início** do primeiro trimestre.
- **Meses** Disponível apenas para os campos de tempo Data ou Registro de Data e Hora.
- **Semanas** Disponível apenas para os campos de tempo Data ou Registro de Data e Hora.
- **Dias** Disponível apenas para os campos de tempo Data ou Registro de Data e Hora.
- **Horas** Disponível apenas para os campos de tempo Hora ou Registro de Data e Hora.
- **Minutos** Disponível apenas para os campos de tempo Hora ou Registro de Data e Hora.
- **Segundos** Disponível apenas para os campos de tempo Hora ou Registro de Data e Hora.

Ao selecionar o **Intervalo de tempo**, será solicitado a preencher campos adicionais. Os campos disponíveis dependem do intervalo de tempo e do tipo de armazenamento. Os campos que puderem ser exibidos são mostrados na lista a seguir.

- **Número de dias por semana**
- **Número de horas em um dia**
- **Semana começa em** O primeiro dia da semana
- **Dia começa às** A hora na qual você considera que um novo dia inicia.
- **Valor do intervalo** É possível escolher uma das seguintes opções: 1, 2, 3, 4, 5, 6, 10, 12, 15, 20 ou 30.
- **Mês de início** O mês no qual o ano fiscal inicia.
- **Período inicial** Se estiver usando **Períodos**, selecione o período inicial.

Dados correspondem às configurações de intervalo de tempo especificadas Se seus dados já contiverem as informações de intervalos de tempo corretas e não precisarem ser convertidos, marque essa caixa de seleção. Ao selecionar essa caixa, os campos na área **Agregação** se tornam indisponíveis.

Agregação

Disponível apenas se desmarcar a caixa de seleção **Dados correspondem às configurações de intervalo de tempo especificadas**; especifique as opções para agregar campos para corresponder ao intervalo

especificado. Por exemplo, se você tiver uma combinação de dados semanais e mensais, será possível agregar, ou "acumular", os valores semanais para atingir um intervalo mensal uniforme. Selecione as configurações padrão a serem utilizadas para agregar tipos de campos diferentes e criar quaisquer configurações customizadas que desejar para quaisquer campos específicos.

- **Contínuo** Configure o método de agregação padrão a ser aplicado a todos os campos contínuos que não forem individualmente especificados. É possível escolher a partir de vários métodos:
  - Soma
  - Média
  - Mínimo
  - Máximo
  - Mediana
  - Primeiro quartil
  - Terceiro quartil

Configurações customizadas para campos especificados Para aplicar uma função de agregação específica a campos individuais, selecione-os nesta tabela e escolha o método de agregação.

- **Campo** Use o botão **Incluir campo** para exibir a caixa de diálogo Selecionar Campos e escolha os campos necessários. Os campos escolhidos são exibidos nessa coluna.
- **Função de agregação** Na lista suspensa, selecione a função de agregação para converter o campo no intervalo de tempo especificado.

Spatio-Temporal Prediction - Opções de Criação Básicas

Utilize as configurações nesta caixa de diálogo para configurar as opções básicas de construção de modelo.

Configurações de modelo

Incluir intercepto

Incluir o intercepto (o termo constante no modelo) pode aumentar a precisão geral da solução. Se você conseguir presumir as passagens de dados por meio da origem, será possível excluir o intercepto.

Máximo de ordem autorregressiva

As ordens autorregressivas especificam quais valores anteriores são utilizados para prever valores atuais. Utilize esta opção para especificar o número de registros anteriores que são utilizados para calcular um novo valor. É possível escolher qualquer número inteiro entre 1 e 5.

Covariância Espacial

Método de estimação

Selecione o método de estimação a ser utilizado; é possível escolher **Paramétrico** ou **Não paramétrico**. Para o método **Paramétrico**, é possível escolher entre um dos três tipos de **Modelo**:

- **Gaussiano**
- **Exponencial**
- **Exponencial** Se selecionar essa opção, também deve-se especificar o nível de **Potência** a ser utilizado. Esse nível pode ser qualquer valor entre 1 e 2, alterado em incrementos de 0,1.

Spatio-Temporal Prediction - Opções de Criação Avançadas

Os usuários com conhecimento detalhado do STP podem utilizar as opções a seguir para fazer um ajuste preciso do processo de construção do modelo.

Porcentagem máxima de valores omissos
Especifique a porcentagem máxima de registros que contêm valores omissos que podem ser incluídos no modelo.

Nível de significância para testar hipóteses na construção de modelo
Especifique o valor do nível de significância a ser utilizado para todos os testes de estimação do modelo do STP, incluindo dois testes de Qualidade do ajuste, testes de Efeito F e testes de Coeficiente T. Esse nível pode ser qualquer valor de 0 a 1, alterado em incrementos de 0,01.

Spatio-Temporal Prediction - Saída

Antes de construir o modelo, utilize as opções nessa área de janela para selecionar a saída que deseja incluir no visualizador de saída.

Informações do modelo

Especificações de modelo
Selecione esta opção para incluir informações de especificação de modelo na saída do modelo.

Sumarização de informações temporais
Selecione esta opção para incluir uma sumarização das informações temporais na saída do modelo.

Avaliação

Qualidade do Modelo
Selecione esta opção para incluir a qualidade do modelo na saída do modelo.

Testes de efeitos no modelo de estrutura média
Selecione esta opção para incluir informações de teste de efeitos na saída do modelo.

Interpretação

Coeficientes do modelo de estrutura média
Selecione esta opção para incluir informações de coeficientes do modelo de estrutura média na saída do modelo.

Coeficientes autorregressivos
Selecione esta opção para incluir informações de coeficientes autorregressivos na saída do modelo.

Testes de deterioração sobre o espaço
Selecione esta opção para incluir informações de teste de covariância espacial ou de redução do espaço na saída do modelo.

Gráfico de parâmetros de modelo de covariância espacial paramétrica
Selecione esta opção para incluir informações do gráfico de parâmetro de modelo de covariância espacial paramétrica na saída do modelo.

Nota: Esta opção estará disponível somente se o método de estimação **Paramétrica** tiver sido selecionado na guia Básicos.

Heat map de correlações
Selecione esta opção para incluir um mapa dos valores de destino na saída do modelo.

Nota: Se houver mais de 500 locais em seu modelo, a saída do mapa não será criada.

Mapa de correlações
Selecione esta opção para incluir um mapa de correlações na saída do modelo.

Nota: Se houver mais de 500 locais em seu modelo, a saída do mapa não será criada.

Clusters de localização

Selecione esta opção para incluir a saída de armazenamento em cluster local na saída do modelo. Apenas a saída que não requer acesso aos dados do mapa é incluída como parte da saída do cluster.

Nota: Esta de saída pode ser criada apenas para um modelo de covariância espacial não paramétrica.

Se escolher essa opção, será possível configurar o seguinte:

- **Limite de similaridade** Selecione o limite no qual clusters de saída devem ser considerados como semelhantes o suficiente para serem mesclados em um único cluster.
- **Número máximo de clusters para exibir** Configure o limite superior para o número de clusters que podem ser incluídos na saída do modelo.

Spatio-Temporal Prediction – Opções de Modelo

Nome do Modelo É possível gerar o nome do modelo automaticamente com base nos campos de destino ou especificar um nome customizado. O nome gerado automaticamente é o nome do campo de destino.

Fator de incerteza (%) O fator de incerteza é um valor de porcentagem que representa o crescimento da incerteza ao realizar previsão no futuro. Os limites superior e inferior da incerteza da previsão aumentarão por essa porcentagem a cada passo no futuro. Configure o fator de incerteza a ser aplicado às suas saídas de modelo; isto configura os limites superior e inferior para os valores preditos.

Nugget do Modelo Spatio-Temporal Prediction

O nugget do modelo Spatio-Temporal Prediction (STP) exibe detalhes do modelo na guia Modelo do Visualizador de Saída. Para obter mais informações sobre como utilizar o visualizador, consulte a seção com o título "Trabalhando com a Saída" no Guia do Usuário do Modelador (ModelerUsersGuide.pdf).

Uma operação de modelagem do Spatio-Temporal Prediction (STP) cria diversos novos campos com o prefixo $STP-, conforme mostrado na tabela a seguir.

Tabela 25. Novos campos criados pela operação de modelagem STP

Nome do campo	Descrição
$STP-<Time>	O campo de tempo criado como parte da construção de modelo. As configurações na área de janela Intervalos de Tempo da guia Opções de Criação determinam o modo como este campo é criado. <Time> é o nome original do campo selecionado como o **Campo Tempo** na guia Campos. **Nota:** Esse campo será criado apenas se você converteu o **Campo de Tempo** original como parte da construção de modelo.
$STP-<Target>	Este campo contém as predições para o valor de resposta. <Target> é o nome do campo **Destino** original para o modelo
$STPVAR-<Target>	Este campo contém os valores de VarianceOfPointPrediction. <Target> é o nome do campo **Destino** original para o modelo
$STPLCI-<Target>	Este campo contém os valores de LowerOfPredictionInterval, ou seja, o limite inferior de confiança. <Target> é o nome do campo **Destino** original para o modelo
$STPUCI-<Target>	Este campo contém os valores de UpperOfPredictionInterval, ou seja, o limite superior de confiança. <Target> é o nome do campo **Destino** original para o modelo

Configurações do Modelo Spatio-Temporal Prediction

Utilize a guia Configurações para controlar o nível de incerteza que você considera aceitável na operação de modelagem.

Fator de incerteza (%) O fator de incerteza é um valor de porcentagem que representa o crescimento da incerteza ao realizar previsão no futuro. Os limites superior e inferior da incerteza da previsão aumentarão por essa porcentagem a cada passo no futuro. Configure o fator de incerteza a ser aplicado às suas saídas de modelo; isto configura os limites superior e inferior para os valores preditos.

Nó TCM

Utilize esse nó para criar um modelo causal temporal (TCM).

Modelos Causais Temporais

A modelagem causal temporal tenta descobrir relacionamentos causais chave nos dados de séries temporais. Na modelagem causal temporal, você especifica um conjunto de séries de resposta e um conjunto de entradas candidatas a essas respostas. Em seguida, o procedimento constrói um modelo de série temporal autorregressivo para cada resposta e inclui somente as entradas que tiverem o relacionamento causal com a resposta. Esta abordagem difere da modelagem tradicional de séries temporais em que se deve especificar explicitamente os preditores para uma série de resposta. Como a modelagem causal temporal geralmente envolve construção de modelos para diversas séries temporais relacionadas, o resultado é referido como um *sistema de modelo*.

No contexto de modelagem causal temporal, o termo *causal* refere-se à causalidade de Granger. Uma série temporal X é considerada "causar no sentido de Granger" outra série temporal Y se a regressão de Y em termos dos valores passados de X e Y resultar em um modelo melhor para Y do que regredir apenas nos valores passados de Y.

Nota: O nó de modelagem causal temporal não suporta os passos Avaliação de Modelo ou Desafiante Campeão no IBM SPSS Collaboration and Deployment Services.

Exemplos

Os tomadores de decisão de negócios usam a modelagem causal temporal para descobrir relacionamentos causais em um conjunto grande de métricas baseadas em tempo que descrevem os negócios. A análise poderá revelar algumas entradas controláveis, que possuem o maior impacto sobre os principais indicadores de desempenho.

Os gerenciadores de sistemas de TI grandes podem usar modelagem causal temporal para detectar anomalias em um grande conjunto de métricas operacionais interrelacionadas. O modelo causal, em seguida, permite ir além da detecção de anomalias e descobrir as causas raiz mais prováveis das anomalias.

Requisitos de campo

Deve haver pelo menos um destino. Por padrão, os campos com um papel predefinido de Nenhum não são utilizados.

Estrutura de dados

A modelagem causal temporal suporta dois tipos de estruturas de dados.

Dados baseados em coluna
Para dados baseados em coluna, cada campo de série temporal contém os dados para uma série

temporal única. Essa estrutura é a estrutura tradicional de dados de séries temporais, conforme utilizado pelo procedimento de modelador de séries temporais.

Dados multidimensionais

Para dados multidimensionais, cada campo de séries temporais contém os dados para diversas séries temporais. Séries temporais separadas, dentro de um determinado campo, são então identificadas por um conjunto de valores de campos categóricos referidos como campos de *dimensão*. Por exemplo, os dados de vendas de dois diferentes canais de vendas (varejo e web) podem ser armazenados em um único campo *sales*. Um campo de dimensão denominado *channel*, com 'valores de varejo' e 'web', identifica os registros que estiverem associados a cada um dos dois canais de vendas.

Séries Temporais para o Modelo

Na guia Campos, utilize as configurações de **Séries Temporais** para especificar as séries a serem incluídas no sistema de modelo.

Selecione a opção para a estrutura de dados que se aplica aos seus dados. Para dados multidimensionais, clique em **Selecionar Dimensões** para especificar os campos de dimensão. A ordem especificada dos campos dimensão define a ordem na qual eles aparecem em todos os diálogos e saída subsequentes. Utilize os botões de seta para cima e para baixo na subdiálogo Selecionar Dimensões para reordenar os campos de dimensão.

Para dados baseados em coluna, o termo *série* tem o mesmo significado que o termo *campo*. Para dados multidimensionais, os campos que contiverem séries temporais são referidos como campos de *métrica*. Uma série temporal para dados multidimensionais é definida por um campo de métrica e por um valor para cada um dos campos da dimensão. As considerações a seguir se aplicam a ambos dados dimensionais e baseados em coluna.

- As séries que forem especificadas como entradas candidatas ou como destino e entrada são consideradas para inclusão no modelo para cada destino. O modelo para cada destino sempre inclui valores em lag do destino em si.
- As séries que forem especificadas como entradas forçadas são sempre incluídas no modelo para cada destino.
- Pelo menos uma série deve ser especificada como um destino ou como destino e entrada.
- Quando **Usar papéis predefinidos** é selecionada, os campos que tiverem um papel de Entrada são configurados como entradas candidatas. Nenhum papel predefinido é mapeado para uma entrada forçada.

Dados multidimensionais

Para dados multidimensionais, especifique os campos de métrica e papéis associados em uma grade, em que cada linha na grade especifica uma única métrica e papel. Por padrão, o sistema de modelo inclui séries de todas as combinações dos campos de dimensão para cada linha na grade. Por exemplo, se houver dimensões para *region* e *brand*, então, por padrão, especificar a métrica *sales* como uma resposta significa que há uma série de respostas de vendas separadas para cada combinação de *region* e *brand*.

Para cada linha na grade, é possível customizar o conjunto de valores para qualquer um dos campos de dimensão clicando no botão de reticências para uma dimensão. Esta ação abre o subdiálogo Selecionar Valores de Dimensão. Também é possível incluir, excluir ou copiar linhas da grade.

A coluna **Contagem de Séries** exibe o número de conjuntos de valores de dimensão que estiverem atualmente especificados para a métrica associada. O valor exibido pode ser maior que o número real de séries (uma série por conjunto). Essa condição ocorre quando algumas das combinações especificadas de valores de dimensão não correspondem à série contida pela métrica associada.

Selecionar Valores de Dimensão: Para dados multidimensionais, é possível customizar a análise ao especificar quais valores de dimensão se aplicam a um campo de métrica específico com um determinado

papel. Por exemplo, se *sales* for um campo de métrica e *channel* for uma dimensão com valores 'retail' e 'web', será possível especificar que as vendas da 'web' representam uma entrada e as vendas 'retail' representam um destino.

Todos os valores
Especifica que todos os valores do campo de dimensão atual são incluídos. Esta opção é a padrão.

Selecionar valores para incluir ou excluir
Utilize esta opção para especificar o conjunto de valores para o campo de dimensão atual. Quando **Incluir** é selecionado para o **Modo**, apenas os valores que forem especificados na lista **Valores Selecionados** são incluídos. Quando **Excluir** é selecionado para o **Modo**, todos os valores diferentes dos valores que são especificados na lista **Valores Selecionados** são incluídos.

É possível filtrar o conjunto de valores a partir do qual escolher. Os valores que atenderem à condição de filtro aparecem na guia **Correspondido** e os valores que não atenderem à condição de filtro aparecem na guia **Não Correspondido** da lista **Valores Desmarcados**. A guia **Tudo** lista todos os valores desmarcados, independentemente de qualquer condição do filtro.

- É possível utilizar asteriscos (*) para indicar caracteres curingas quando especificar um filtro.
- Para limpar o filtro atual, especifique um valor vazio para o termo de procura no diálogo Filtrar Valores Exibidos.

Observações

Na guia Campos, utilize as configurações de **Observações** para especificar os campos que definem as observações.

Observações que são definidas pelas datas/horas

É possível especificar que as observações sejam definidas por um campo de data, de hora ou de registro de data e hora. Além do campo que define as observações, selecione o intervalo de tempo apropriado que descreve as observações. Dependendo do intervalo de tempo especificado, também é possível especificar outras configurações, como o intervalo entre as observações (incremento) ou o número de dias por semana. As considerações a seguir se aplicam ao intervalo de tempo:

- Utilize o valor **Irregular** quando as observações estiverem espaçadas irregularmente no tempo, como o horário no qual uma ordem de vendas é processada. Quando **Irregular** é selecionada, deve-se especificar o intervalo de tempo que é usado para a análise, a partir das configurações de **Intervalo de Tempo** na guia Especificações de Dados.
- Quando as observações representam uma data e hora e o intervalo de tempo é horas, minutos ou segundos, então utilize **Horas por dia**, **Minutos por dia** ou **Segundos por dia**. Quando as observações representam um tempo (duração) sem referência a uma data e o intervalo de tempo é horas, minutos ou segundos, então utilize **Horas (não periódico)**, **Minutos (não periódico)** ou **Segundos (não periódico)**.
- Com base no intervalo de tempo selecionado, o procedimento poderá detectar observações omissas. Detectar observações omissas é necessário porque o procedimento supõe que todas as observações estão igualmente espaçadas no tempo e que não há observações omissas. Por exemplo, se o intervalo de tempo for Dias e a data 2014-10-27 for seguida por 2014-10-29, então há uma observação omissa para 2014-10-28. Os valores são imputados para quaisquer observações omissas. Configurações para manipular valores omissos podem ser especificadas a partir da guia Especificações de Dados.
- O intervalo de tempo especificado permite que o procedimento detecte diversas observações no mesmo intervalo de tempo que precisam ser agregadas e alinhe essas observações em um limite de intervalo, como o primeiro do mês, para assegurar que as observações sejam igualmente espaçadas. Por exemplo, se o intervalo de tempo for Meses, então diversas datas no mesmo mês serão agregadas. Esse tipo de agregação é referido como *agrupamento*. Por padrão,

as observações são sumarizadas quando agrupadas. É possível especificar um método diferente para agrupamento, como a média das observações, a partir das configurações de **Agregação e Distribuição** na guia Especificações de Dados.

- Para alguns intervalos de tempo, configurações adicionais podem definir quebras nos intervalos normais igualmente espaçados. Por exemplo, se o intervalo de tempo for Dias, mas apenas dias da semana forem válidos, será possível especificar que há cinco dias em uma semana, com a semana iniciando na segunda-feira.

Observações que são definidas por períodos ou períodos cíclicos

Observações podem ser definidas por um ou mais campos de número inteiro que representam períodos ou ciclos repetitivos de períodos, até um número arbitrário de níveis de ciclo. Com esta estrutura, é possível descrever uma série de observações que não se enquadram em um dos intervalos de tempo padrão. Por exemplo, um ano fiscal com apenas 10 meses pode ser descrito com um campo de ciclo que representa anos e com um campo de período que representa meses, em que o comprimento de um ciclo é 10.

Os campos que especificam períodos cíclicos definem uma hierarquia de níveis periódicos, em que o nível mais baixo é definido pelo campo **Período**. O próximo nível mais alto é especificado por um campo de ciclo cujo nível é 1, seguido por um campo de ciclo cujo nível é 2, e assim por diante. Os valores de campo de cada nível, exceto o mais alto, devem ser periódicos com relação ao próximo nível mais alto. Os valores do nível mais alto não podem ser periódicos. Por exemplo, no caso do ano fiscal de 10 meses, os meses são periódicos dentro dos anos e os anos não são periódicos.

- O comprimento de um ciclo em um determinado nível é a periodicidade do próximo nível mais baixo. Para o exemplo do ano fiscal, há apenas um nível de ciclo e o comprimento do ciclo é 10, já que o próximo nível mais baixo representa meses e há 10 meses no ano fiscal especificado.
- Especificar o valor inicial para qualquer campo periódico que não iniciar a partir de 1. Essa configuração é necessária para detectar valores omissos. Por exemplo, se um campo periódico iniciar a partir de 2, mas o valor inicial estiver especificado como 1, então o procedimento irá supor que há um valor omisso para o primeiro período em cada ciclo desse campo.

Intervalo de Tempo para Análise

O intervalo de tempo que é utilizado para a análise pode diferir do intervalo de tempo das observações. Por exemplo, se o intervalo de tempo das observações for Dias, será possível escolher Meses para o intervalo de tempo para análise. Os dados são então agregados de dados diários para dados mensais antes de o modelo ser construído. Também é possível optar por distribuir os dados de um intervalo de tempo maior para um intervalo de tempo menor. Por exemplo, se as observações forem trimestrais, será possível distribuir os dados de trimestrais para mensais.

As opções disponíveis para o intervalo de tempo no qual a análise é feita dependem de como as observações são definidas e do intervalo de tempo dessas observações. Em particular, quando as observações são definidas por períodos cíclicos, então apenas a agregação é suportada. Nesse caso, o intervalo de tempo da análise deve ser maior ou igual ao intervalo de tempo das observações.

O intervalo de tempo para a análise é especificado nas configurações de **Intervalo de Tempo** na guia Especificações de Dados. O método pelo qual os dados são agregados ou distribuídos é especificado nas configurações de **Agregação e Distribuição** na guia Especificações de Dados.

Agregação e Distribuição

Funções de agregação

Quando o intervalo de tempo que é utilizado para a análise é maior que o intervalo de tempo das observações, os dados de entrada são agregados. Por exemplo, a agregação é feita quando o intervalo de tempo das observações for Dias e o intervalo de tempo para análise for Meses. As funções de agregação a seguir estão disponíveis: mean, sum, mode, min ou max.

Funções de distribuição

Quando o intervalo de tempo que é utilizado para a análise é menor que o intervalo de tempo das observações, os dados de entrada são distribuídos. Por exemplo, a distribuição é feita quando o intervalo de tempo das observações for Trimestres e o intervalo de tempo para análise for Meses. As funções de distribuição a seguir estão disponíveis: mean ou sum.

Funções de agrupamento

O agrupamento é aplicado quando as observações são definidas por datas/horas e diversas observações ocorrem no mesmo intervalo de tempo. Por exemplo, se o intervalo de tempo das observações for Meses, então diversas datas no mesmo mês serão agrupadas e associadas ao mês em que elas ocorrem. As funções de agrupamento a seguir estão disponíveis: mean, sum, mode, min ou max. O agrupamento é feito sempre quando as observações são definidas por datas/horas e o intervalo de tempo das observações é especificado como Irregular.

Nota: Embora o agrupamento seja uma forma de agregação, ele é feito antes de qualquer manipulação de valores omissos, ao passo que a agregação formal é feita após qualquer manipulação de valores omissos. Quando o intervalo de tempo das observações é especificado como Irregular, a agregação é feita apenas com a função de agrupamento.

Agregar observações entre dias com as do dia anterior

Especifica se as observações com tempos que ultrapassam um limite de um dia são agregados aos valores para o dia anterior. Por exemplo, para observações por hora com um dia de oito horas que começa às 20h, essa configuração especifica se as observações entre 0h e 4h são incluídas nos resultados agregados para o dia anterior. Essa configuração se aplicará apenas se o intervalo de tempo das observações for Horas por dia, Minutos por dia ou Segundos por dia e o intervalo de tempo para análise for Dias.

Configurações customizadas para campos especificados

É possível especificar as funções de agregação, distribuição e agrupamento em uma base de campo por campo. Essas configurações substituem as configurações padrão para as funções de agregação, distribuição e agrupamento.

Valores Omissos

Os valores omissos nos dados de entrada são substituídos por um valor imputado. Os métodos de substituição a seguir estão disponíveis:

Interpolação linear

Substitui valores omissos utilizando um interpolação linear. O último valor válido antes do valor omisso e o primeiro valor válido após o valor omisso são utilizados para a interpolação. Se a primeira ou a última observação na série tiver um valor omisso, então os dois valores não omissos mais próximos no início ou no término da série serão utilizados.

Média de série

Substitui valores omissos pela média para toda a série.

Média de pontos próximos

Substitui valores omissos pela média de valores circundantes válidos. O span de pontos próximos é o número de valores válidos antes e após o valor omisso que são utilizados para calcular a média.

Mediana de pontos próximos

Substitui valores omissos pela média de valores válidos circundantes. O span de pontos próximos é o número de valores válidos antes e após o valor omisso que são utilizados para calcular a mediana.

Tendência linear

Esta opção utiliza todas as observações não omissas na série para ajustar um modelo de regressão linear simples, que é então utilizado para imputar os valores omissos.

Outras configurações:

Porcentagem máxima de valores omissos (%)
Especifica a porcentagem máxima de valores omissos que são permitidos para qualquer série. As séries com mais valores omissos que o máximo especificado são excluídas da análise.

Opções de Dados Gerais

Número máximo de valores distintos por campo de dimensão
Esta configuração se aplica a dados multidimensionais e especifica o número máximo de valores distintos que são permitidos para qualquer campo de dimensão. Por padrão, este limite é configurado como 10000, mas poderá ser aumentado para um número arbitrariamente grande.

Opções de Criação Gerais

Largura de intervalo de confiança (%)
Esta configuração controla os intervalos de confiança para ambas as previsões e parâmetros do modelo. É possível especificar quaisquer valores positivos menores que 100. Por padrão, um intervalo de confiança de 95% é utilizado.

Número máximo de entradas por cada resposta
Essa configuração especifica o número máximo de entradas que são permitidas no modelo para cada resposta. É possível especificar um número inteiro no intervalo de 1 a 20. O modelo para cada resposta sempre inclui valores de lag de si mesmo, portanto, configurar esse valor para 1 especifica que a única entrada é a própria resposta.

Tolerância do modelo
Essa configuração controla o processo iterativo que é utilizado para determinar o melhor conjunto de entradas para cada resposta. É possível especificar qualquer valor que seja maior que zero. O padrão é 0,001.

Limite de valor discrepante (%)
Uma observação será sinalizada como um valor discrepante se a probabilidade de que o modelo seja um valor discrepante, conforme calculado a partir do mesmo, exceder esse limite. É possível especificar um valor no intervalo de 50 a 100.

Número de Lags para Cada Entrada
Esta configuração especifica o número de termos de lag para cada entrada no modelo para cada resposta. Por padrão, o número de termos de lag é determinado automaticamente a partir do intervalo de tempo que é utilizado para a análise. Por exemplo, se o intervalo de tempo for meses (com um incremento de um mês), então o número de lags será 12. Opcionalmente, é possível especificar explicitamente o número de lags. O valor especificado deve ser um número inteiro no intervalo de 1 a 20.

Continuar a estimação usando modelos existentes
Se você já gerou um modelo causal temporal, selecione esta opção para reutilizar as configurações de critérios que forem especificadas para esse modelo, ao invés de construir um novo modelo. Dessa forma, é possível economizar tempo ao reestimar e produzir uma nova previsão que seja baseada nas mesmas configurações de modelo como antes, porém utilizando dados mais recentes.

Série para Exibir

Estas opções especificam as séries (destinos ou entradas) para as quais a saída é exibida. O conteúdo da saída para a série especificada é determinado pelas configurações de **Opções de Saída**.

Exibir destinos associados com modelos de melhor ajuste
Por padrão, a saída é exibida para os destinos que estiverem associados aos 10 modelos de melhor ajuste, conforme determinado pelo valor de R-quadrado. É possível especificar um número fixo diferente de modelos de melhor ajuste ou especificar uma porcentagem de modelos de melhor ajuste. Também é possível escolher a partir da Qualidade do ajuste a seguir de medidas de ajuste:

R-quadrado
Medida de qualidade de ajuste de um modelo linear, às vezes chamada de coeficiente de determinação. É a proporção de variação na variável de destino explicada pelo modelo. O valor varia de 0 a 1. Valores pequenos indicam que o modelo não ajusta bem os dados.

Porcentagem de erro quadrático médio raiz
A medida do quanto os valores preditos pelo modelo diferem dos valores observados da série. Ela é independente das unidades que são utilizadas e pode, portanto, ser utilizada para comparar séries com unidades diferentes.

Erro quadrático médio raiz
A raiz quadrada do erro quadrático médio. Uma medida de quanto uma série dependente varia a partir de seu nível predito pelo modelo, expressa nas mesmas unidades que a série dependente.

BIC Critério de Informação Bayesiano. Uma medida para selecionar e comparar modelos com base no log da verossimilhança -2 reduzido. Valores menores indicam melhores modelos. O BIC também "penaliza" modelos sobreparametrizados (por exemplo, modelos complexos com um número grande de entradas), mas é mais rígido do que o AIC.

AIC Akaike Information Criterion. Uma medida para selecionar e comparar modelos com base no log da verossimilhança -2 reduzido. Valores menores indicam melhores modelos. O AIC "penaliza" modelos sobreparametrizados (modelos complexos com um número grande de entradas, por exemplo).

Especificar série individual

É possível especificar séries individuais para as quais você deseja gerar.

- Para dados baseados em coluna, especifique os campos que contêm a série desejada. A ordem dos campos especificados define a ordem na qual eles aparecem na saída.
- Para dados multidimensionais, especifique uma série determinada ao incluir uma entrada na grade para o campo de métrica que contém a série. Em seguida, especifique os valores dos campos de dimensão que definem a série.
  - É possível inserir o valor para cada campo de dimensão diretamente na grade ou selecionar a partir da lista de valores da dimensão disponíveis. Para selecionar a partir da lista de valores da dimensão disponíveis, clique no botão de reticências na célula para a dimensão desejada. Esta ação abre o subdiálogo Selecionar Valor da Dimensão.
  - É possível procurar a lista de valores da dimensão, no subdiálogo Selecionar Valor da Dimensão, clicando no ícone de binóculo e especificando um termo de procura. Espaços são tratados como parte do termo de procura. Asteriscos (*) no termo de procura não indicam caracteres curinga.
  - A ordem das séries na grade define a ordem na qual elas aparecem na saída.

Tanto para dados baseados em coluna quanto para dados multidimensionais, a saída é limitada a 30 séries. Este limite inclui séries individuais (entradas ou destinos) que você especificar e destinos que estiverem associados aos modelos de melhor ajuste. As séries especificadas individualmente têm precedência sobre os destinos que estiverem associados aos modelos de melhor ajuste.

Opções de Saída

Estas opções especificam o conteúdo da saída. As opções no grupo **Saída para destinos** geram uma saída para os destinos que estiverem associados aos modelos de melhor ajuste nas configurações de **Série para Exibir**. As opções no grupo **Saída para série** geram uma saída para as séries individuais que estiverem especificadas nas configurações de **Série para Exibir**.

Sistema de modelo global

Exibe uma representação gráfica das relações causais entre séries no sistema de modelo. Tabelas de estatísticas de ajuste de modelo e de valores discrepantes dos destinos exibidos são incluídas como parte do item de saída. Quando essa opção é selecionada no grupo **Saída para série**, um item de saída separado é criado para cada série individual que estiver especificada nas configurações de **Série para Exibir**.

As relações causais entre as séries têm um nível de significância associado, no qual um nível de significância menor indica uma conexão mais significativa. É possível optar por ocultar relações com um nível de significância que seja maior que um valor especificado.

Valores discrepantes e estatísticas de ajuste do modelo
Tabelas de estatísticas de ajuste de modelo e de valores discrepantes para as séries de destino que estiverem selecionadas para exibição. Essas tabelas contêm as mesmas informações que as tabelas na visualização Sistema de Modelo Geral. Essas tabelas suportam todas as variáveis padrão para editar e definir tabelas como dinâmicas.

Efeitos do modelo e parâmetros do modelo
Tabelas de testes de efeitos do modelo e de parâmetros de modelo para as séries de destino que estiverem selecionadas para exibição. Os testes de efeitos do modelo incluem a estatística de F e o valor de significância associado a cada entrada incluída no modelo.

Diagrama de impacto

Exibe uma representação gráfica das relações causais entre uma série de interesse e outras séries que ela afeta ou vice-versa. As séries que afetam a série de interesse são referidas como *causas*. Selecionar **Efeitos** gera um diagrama de impacto que é inicializado para exibir os efeitos. Selecionar **Causas** gera um diagrama de impacto que é inicializado para exibir as causas. Selecionar **Ambos causas e efeitos** gera dois diagramas de impacto separados, um que é inicializado para causas e outro que é inicializado para efeitos. Em seguida, é possível alternar interativamente entre as causas e efeitos no item de saída que exibe o diagrama de impacto.

É possível especificar o número de níveis de causas e efeitos para exibir, em que o primeiro nível é apenas a série de interesse. Cada nível adicional mostra causas ou efeitos mais indiretos da série de interesse. Por exemplo, o terceiro nível na exibição de efeitos consiste nas séries que contêm a série no segundo nível como uma entrada direta. As séries no terceiro nível são, então, afetadas indiretamente pela série de interesse, desde que a série de interesse seja uma entrada direta para as séries no segundo nível.

Gráfico de série
Representa valores observados e preditos para as séries de destino que estiverem selecionadas para exibição. Quando as previsões são solicitadas, o gráfico também mostra os valores previstos e os intervalos de confiança para as previsões.

Gráfico de resíduos
Representa os resíduos do modelo para as séries de destino que estiverem selecionadas para exibição.

Entradas principais
Representa cada destino exibido, ao longo do tempo, com as 3 principais entradas para o destino. As entradas principais são as entradas com o valor de significância mais baixo. Para acomodar escalas diferentes para as entradas e o destino, o eixo y representa o escore z para cada série.

Tabela de previsão
Tabelas de valores previstos e de intervalos de confiança das previsões das series de destino que estiverem selecionadas para exibição.

Análise de causa raiz de valor discrepante
Determina qual série tem maior probabilidade de ser a causa de cada valor discrepante em uma série de interesse. A análise de causa raiz de valor discrepante é feita para cada série de destino que estiver incluída na lista de séries individuais nas configurações de **Série para Exibir**.

Saída

Tabela e gráfico de valores discrepantes interativos
Tabela e gráfico de valores discrepantes e as causas raiz desses valores discrepantes para cada série de interesse. A tabela contém uma única linha para cada valor discrepante. O gráfico é um diagrama de impacto. Selecionar uma linha na tabela destaca o caminho, no diagrama de impacto, da série de interesse para a série que mais provavelmente causa o valor discrepante associado.

Tabela dinâmica de valores discrepantes
Tabela de valores discrepantes e as causas raiz desses valores para cada série de interesse. Esta tabela contém as mesmas informações que a tabela na exibição interativa. Esta tabela suporta todas as variáveis padrão para editar e definir tabelas como dinâmicas.

Níveis Causais
É possível especificar o número de níveis a serem incluídos na procura das causas raiz. O conceito de níveis usado aqui é o mesmo descrito para diagramas de impacto.

Ajuste do modelo por todos os modelos
Histograma de ajuste do modelo para todos os modelos e para estatísticas de ajuste selecionadas. As estatísticas de ajuste a seguir estão disponíveis:

R-quadrado
Medida de qualidade de ajuste de um modelo linear, às vezes chamada de coeficiente de determinação. É a proporção de variação na variável de destino explicada pelo modelo. O valor varia de 0 a 1. Valores pequenos indicam que o modelo não ajusta bem os dados.

Porcentagem de erro quadrático médio raiz
A medida do quanto os valores preditos pelo modelo diferem dos valores observados da série. Ela é independente das unidades que são utilizadas e pode, portanto, ser utilizada para comparar séries com unidades diferentes.

Erro quadrático médio raiz
A raiz quadrada do erro quadrático médio. Uma medida de quanto uma série dependente varia a partir de seu nível predito pelo modelo, expressa nas mesmas unidades que a série dependente.

BIC Critério de Informação Bayesiano. Uma medida para selecionar e comparar modelos com base no log da verossimilhança -2 reduzido. Valores menores indicam melhores modelos. O BIC também "penaliza" modelos sobreparametrizados (por exemplo, modelos complexos com um número grande de entradas), mas é mais rígido do que o AIC.

AIC Akaike Information Criterion. Uma medida para selecionar e comparar modelos com base no log da verossimilhança -2 reduzido. Valores menores indicam melhores modelos. O AIC "penaliza" modelos sobreparametrizados (modelos complexos com um número grande de entradas, por exemplo).

Valores Discrepantes ao longo do tempo
Gráfico de barras do número de valores discrepantes, em todos os destinos, para cada intervalo de tempo no período de estimação.

Transformações de série
Tabela de quaisquer transformações que foram aplicadas à série no sistema de modelo. As transformações possíveis são imputação, agregação e distribuição de valor omisso.

Período de Estimação

Por padrão, o período de estimação começa no momento da primeira observação e termina no momento da observação mais recente em todas as séries.

Por horários de início e de encerramento
É possível especificar o início e o término do período de estimação ou especificar apenas o início ou o término. Se você omitir o início ou o término do período de estimação, o valor padrão será utilizado.

- Se as observações forem definidas por um campo de data/hora, então insira valores para o início e o término no mesmo formato que é utilizado para o campo de data/hora.
- Para as observações que forem definidas por períodos cíclicos, especifique um valor para cada um dos campos de períodos cíclicos. Cada campo é exibido em uma coluna separada.

Por intervalos de tempo mais recentes ou mais antigos

Define o período de estimação como um número especificado de intervalos de tempo que começam no primeiro intervalo de tempo ou terminam no último intervalo de tempo nos dados, com um offset opcional. Neste contexto, o intervalo de tempo refere-se ao intervalo de tempo da análise. Por exemplo, suponha que as observações sejam mensais, mas o intervalo de tempo da análise seja trimestral. Especificar **Mais recente** e um valor de 24 para o **Número de intervalos de tempo** significa os últimos 24 trimestres.

Opcionalmente, é possível excluir um número especificado de intervalos de tempo. Por exemplo, especificar os últimos 24 intervalos de tempo e 1 para o número para excluir significa que o período de estimação consiste em 24 intervalos que precedem o último.

Opções de Modelo

Nome do modelo

É possível especificar um nome customizado para o modelo ou aceitar o nome gerado automaticamente, que é *TCM*.

Previsão

A opção **Estender registros para o futuro** configura o número de intervalos de tempo para prever além do término do período de estimação. O intervalo de tempo nesse caso é o intervalo de tempo da análise, que é especificado na guia Especificações de Dados. Quando previsões são solicitadas, modelos autorregressivos são construídos automaticamente para quaisquer séries de entrada que não sejam também respostas. Esses modelos são, então, utilizados para gerar valores para essas séries de entrada no período de previsão. Não há limite máximo para essa configuração.

Resultado Interativo

A saída de modelagem causal temporal inclui um número de objetos de saída interativos. As variáveis interativas estão disponíveis ao ativar (clicar duas vezes) o objeto no Visualizador de Saída.

Sistema de modelo global

Exibe as relações causais entre séries no sistema de modelo. Todas as linhas que conectam uma resposta específica às suas entradas possuem a mesma cor. A espessura da linha indica a significância da conexão causal, em que as linhas mais espessas representam uma conexão mais significativa. As entradas que também não forem respostas são indicadas com um quadrado preto.

- É possível exibir as relações para os principais modelos, para uma série especificada, para todas as séries ou para modelos sem entradas. Os modelos principais são os modelos que atendem aos critérios que foram especificados para modelos de melhor ajuste nas configurações de **Séries para Exibir**.
- É possível gerar diagramas de impacto para uma ou mais séries ao selecionar os nomes das séries no gráfico, clicar com o botão direito e, em seguida, escolher **Criar Diagrama de Impacto** no menu de contexto.
- É possível optar por ocultar relações causais que possuem um nível de significância que seja maior que um valor especificado. Níveis de significância menores indicam uma relação causal mais significativa.
- É possível exibir as relações de uma determinada série ao selecionar o nome da série no gráfico, clicar com o botão direito e, em seguida, escolher **Destacar relações para série** no menu de contexto.

Diagrama de impacto
Exibe uma representação gráfica das relações causais entre uma série de interesse e outras séries que ela afeta ou vice-versa. As séries que afetam a série de interesse são referidas como *causas*.

- É possível alterar a série de interesse ao especificar o nome da série desejada. Clicar duas vezes em qualquer nó no diagrama de impacto altera a série de interesse para a série associada a esse nó.
- É possível alternar a exibição entre as causas e efeitos e também alterar o número de níveis de causas e efeitos para exibir.
- Clicar uma vez em qualquer nó abre um diagrama de sequência detalhado para a série que está associada ao nó.

Análise de causa raiz de valor discrepante
Determina qual série tem maior probabilidade de ser a causa de cada valor discrepante em uma série de interesse.

- É possível exibir a causa raiz para qualquer valor discrepante ao selecionar a linha para o valor discrepante na tabela Valores Discrepantes. Também é possível exibir a causa raiz ao clicar no ícone para o valor discrepante no gráfico de sequência.
- Clicar uma vez em qualquer nó abre um diagrama de sequência detalhado para a série que está associada ao nó.

Qualidade do modelo global
Histograma de ajuste do modelo para todos os modelos, para uma estatística de ajuste específica. Clicar em uma barra no gráfico de barras filtra o gráfico de pontos para que ele exiba apenas os modelos que estiverem associados à barra selecionada. É possível localizar o modelo para uma série de respostas específica no gráfico de pontos ao especificar o nome da série.

Distribuição de valor discrepante
Gráfico de barras do número de valores discrepantes, em todos os destinos, para cada intervalo de tempo no período de estimação. Clicar em uma barra no gráfico de barras filtra o gráfico de pontos para que ele exiba apenas os valores discrepantes que estiverem associados à barra selecionada.

Nugget do Modelo TCM

Uma operação de modelagem TCM cria diversos novos campos com o prefixo $TCM-, conforme mostrado na tabela a seguir.

Tabela 26. Novos campos criados pela operação de modelagem do TCM

Nome do campo	Descrição
$TCM-*colname*	O valor previsto pelo modelo para cada série de resposta.
$TCMLCI-*colname*	Os intervalos de confiança inferiores para cada série prevista.
$TSUCI-*colname*	Os intervalos de confiança superiores para cada série prevista.
$TCMResidual-*colname*	O valor residual de ruído para cada coluna dos dados de modelo gerados.

Configurações do Nugget do Modelo TCM

A guia Configurações fornece opções adicionais para o nugget do modelo TCM.

Previsão

A opção **Estender registros para o futuro** configura o número de intervalos de tempo para prever além do término do período de estimação. O intervalo de tempo nesse caso é o intervalo de tempo da análise, que é especificado na guia Especificações de Dados do nó TCM. Quando previsões são solicitadas, modelos autorregressivos são construídos automaticamente para quaisquer séries de entrada que não sejam também respostas. Esses modelos são, então, utilizados para gerar valores para essas séries de entrada no período de previsão.

Tornar disponível para escoragem

Criar novos campos para cada modelo a ser escorado. Permite especificar os novos campos a serem criados para cada modelo a ser escorado.

- **Resíduos de Ruído.** Se selecionada, cria um novo campo (com o prefixo $TCM- padrão) para os resíduos do modelo para cada campo de destino, com um total desses valores.
- **Limites de confiança superiores e inferiores.** Se selecionada, cria novos campos (com o prefixo $TCM- padrão) para intervalos de confiança inferiores e superiores, respectivamente, para cada campo de destino, com os totais desses valores.

Respostas incluídas para escoragem. Selecione as respostas disponíveis a serem incluídas na escoragem do modelo.

Cenários de modelo causal temporal

O procedimento de Cenários de Modelo Causal Temporal executa cenários definidos pelo usuário para um sistema de modelo causal temporal, com os dados do conjunto de dados ativo. Um *cenário* é definido por uma série temporal, que é referida como *série raiz*, e por um conjunto de valores definidos pelo usuário para essa série ao longo de um intervalo de tempo especificado. Os valores especificados são, então, utilizados para gerar predições para as séries temporais que forem afetadas pela série raiz. O procedimento requer um arquivo do sistema de modelo que foi criado pelo procedimento Modelagem Causal Temporal. Supõe-se que o conjunto de dados ativo sejam os mesmos dados que foram utilizados para criar o arquivo do sistema de modelo.

Exemplo

Utilizando o procedimento Modelagem Causal Temporal, um tomador de decisões de negócios descobriu uma métrica chave que afeta um número de indicadores de desempenho importantes. Como a métrica é controlável, o tomador de decisões quer investigar o efeito dos vários conjuntos de valores da métrica para o próximo trimestre. A investigação é facilmente realizada ao carregar o arquivo do sistema de modelo no procedimento de Cenários de Modelo Causal Temporal e especificar os conjuntos de valores para a métrica chave.

Definindo o Período de Cenário

O período de cenário é o período durante o qual você especifica os valores que são utilizados para executar seus cenários. Ele pode iniciar antes ou após o término do período de estimação. É possível, opcionalmente, especificar para prever além do término do período do cenário. Por padrão, as predições são geradas no término do período do cenário. Todos os cenários utilizam o mesmo período de cenário e as mesmas especificações para até quando prever no futuro.

Nota: As predições iniciam no primeiro período de tempo após o início do período de cenário. Por exemplo, se o período de cenário iniciar em 2014-11-01 e o intervalo de tempo for meses, então a primeira predição será para 2014-12-01.

Especificar por início, de término e de predição através de tempos

- Se as observações forem definidas por um campo de data/hora, então insira valores para o início, término e predição no mesmo formato que é utilizado para o campo de data/hora. Os valores para campos de data/hora são alinhados com o início do intervalo de tempo associado. Por exemplo, se o intervalo de tempo da análise for meses, o valor 10/10/2014 será ajustado para 01/10/2014, que é o início do mês.
- Para as observações que forem definidas por períodos cíclicos, especifique um valor para cada um dos campos de períodos cíclicos. Cada campo é exibido em uma coluna separada.

Especificar por intervalos de tempo relativos ao término do período de estimação

Define o início e o término em termos do número de intervalos de tempo relativo ao término do período de estimação, onde o intervalo de tempo é o intervalo de tempo da análise. O término do

período de estimação é definido como um intervalo de tempo de 0. Intervalos de tempo anteriores ao término do período de estimação possuem valores negativos e intervalos posteriores ao término do período de estimação possuem valores positivos. Também é possível especificar quantos intervalos prever além do término do período do cenário. O padrão é 0.

Por exemplo, suponha que o intervalo de tempo da análise seja meses e que você especifique 1 para o intervalo inicial, 3 para o intervalo final e 1 para até quando prever além disso. O período de cenário será, então, os 3 meses após o término do período de estimação. As predições são geradas para o segundo e terceiro meses do período de cenário e para mais 1 mês após o término do período de cenário.

Incluindo Cenários e Grupos de Cenários

A guia Cenários especifica os cenários que deverão ser executados. Para definir cenários, deve-se primeiro definir o período de cenário clicando em **Definir Período de Cenário**. Os cenários e grupos de cenários (aplicam-se apenas a dados multidimensionais) são criados clicando no botão **Incluir Cenário** ou **Incluir Grupo de Cenários** associado. Ao selecionar um cenário ou um grupo de cenários na grade associada, é possível editá-lo, fazer uma cópia dele ou excluí-lo.

Dados baseados em coluna

A coluna **Campo raiz** na grade especifica o campo de séries temporais cujos valores são substituídos pelos valores de cenário. A coluna **Valores de cenário** exibe os valores de cenário especificados na ordem do mais antigo ao mais recente. Se os valores de cenário forem definidos por uma expressão, então a coluna exibirá a expressão.

Dados multidimensionais

Cenários individuais

Cada linha na grade Cenários Individuais especifica uma série temporal cujos valores são substituídos pelos valores de cenário especificados. A série é definida pela combinação do campo que é especificado na coluna **Métrica Raiz** e do valor especificado para cada um dos campos de dimensão. O conteúdo da coluna **Valores de cenário** é o mesmo que para dados baseados em coluna.

Grupos de cenários

Um *grupo de cenários* define um conjunto de cenários que se baseiam em um campo de métrica raiz único e em diversos conjuntos de valores de dimensão. Cada conjunto de valores de dimensão (um valor por campo de dimensão), para o campo de métrica especificado, define uma série temporal. Um cenário individual é, então, gerado para cada uma dessas séries temporais, cujos valores são, em seguida, substituídos pelos valores de cenário. Os valores de cenário para um grupo de cenários são especificados por uma expressão, que é, então, aplicada a cada série temporal no grupo.

A coluna **Contagem de Séries** exibe o número de conjuntos de valores de dimensão que estiverem associados a um grupo de cenários. O valor exibido pode ser maior que o número real de séries temporais que estiverem associadas ao grupo de cenários (uma série por conjunto). Essa condição ocorre quando algumas das combinações especificadas de valores de dimensão não correspondem à série contida pela métrica raiz do grupo.

Como exemplo de um grupo de cenários, considere um campo de métrica *advertising* e dois campos de dimensão *region* e *brand*. É possível definir um grupo de cenário com base em *advertising* como a métrica raiz e que inclua todas as combinações de *region* e *brand*. Em seguida, é possível especificar `advertising*1.2` como a expressão para investigar o efeito de aumentar *advertising* em 20 por cento para cada série temporal que estiver associada ao campo *advertising*. Se houver 4 valores de *region* e 2 valores de *brand*, então haverá 8 séries temporais e, com isso, 8 cenários definidos pelo grupo.

Definição de Cenário: As configurações para definir um cenário dependem se seus dados são multidimensionais ou baseados em colunas.

Série raiz

Especifica a série raiz para o cenário. Cada cenário baseia-se em uma série raiz única. Para dados baseados em coluna, selecione o campo que define a série raiz. Para dados multidimensionais, especifique a série raiz ao incluir uma entrada na grade para o campo de métrica que contém a série. Em seguida, especifique os valores dos campos de dimensão que definem a série raiz. O seguinte se aplica para especificar os valores de dimensão:

- É possível inserir o valor para cada campo de dimensão diretamente na grade ou selecionar a partir da lista de valores da dimensão disponíveis. Para selecionar a partir da lista de valores da dimensão disponíveis, clique no botão de reticências na célula para a dimensão desejada. Esta ação abre o subdiálogo Selecionar Valor da Dimensão.
- É possível procurar a lista de valores da dimensão, no subdiálogo Selecionar Valor da Dimensão, clicando no ícone de binóculo e especificando um termo de procura. Espaços são tratados como parte do termo de procura. Asteriscos (*) no termo de procura não indicam caracteres curinga.

Especificar respostas afetadas

Use essa opção quando você souber quais respostas específicas são afetadas pela série raiz e desejar investigar os efeitos apenas nessas respostas. Por padrão, as respostas que forem afetadas pela série raiz são determinadas automaticamente. É possível especificar a amplitude das séries que são afetadas pelo cenário com as configurações na guia Opções.

Para dados baseados em coluna, selecione as respostas desejadas. Para dados multidimensionais, especifique a série de resposta ao incluir uma entrada na grade para o campo de métrica de resposta que contém a série. Por padrão, todas as séries que estiverem contidas no campo de métrica especificado são incluídas. É possível customizar o conjunto de séries incluídas ao customizar os valores incluídos em um ou mais campos de dimensão. Para customizar os valores de dimensão que estão incluídos, clique no botão de reticências da dimensão desejada. Esta ação abre o diálogo Selecionar Valores de Dimensão.

A coluna **Contagem de Séries** (para dados multidimensionais) exibe o número de conjuntos de valores de dimensão que estão atualmente especificados para a métrica de resposta associada. O valor exibido pode ser maior que o número real de séries de resposta afetadas (uma série por conjunto). Essa condição ocorre quando algumas das combinações de valores de dimensão especificadas não correspondem à série contida pela métrica de resposta associada.

ID do Cenário

Cada cenário deve ter um identificador exclusivo. O identificador é exibido na saída que está associada ao cenário. Não há nenhuma restrição, a não ser exclusividade, sobre o valor do identificador.

Especificar valores de cenário para a série raiz

Utilize essa opção para especificar valores explícitos para a série raiz no período de cenário. Deve-se especificar um valor numérico para cada intervalo de tempo que estiver listado na grade. É possível obter os valores das séries raiz (reais ou previstos) para cada intervalo no período de cenário clicando em **Leitura**, **Previsão** ou **Leitura\Previsão**.

Especificar expressão para valores de cenário para a série raiz

É possível definir uma expressão para calcular os valores das séries raiz no período de cenário. É possível inserir a expressão diretamente ou clicar no botão de calculadora e criar a expressão no Construtor de Expressão de Valores de Cenário.

- A expressão pode conter qualquer resposta ou entrada no sistema de modelo.
- Quando o período de cenário ultrapassa os dados existentes, a expressão é aplicada aos valores previstos dos campos na expressão.
- Para dados multidimensionais, cada campo na expressão especifica uma série temporal que é definida pelo campo e os valores de dimensão que foram especificados para a métrica raiz. São essas séries temporais que são utilizadas para avaliar a expressão.

Como exemplo, suponha que o campo raiz seja *advertising* e a expressão seja `advertising*1.2`. Os valores para *advertising* que são utilizados no cenário representam um aumento de 20 por cento sobre os valores existentes.

Nota: Os cenários são criados clicando em **Incluir Cenário** na guia Cenários.

Selecionar Valores de Dimensão: Para dados multidimensionais, é possível customizar os valores de dimensão que definem as respostas que são afetadas por um cenário ou por um grupo de cenários. Também é possível customizar os valores de dimensão que definem o conjunto de séries raiz para um grupo de cenários.

Todos os valores
Especifica que todos os valores do campo de dimensão atual são incluídos. Esta opção é a padrão.

Selecionar valores
Utilize esta opção para especificar o conjunto de valores para o campo de dimensão atual. É possível filtrar o conjunto de valores a partir do qual escolher. Os valores que atenderem à condição de filtro aparecem na guia **Correspondido** e os valores que não atenderem à condição de filtro aparecem na guia **Não Correspondido** da lista **Valores Desmarcados**. A guia **Tudo** lista todos os valores desmarcados, independentemente de qualquer condição do filtro.

- É possível utilizar asteriscos (*) para indicar caracteres curingas quando especificar um filtro.
- Para limpar o filtro atual, especifique um valor vazio para o termo de procura no diálogo Filtrar Valores Exibidos.

Para customizar valores de dimensão para respostas afetadas:

1. No diálogo Definição de Cenário ou Definição de Grupo de Cenários, selecione a métrica de resposta para a qual deseja customizar valores de dimensão.
2. Clique no botão de reticências na coluna da dimensão que deseja customizar.

Para customizar valores de dimensão para a série raiz de um grupo de cenários:

1. No diálogo Definição de Grupo de Cenários, clique no botão de reticências (na grade da série raiz) da dimensão que deseja customizar.

Definição de Grupo de Cenários:

Série raiz
Especifica o conjunto de série raiz para o grupo de cenários. Um cenário individual é gerado para cada série temporal no conjunto. Especifique a série raiz ao incluir uma entrada na grade para o campo de métrica que contém a série desejada. Em seguida, especifique os valores dos campos de dimensão que definem o conjunto. Por padrão, todas as séries que estiverem contidas no campo de métrica raiz especificado são incluídas. É possível customizar o conjunto de séries incluídas ao customizar os valores incluídos em um ou mais campos de dimensão. Para customizar os valores de dimensão que estão incluídos, clique no botão de reticências para uma dimensão. Esta ação abre o diálogo Selecionar Valores de Dimensão.

A coluna **Contagem de Séries** exibe o número de conjuntos de valores de dimensão que estão atualmente incluídos na métrica raiz associada. O valor exibido pode ser maior que o número real de séries raiz para o grupo de cenários (uma série por conjunto). Essa condição ocorre quando algumas das combinações especificadas de valores de dimensão não correspondem à série contida pela métrica raiz.

Especificar série de resposta afetada
Use essa opção quando você souber quais respostas específicas são afetadas pelo conjunto de séries raiz e desejar investigar os efeitos apenas nessas respostas. Por padrão, as respostas que

forem afetadas por cada série raiz são determinadas automaticamente. É possível especificar a amplitude das séries que são afetadas por cada cenário individual com as configurações na guia Opções.

Especifique a série de resposta ao incluir uma entrada na grade para o campo de métrica que contém a série.Por padrão, todas as séries que estiverem contidas no campo de métrica especificado são incluídas. É possível customizar o conjunto de séries incluídas ao customizar os valores incluídos em um ou mais campos de dimensão. Para customizar os valores de dimensão que estão incluídos, clique no botão de reticências da dimensão desejada. Esta ação abre o diálogo Selecionar Valores de Dimensão.

A coluna **Contagem de Séries** exibe o número de conjuntos de valores de dimensão que estão atualmente especificados para a métrica de resposta associada. O valor exibido pode ser maior que o número real de séries de resposta afetadas (uma série por conjunto). Essa condição ocorre quando algumas das combinações de valores de dimensão especificadas não correspondem à série contida pela métrica de resposta associada.

Prefixo de ID de cenário
Cada grupo de cenário deve ter um prefixo exclusivo. O prefixo é utilizado para construir um identificador que é exibido na saída que está associada a cada cenário individual no grupo de cenários. O identificador para um cenário individual é o prefixo, seguido por um sublinhado e depois pelo valor de cada campo de dimensão que identifica a série raiz. Os valores de dimensão são separados por sublinhados. Não há nenhuma restrição, a não ser exclusividade, sobre o valor do prefixo.

Expressão para valores de cenário para série de raiz
Os valores de cenário para um grupo de cenário são especificados por uma expressão, que é, então, utilizada para calcular os valores de cada uma das séries raiz no grupo. É possível inserir uma expressão diretamente ou clicar no botão de calculadora e criar a expressão no Construtor de Expressão de Valores de Cenário.

- A expressão pode conter qualquer resposta ou entrada no sistema de modelo.
- Quando o período de cenário ultrapassa os dados existentes, a expressão é aplicada aos valores previstos dos campos na expressão.
- Para cada série raiz no grupo, os campos na expressão especificam séries temporais que são definidas por esses campos e os valores de dimensão que definem a série raiz. São essas séries temporais que são utilizadas para avaliar a expressão. Por exemplo, se uma série raiz for definida por `region='north'` e `brand='X'`, então as séries temporais que são utilizadas na expressão serão definidas por esses mesmos valores de dimensão.

Como exemplo, suponha que o campo de métrica raiz seja *advertising* e que haja dois campos de dimensão *region* e *brand*. Além disso, suponha que o grupo de cenários inclua todas as combinações de valores de campo de dimensão. Em seguida, é possível especificar `advertising*1.2` como a expressão para investigar o efeito de aumentar *advertising* em 20 por cento para cada série temporal que estiver associada ao campo *advertising*.

Nota: Os grupos de cenários se aplicam apenas aos dados multidimensionais e são criados ao clicar em **Incluir Grupo de Cenários** na guia Cenários.

Opções

Nível máximo para respostas afetadas
Especifica o número máximo de níveis de respostas afetadas. Cada nível sucessivo, até um máximo de 5, inclui respostas que são mais indiretamente afetadas pela série raiz. Especificamente, o primeiro nível inclui as respostas que possuem a série raiz como uma entrada direta. As respostas no segundo nível possuem respostas no primeiro nível como uma entrada direta, e assim por diante. Aumentar o valor dessa configuração aumenta a complexidade do cálculo e pode afetar o desempenho.

Máximo de respostas detectadas automaticamente
Especifica o número máximo de respostas afetadas que são detectadas automaticamente para cada série raiz. Aumentar o valor dessa configuração aumenta a complexidade do cálculo e pode afetar o desempenho.

Diagrama de impacto
Exibe uma representação gráfica das relações causais entre a série raiz para cada cenário e a série de destino que afeta essa série raiz. Tabelas de ambos os valores de cenário e valores preditos para as respostas afetadas são incluídas como parte do item de saída. O diagrama inclui gráficos dos valores preditos das respostas afetadas. Clicar uma vez em qualquer nó no diagrama de impacto abre um diagrama de sequência detalhado para a série que está associada ao nó. Um diagrama de impacto separado é gerado para cada cenário.

Gráficos de séries
Gera gráficos de série dos valores preditos para cada uma das variáveis afetadas em cada cenário.

Tabelas de previsão e cenário
Tabelas de valores preditos e de valores de cenário para cada cenário. Essas tabelas contêm as mesmas informações que as tabelas no diagrama de impacto. Essas tabelas suportam todas as variáveis padrão para editar e definir tabelas como dinâmicas.

Incluir intervalos de confiança nos gráficos e tabelas
Especifica se os intervalos de confiança para as predições de cenário estão incluídos na saída do gráfico e da tabela.

Largura de intervalo de confiança (%)
Esta configuração controla os intervalos de confiança para as predições de cenário. É possível especificar quaisquer valores positivos menores que 100. Por padrão, um intervalo de confiança de 95% é utilizado.

Capítulo 14. Modelos do Nó de Resposta de Autoaprendizado

Nó SLRM

O nó **Self-Learning Response Model** (SLRM) permite construir um modelo que você possa atualizar, ou reestimar, continuamente, conforme um conjunto de dados cresce sem a necessidade de reconstruir o modelo toda vez que usar o conjunto de dados completo. Por exemplo, isso é útil quando você tiver vários produtos e desejar identificar qual produto um cliente mais provavelmente comprará se você oferecer a ele. Esse modelo permite prever quais ofertas são mais apropriadas para clientes e a probabilidade dessas ofertas serem aceitas.

O modelo pode inicialmente ser construído utilizando um conjunto de dados pequeno com ofertas feitas aleatoriamente e as respostas a essas ofertas. Conforme o conjunto de dados cresce, o modelo pode ser atualizado e, portanto, se tornar mais apto para prever as ofertas mais adequadas para clientes e a probabilidade da aceitação dessas ofertas com base em outros campos de entrada, como idade, sexo, profissão e renda. As ofertas disponíveis podem ser alteradas ao incluí-las ou removê-las da caixa de diálogo do nó, ao invés de ter que alterar o campo de destino do conjunto de dados.

Quando acoplado ao IBM SPSS Collaboration and Deployment Services, é possível configurar atualizações regulares automáticas para o modelo. Esse processo, que não necessita de supervisão ou ação manual, fornece uma solução flexível e de baixo custo para organizações e aplicativos onde uma intervenção customizada por um minerador de dados não for possível ou necessária.

Exemplo. Uma instituição financeira deseja obter resultados mais rentáveis ao corresponder a oferta que tem maior probabilidade de ser aceita por cada cliente. É possível utilizar um modelo de autoaprendizado para identificar as características de clientes com maior probabilidade de responderem favoravelmente com base em promoções anteriores e atualizar o modelo em tempo real com base nas respostas mais recentes do cliente.

Opções de Campo do Nó SLRM

Antes de executar um nó SLRM, deve-se especificar os campos de destino e de resposta de destino na guia Campos do nó.

Campo de destino. Selecione o campo de destino na lista, por exemplo, um campo nominal (conjunto) que contendo os produtos diferentes que desejar oferecer aos clientes.

Nota: o campo de destino deve ter armazenamento de sequência de caracteres, não numérico.

Campo de resposta de destino. Selecione o campo de resposta de destino na lista. Por exemplo, Aceito ou Rejeitado.

Nota: este campo deve ser um Flag. O valor true do flag indica aceitação da oferta e o valor false indica recusa da oferta.

Os campos restantes nessa caixa de diálogo são aqueles padrão utilizados em todo o IBM SPSS Modeler. Consulte o tópico “Opções de Campos do Nó de Modelagem” na página 31 para obter mais informações.

Nota: se os dados de origem incluírem intervalos que devem ser utilizados como campos de entrada contínuos (intervalo numérico), assegure-se de que os metadados incluam detalhes mínimo e máximo para cada intervalo.

Opções de Modelo do Nó SLRM

Nome do modelo É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Utilizar dados particionados. Se um campo de partição for definido, essa opção assegurará que apenas os dados da partição de treinamento sejam utilizados para construir o modelo.

Continuar treinando o modelo existente. Por padrão, um modelo completamente novo é criado sempre que um nó de modelagem é executado. Se essa opção for selecionada, o treinamento continua com o último modelo produzido com sucesso pelo nó. Isso permite atualizar ou renovar um modelo existente sem precisar acessar os dados originais e poderá resultar em um desempenho significativamente mais rápido desde que *apenas* os registros novos ou atualizados sejam alimentados no fluxo. Detalhes do modelo anterior são armazenados com o nó de modelagem, o que permite utilizar essa opção mesmo se o nugget do modelo anterior não estiver mais disponível na paleta de fluxo ou de Modelos.

Valores do campo de destino Por padrão, isto é configurado para **Usar tudo**, o que significa que um modelo será construído contendo cada oferta associada ao valor do campo de destino selecionado. Se desejar gerar um modelo que contém apenas algumas das ofertas do campo de destino, clique em **Especificar** e utilize os botões **Incluir**, **Editar** e **Excluir** para inserir ou corrigir os nomes das ofertas para as quais deseja construir um modelo. Por exemplo, se você escolher um destino que lista todos os produtos que fornecer, será possível utilizar este campo para limitar os produtos oferecidos a apenas alguns que forem inseridos aqui.

Avaliação do Modelo. Os campos nesse painel são independentes do modelo por não afetarem o escore. Ao invés disso, eles permitem criar uma representação visual do quão bem o modelo irá prever os resultados.

Nota: para exibir os resultados da avaliação do modelo no nugget do modelo, deve-se selecionar também a caixa **Exibir avaliação do modelo**.

- **Incluir avaliação de modelo.** Selecione esta caixa para criar gráficos que mostram a precisão predita do modelo para cada oferta selecionada.
- **Configurar semente aleatória.** Ao estimar a precisão de um modelo com base em uma porcentagem aleatória, esta opção permite duplicar os mesmos resultados em outra sessão. Ao especificar o valor inicial utilizado pelo gerador de número aleatório, é possível assegurar que os mesmos registros sejam designados toda vez que o nó for executado. Insira o valor semente desejado. Se essa opção não estiver selecionada, uma amostra diferente será gerada toda vez que o nó for executado.
- **Tamanho simulado da amostra.** Especifique o número de registros a serem utilizados na amostra ao avaliar o modelo. O padrão é 100.
- **Número de iterações.** Permite parar a construção da avaliação do modelo após o número de iterações especificado. Especifique o número máximo de iterações; o padrão é 20.

  Nota: lembre-se de que tamanhos de amostra maiores e números altos de iterações aumentam a quantidade de tempo necessária para construir o modelo.

Exibir avaliação de modelo. Selecione esta opção para exibir uma representação gráfica dos resultados no nugget do modelo.

Opções de Configurações do Nó SLRM

As opções de configurações de nó permitem fazer um ajuste preciso do processo de construção de modelo.

Número máximo de preditores por registro Essa opção permite limitar o número de predições feitas para cada registro no conjunto de dados. O padrão é 3.

Por exemplo, você pode ter seis ofertas (como economias, hipoteca, empréstimo para compra de carro, pensão, cartão de crédito e seguro), mas deseja saber as duas melhores para recomendar; nesse caso, você configura esse campo para 2. Ao construir o modelo e anexá-lo a uma tabela, você verá duas colunas de predição (e a confiança associada na probabilidade da oferta que está sendo aceita) por registro. As predições podem ser compostas de qualquer uma das seis ofertas possíveis.

Nível de aleatorização Para evitar quaisquer vieses, por exemplo, em um conjunto de dados pequeno ou incompleto, e tratar todas as possíveis ofertas igualmente, será possível incluir um nível de aleatorização na seleção de ofertas e a probabilidade de elas serem incluídas como ofertas recomendadas. A aleatorização é expressa como uma porcentagem, mostrada como valores decimais entre 0,0 (nenhuma aleatorização) e 1,0 (completamente aleatorizada). O padrão é 0,0.

Configurar semente aleatória Ao incluir um nível de aleatorização para seleção de uma oferta, esta opção permite duplicar os mesmos resultados em outra sessão. Ao especificar o valor inicial utilizado pelo gerador de número aleatório, é possível assegurar que os mesmos registros sejam designados toda vez que o nó for executado. Insira o valor semente desejado. Se essa opção não estiver selecionada, uma amostra diferente será gerada toda vez que o nó for executado.

Nota: Ao usar a opção **Configurar semente aleatória** com registros lidos a partir de um banco de dados, um nó Ordenar poderá ser necessário antes da amostragem para assegurar o mesmo resultado sempre que o nó for executado. Isso ocorre porque a semente aleatória depende da ordem de registros, que não é garantido que ela permaneça a mesma em um banco de dados relacional.

Ordenação Selecione a ordem na qual as ofertas devem ser exibidas no modelo construído:

- **Decrescente** O modelo exibe as ofertas com os escores mais altos primeiro. Essas são as ofertas que possuem a probabilidade maior de aceitação.
- **Crescente** O modelo exibe as ofertas com os escores mais baixos primeiro. Essas são as ofertas que possuem a probabilidade maior de rejeição. Por exemplo, isso pode ser útil quando decidir quais clientes deverão ser removidos de uma campanha de marketing para uma oferta específica.

Preferências para campos de destino Ao construir um modelo, poderá haver determinados aspectos dos dados que você deseja promover ou remover ativamente. Por exemplo, se estiver construindo um modelo que seleciona a melhor oferta financeira para promover para um cliente, você poderá querer assegurar que uma oferta específica seja sempre incluída, independentemente do quão bem ela for escorada com relação a cada cliente.

Para incluir uma oferta neste painel e editar suas preferências, clique em **Incluir**, insira o nome da oferta (por exemplo, Economias ou Hipoteca), e clique em **OK**.

- **Valor** Mostra o nome da oferta que você incluiu.
- **Preferências** Especifique o nível de preferência a ser aplicado à oferta. A preferência é expressa como uma porcentagem, mostrada como valores decimais entre 0,0 (não preferencial) e 1,0 (mais preferencial). O padrão é 0,0.
- **Sempre incluir** Para assegurar que uma oferta específica seja sempre incluída nas predições, selecione esta caixa.

  Nota: Se **Preferência** for configurada como 0,0, a configuração **Sempre incluir** será ignorada.

Levar em consideração a confiabilidade do modelo Um modelo bem estruturado rico em dados que tiver sido ajustado através de várias regenerações deve sempre produzir resultados mais precisos com relação a um novo modelo com poucos dados. Para aproveitar a confiabilidade maior do modelo mais maduro, selecione esta caixa.

Nuggets do Modelo SLRM

Nota: os resultados serão mostrados nesta guia apenas se você selecionar as opções **Incluir avaliação de modelo** e **Exibir avaliação de modelo** na guia Opções de modelo.

Ao executar um fluxo que contém um modelo SLRM, o nó estima a precisão das predições para cada valor de campo de destino (oferta) e a importância de cada preditor utilizado.

Nota: se você selecionou **Continuar treinamento de modelo existente** na guia Modelo do nó de modelagem, as informações mostradas no nugget do modelo serão atualizadas toda vez que o modelo for gerado novamente.

Para os modelos construídos usando o IBM SPSS Modeler 12.0 ou posterior, a guia Modelo do nugget do modelo é dividida em duas colunas:

Coluna esquerda.

- **Visualização.** Quando você tiver mais de uma oferta, selecione aquela para a qual deseja exibir os resultados.
- **Desempenho do modelo.** Isso mostra a precisão do modelo estimada de cada oferta. O conjunto de testes é gerado por meio de simulação.

Coluna direita.

- **Visualização.** Selecione se deseja exibir os detalhes de **Associação com Resposta** ou de **Importância de Variável**.
- **Associação com Resposta.** Exibe a associação (correlação) de cada preditor com a variável de destino.
- **Importância do Preditor.** Indica a importância relativa de cada preditor na estimativa do modelo. Geralmente você desejará focar seus esforços de modelagem nos preditores que forem mais importantes e considerar descartar ou ignorar aqueles que forem menos importantes. Este gráfico pode ser interpretado da mesma forma que outros modelos que exibem a importância do preditor, embora no caso do SLRM, o gráfico é gerado por meio de simulação pelo algoritmo SLRM. Isso é feito ao remover cada preditor por vez do modelo e ver como isso afeta a precisão do modelo. Consulte o tópico "Importância do preditor" na página 44 para obter mais informações.

Configurações do Modelo SLRM

A guia Configurações de um nugget do modelo SLRM especifica opções para modificar o modelo construído. Por exemplo, é possível utilizar o nó SLRM para construir vários modelos diferentes usando os mesmos dados e configurações e, em seguida, utilizar essa guia em cada modelo para modificar um pouco as configurações para ver como isso afeta os resultados.

Nota: Esta guia estará disponível somente após o nugget do modelo ter sido incluído em um fluxo.

Número máximo de preditores por registro Essa opção permite limitar o número de predições feitas para cada registro no conjunto de dados. O padrão é 3.

Por exemplo, você pode ter seis ofertas (como economias, hipoteca, empréstimo para compra de carro, pensão, cartão de crédito e seguro), mas deseja saber as duas melhores para recomendar; nesse caso, você configura esse campo para 2. Ao construir o modelo e anexá-lo a uma tabela, você verá duas colunas de predição (e a confiança associada na probabilidade da oferta que está sendo aceita) por registro. As predições podem ser compostas de qualquer uma das seis ofertas possíveis.

Nível de aleatorização Para evitar quaisquer vieses, por exemplo, em um conjunto de dados pequeno ou incompleto, e tratar todas as possíveis ofertas igualmente, será possível incluir um nível de aleatorização na seleção de ofertas e a probabilidade de elas serem incluídas como ofertas recomendadas. A

aleatorização é expressa como uma porcentagem, mostrada como valores decimais entre 0,0 (nenhuma aleatorização) e 1,0 (completamente aleatorizada). O padrão é 0,0.

Configurar semente aleatória Ao incluir um nível de aleatorização para seleção de uma oferta, esta opção permite duplicar os mesmos resultados em outra sessão. Ao especificar o valor inicial utilizado pelo gerador de número aleatório, é possível assegurar que os mesmos registros sejam designados toda vez que o nó for executado. Insira o valor semente desejado. Se essa opção não estiver selecionada, uma amostra diferente será gerada toda vez que o nó for executado.

Nota: Ao usar a opção **Configurar semente aleatória** com registros lidos a partir de um banco de dados, um nó Ordenar poderá ser necessário antes da amostragem para assegurar o mesmo resultado sempre que o nó for executado. Isso ocorre porque a semente aleatória depende da ordem de registros, que não é garantido que ela permaneça a mesma em um banco de dados relacional.

Ordenação Selecione a ordem na qual as ofertas devem ser exibidas no modelo construído:

- **Decrescente** O modelo exibe as ofertas com os escores mais altos primeiro. Essas são as ofertas que possuem a probabilidade maior de aceitação.
- **Crescente** O modelo exibe as ofertas com os escores mais baixos primeiro. Essas são as ofertas que possuem a probabilidade maior de rejeição. Por exemplo, isso pode ser útil quando decidir quais clientes deverão ser removidos de uma campanha de marketing para uma oferta específica.

Preferências para campos de destino Ao construir um modelo, poderá haver determinados aspectos dos dados que você deseja promover ou remover ativamente. Por exemplo, se estiver construindo um modelo que seleciona a melhor oferta financeira para promover para um cliente, você poderá querer assegurar que uma oferta específica seja sempre incluída, independentemente do quão bem ela for escorada com relação a cada cliente.

Para incluir uma oferta neste painel e editar suas preferências, clique em **Incluir**, insira o nome da oferta (por exemplo, Economias ou Hipoteca), e clique em **OK**.

- **Valor** Mostra o nome da oferta que você incluiu.
- **Preferências** Especifique o nível de preferência a ser aplicado à oferta. A preferência é expressa como uma porcentagem, mostrada como valores decimais entre 0,0 (não preferencial) e 1,0 (mais preferencial). O padrão é 0,0.
- **Sempre incluir** Para assegurar que uma oferta específica seja sempre incluída nas predições, selecione esta caixa.

  Nota: Se **Preferência** for configurada como 0,0, a configuração **Sempre incluir** será ignorada.

Levar em consideração a confiabilidade do modelo Um modelo bem estruturado rico em dados que tiver sido ajustado através de várias regenerações deve sempre produzir resultados mais precisos com relação a um novo modelo com poucos dados. Para aproveitar a confiabilidade maior do modelo mais maduro, selecione esta caixa.

Gerar SQL para este modelo Ao usar dados de um banco de dados, código SQL pode ser enviado por push de volta para o banco de dados para execução, fornecendo desempenho superior para muitas operações.

Selecione uma das opções a seguir para especificar como a geração de SQL é executada.

- **Padrão: Escorar usando o Server Scoring Adapter (se instalado) no processo** Se conectado a um banco de dados com um adaptador de escoragem instalado, gera a SQL utilizando o adaptador de escoragem e funções definidas pelo usuário (UDF) associadas e escora seu modelo no banco de dados. Quando nenhum adaptador de escoragem estiver disponível, essa opção busca seus dados novamente a partir do banco de dados e os escora no SPSS Modeler.

- **Escorar fora do Banco de dados** Se selecionada, esta opção busca seus dados novamente a partir do banco de dados e os escora no SPSS Modeler.

Capítulo 15. Modelos de Support Vector Machine

Sobre o SVM

O Support Vector Machine (SVM) é uma técnica robusta de classificação e regressão que maximiza a precisão preditiva de um modelo sem causar super ajuste dos dados de treinamento. O SVM é particularmente adequado para analisar dados com números muito grandes (por exemplo, milhares) de campos preditores.

O SVM possui aplicações em várias disciplinas, dentre elas gerenciamento de relacionamento com o cliente (CRM), reconhecimento facial ou de outras imagens, bioinformática, extração de conceito de mineração de texto, detecção de intrusão, predição de estrutura proteica e reconhecimento de voz.

Como o SVM Funciona

O SVM funciona ao mapear dados para um espaço de variável altamente dimensional para que os pontos de dados possam ser categorizados, mesmo quando os dados não forem de outra forma linearmente separáveis. Após um separador entre as categorias ser localizado, os dados serão transformados de modo que o separador possa ser desenhado como um hiperplano. Após disso, as características dos novos dados podem ser utilizadas para prever o grupo ao qual um novo registro deve pertencer.

Por exemplo, considere a figura a seguir, na qual os pontos de dados caem em duas categorias diferentes.

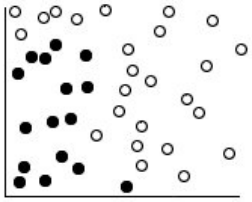


Figura 60. Conjunto de dados original

As duas categorias podem ser separadas com uma curva, conforme mostrado na figura a seguir.

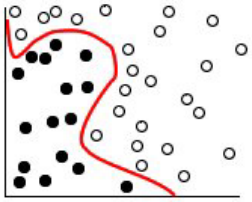


Figura 61. Dados com separador incluído

Após a transformação, o limite entre as duas categorias poderá ser definido por um hiperplano, conforme mostrado na figura a seguir.

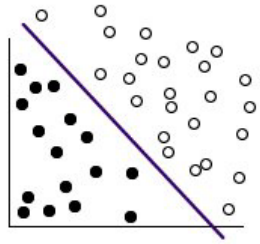


Figura 62. Dados transformados

A função matemática usada para a transformação é conhecida como a função **kernel**. O SVM no IBM SPSS Modeler suporta os tipos de kernel a seguir:

- Linear
- Polinomial
- Função de base radial (RBF)
- Curva sigmoide

Uma função kernel linear é recomendada quando a separação linear dos dados for direta. Em outros casos, uma das outras funções deve ser utilizada. Será necessário experimentar diferentes funções para obter o melhor modelo em cada caso, já que cada uma delas utiliza diferentes algoritmos e parâmetros.

Ajustando um Modelo de SVM

Além da linha de separação entre as categorias, um modelo SVM de classificação também localiza linhas marginais que definem o espaço entre as duas categorias.

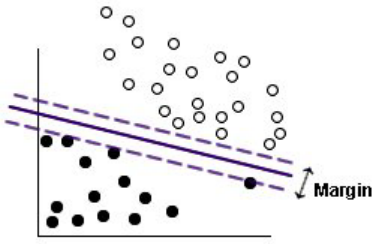


Figura 63. Dados com um modelo preliminar

Os pontos de dados que se encontram nas margens são conhecidos como os **vetores de suporte**.

Quanto maior for a margem entre as duas categorias, melhor o modelo será na previsão da categoria para novos registros. No exemplo anterior, a margem não é muito ampla e o modelo é considerado **super ajustado**. Uma pequena quantia de classificação errada pode ser aceita para ampliar a margem; um exemplo disso é mostrado na figura a seguir.

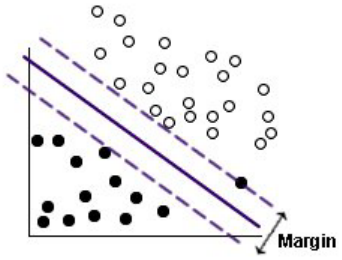


Figura 64. Dados com um modelo melhorado

Em alguns casos, a separação linear é mais difícil; um exemplo disso é mostrado na figura a seguir.

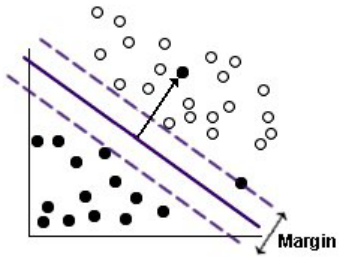


Figura 65. Um problema para separação linear

Em um caso como esse, o objetivo é encontrar o equilíbrio ideal entre uma margem larga e um número pequeno de pontos de dados classificados incorretamente. A função de kernel possui um **parâmetro de regularização** (conhecido como C) que controla o trade-off entre esses dois valores. Provavelmente será necessário experimentar valores diferentes deste e de outros parâmetros do kernel para localizar o melhor modelo.

Nó SVM

O nó SVM permite utilizar uma Support Vector Machine para classificar os dados. O SVM é particularmente adequado para uso com amplos conjuntos de dados, ou seja, aqueles com um grande número de campos preditores. É possível utilizar as configurações padrão no nó para produzir um modelo básico relativamente rápido, ou utilizar as configurações de Especialista para experimentar com diferentes tipos de modelo SVM.

Quando o modelo for construído, será possível:

- Procurar o nugget do modelo para exibir a importância relativa dos campos de entrada na construção do modelo.
- Anexar um nó Tabela ao nugget do modelo para visualizar a saída do modelo.

Exemplo. Um pesquisador médico obteve um conjunto de dados contendo características de um número de amostras de células humanas extraídas de pacientes que se acredita que estejam correndo risco de desenvolver câncer. A análise dos dados originais demonstrou que muitas das características diferiram significativamente entre amostras benignas e malignas. O pesquisador deseja desenvolver um modelo SVM que possa utilizar os valores de características de células semelhantes em amostras de outros pacientes para fornecer uma indicação precoce se suas amostras podem ser benignas ou malignas.

Opções do Modelo do Nó SVM

Nome do modelo É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Utilizar dados particionados. Se um campo de partição for definido, essa opção assegurará que apenas os dados da partição de treinamento sejam utilizados para construir o modelo.

Criar modelos de divisão. Constrói um modelo separado para cada valor possível de campos de entrada que são especificados como campos de divisão. Consulte o tópico "Construindo Modelos de Divisão" na página 28 para obter mais informações.

Opções Avançadas do Nó SVM

Se você tiver conhecimento detalhado do Support Vector Machine, as opções avançadas permitem fazer um ajuste preciso do processo de treinamento. Para acessar as opções avançadas, configure o Modo para **Especialista** na guia Especialista.

Incluir todas as probabilidades (válido apenas para variáveis resposta categóricas). Se selecionada (marcada), especifica que as probabilidades de cada valor possível de um campo nominal ou de destino de sinalização são exibidas para cada registro processado pelo nó. Se essa opção não estiver selecionada, a probabilidade somente do valor predito é exibida para os campos nominal ou de destino de sinalização. A configuração dessa caixa de seleção determina o estado padrão da caixa de seleção correspondente na exibição de nugget do modelo.

Critérios de parada. Determina quando parar o algoritmo de otimização. O intervalo de valores 1.0E-1 a 1.0E-6; o padrão é 1.0E-3. Reduzir o valor resulta em um modelo mais preciso, mas o modelo levará mais tempo para treinar.

Parâmetro de regularização (C). Controla o trade-off entre maximizar a margem e minimizar o termo de erro de treinamento. O valor deve estar normalmente entre 1 e 10, inclusive; o padrão é 10. Aumentar o valor melhora a precisão da classificação (ou reduz o erro de regressão) dos dados de treinamento, mas isso também pode levar a super ajuste.

Precisão de regressão (epsilon). Utilizado apenas se o nível de medição do campo de destino for *Contínuo*. Faz com que erros sejam aceitos, contanto que eles sejam menores do que o valor especificado aqui. Aumentar o valor pode resultar em uma modelagem mais rápida, mas à custa da precisão.

Tipo de Kernel. Determina o tipo de função de kernel utilizado para a transformação. Diferentes tipos de kernel fazem com que o separador seja calculado em diferentes formas, portanto, é aconselhável experimentar com as várias opções. O padrão é **RBF** (Função de Base Radial).

Gama RBF. Ativado apenas se o tipo de kernel for configurado para **RBF**. O valor deve estar normalmente entre $3/k$ e $6/k$, em que k é o número de campos de entrada. Por exemplo, se houver 12 campos de entrada, os valores entre 0,25 e 0,5 já valeriam a pena tentar. Aumentar o valor melhora a precisão da classificação (ou reduz o erro de regressão) dos dados de treinamento, mas isso também pode levar a super ajuste.

Gama. Ativado apenas se o tipo de kernel for configurado para **Polinomial** ou **Sigmoide**. Aumentar o valor melhora a precisão da classificação (ou reduz o erro de regressão) dos dados de treinamento, mas isso também pode levar a super ajuste.

Viés. Ativado apenas se o tipo de kernel for configurado para **Polinomial** ou **Sigmoide**. Configura o valor de `coef0` na função de kernel. O valor padrão 0 é apropriado na maioria dos casos.

Grau. Ativado somente se o tipo de Kernel for configurado para **Polinomial**. Controla a complexidade (dimensão) do espaço de mapeamento. Normalmente, você não utilizaria um valor maior que 10.

Nugget do Modelo SVM

O modelo SVM cria um número de novos campos. O mais importante deles é o campo **$S-fieldname**, que mostra o valor do campo de destino predito pelo modelo.

O número e os nomes dos novos campos criados pelo modelo dependem do nível de medição do campo de destino (esse campo será indicado nas tabelas a seguir por *fieldname*).

Para ver esses campos e seus valores, inclua um nó Tabela no nugget do modelo SVM e execute o nó Tabela.

Tabela 27. O nível de medição do campo de destino é 'Nominal' ou 'Flag'

Novo nome de campo	Descrição
$S-*fieldname*	Valor predito do campo de destino.
$SP-*fieldname*	Probabilidade do valor predito.
$SP-*value*	A probabilidade de cada valor possível do valor nominal ou de flag (exibido apenas se **Incluir todas as probabilidades** estiver selecionada na guia Configurações do nugget do modelo).
$SRP-*value*	(Apenas respostas de flag) Escores de propensão Bruta (SRP) e ajustada (SAP), indicando a probabilidade de um resultado "true" para o campo de destino. Esses scores serão exibidos somente se as caixas de seleção correspondentes forem selecionadas na guia Analisar do nó de modelagem SVM antes de o modelo ser gerado. Consulte o tópico "Opções de Análise do Nó de Modelagem" na página 35 para obter mais informações.
$SAP-*value*	

Tabela 28. O nível de medição do campo de destino é 'Contínuo'

Novo nome de campo	Descrição
$S-*fieldname*	Valor predito do campo de destino.

Importância do preditor

Opcionalmente, um gráfico que indica a importância relativa de cada preditor na estimativa do modelo também pode ser exibido na guia Modelo. Geralmente você desejará focar seus esforços de modelagem nos preditores que forem mais importantes e considerar descartar ou ignorar aqueles que forem menos importantes. Observe que este gráfico estará disponível apenas se **Calcular a importância do preditor** estiver selecionada na guia Análise antes de gerar o modelo. Consulte o tópico "Importância do preditor" na página 44 para obter mais informações.

Nota: a importância do preditor pode levar mais tempo para calcular para SVM do que para outros tipos de modelos, e não é selecionada na guia Analisar por padrão. Selecionar essa opção pode diminuir o desempenho, principalmente com grandes conjuntos de dados.

Configurações do Modelo de SVM

A guia Configurações permite especificar campos extras a serem exibidos quando visualizar os resultados (por exemplo, ao executar um nó Tabela anexado ao nugget). É possível ver o efeito de cada uma dessas opções ao selecioná-las e clicar no botão Visualizar -- role para a direita da saída Visualizar para ver os campos extras.

Incluir todas as probabilidades (válido apenas para variáveis resposta categóricas). Se essa opção for selecionada, as probabilidades para cada valor possível de um campo de destino nominal ou de flag são

exibidas para cada registro processado pelo nó. Se essa opção estiver desmarcada, apenas o valor predito e sua probabilidade serão exibidos para campos de destino nominal ou de flag.

A configuração padrão dessa caixa de seleção é determinada pela caixa de seleção correspondente no nó de modelagem.

Calcular escores de propensão bruta. Para modelos com uma resposta de flag (que retornam uma predição de sim ou não), é possível solicitar os escores de propensão que indicam a probabilidade do resultado real especificado para o campo de destino. Esses são um complemento dos outros valores de predição e de confiança que podem ser gerados durante a escoragem.

Calcular escores de propensão ajustada. Os escores de propensão bruta baseiam-se apenas nos dados de treinamento e esses podem ser altamente otimistas devido à tendência de muitos modelos a super ajustar desses dados. As propensões ajustadas tentam compensar ao avaliar o desempenho do modelo com relação à partição de teste ou de validação. Essa opção requer que um campo de partição seja definido no fluxo e que os escores de propensão ajustada sejam ativados no nó de modelagem antes de gerar o modelo.

Gerar SQL para este modelo Ao usar dados de um banco de dados, código SQL pode ser enviado por push de volta para o banco de dados para execução, fornecendo desempenho superior para muitas operações.

Selecione uma das opções a seguir para especificar como a geração de SQL é executada.

- **Padrão: Escorar usando o Server Scoring Adapter (se instalado) no processo** Se conectado a um banco de dados com um adaptador de escoragem instalado, gera a SQL utilizando o adaptador de escoragem e funções definidas pelo usuário (UDF) associadas e escora seu modelo no banco de dados. Quando nenhum adaptador de escoragem estiver disponível, essa opção busca seus dados novamente a partir do banco de dados e os escora no SPSS Modeler.
- **Escorar fora do Banco de dados** Se selecionada, esta opção busca seus dados novamente a partir do banco de dados e os escora no SPSS Modeler.

Capítulo 16. Modelos de Vizinho Mais Próximo

Nó KNN

A Análise do Vizinho Mais Próximo é um método de classificação de casos com base na sua similaridade com outros casos. Em aprendizado por máquina, ela foi desenvolvida como uma maneira de reconhecer padrões de dados sem requerer uma correspondência exata com nenhum dos padrões ou casos armazenados. Os casos semelhantes estão próximos uns dos outros e os casos dissimilares estão distantes. Portanto, a distância entre dois casos é uma medida de sua dissimilaridade.

Os casos que estiverem próximos uns dos outros são chamados de "vizinhos". Quando um novo caso (validação) é apresentado, sua distância de cada um dos casos no modelo é calculada. As classificações dos casos mais similares – os vizinhos mais próximos – são verificadas e o novo caso é colocado na categoria que contiver o maior número de vizinhos mais próximos.

É possível especificar o número de vizinhos mais próximos para examinar; este valor é chamado de *k*. As imagens mostram como um novo caso será classificado utilizando dois valores diferentes de *k*. Quando *k* = 5, o novo caso é colocado na categoria *1* porque a maioria dos vizinhos mais próximos pertence à categoria *1*. No entanto, quando *k* = 9, o novo caso é colocado na categoria *0* porque a maioria dos vizinhos mais próximos pertence à categoria *0*.

A análise do vizinho mais próximo também pode ser utilizada para calcular valores para uma variável resposta contínua. Nesta situação, a média ou mediana do valor dos vizinhos mais próximos é utilizada para obter o valor predito para o novo caso.

Opções Objetivas do Nó KNN

Na guia Objetivos, é possível escolher se deseja construir um modelo que prediz o valor de um campo de destino em seus dados de entrada com base nos valores de seus vizinhos mais próximos, ou apenas localizar quais são os vizinhos mais próximos para um determinado caso de interesse.

Que tipo de análise você deseja realizar?

Prever um campo de destino. Escolha esta opção se você desejar prever o valor de um campo de destino com base nos valores de seus vizinhos mais próximos.

Identificar apenas os vizinhos mais próximos. Escolha esta opção se desejar apenas ver quais são os vizinhos mais próximos para um campo de entrada específico.

Se escolher identificar apenas os vizinhos mais próximos, o restante das opções nesta guia relacionadas à precisão e à velocidade será desativado já que elas são relevantes apenas para prever respostas.

Qual é o seu objetivo?

Ao prever um campo de destino, esse grupo de opções permite decidir se a velocidade, a precisão, ou uma combinação de ambos, são os fatores mais importantes ao prever um campo de destino. Como alternativa, é possível optar por customizar as configurações manualmente.

Se você escolher a opção Balanceamento, Velocidade ou Precisão, o algoritmo pré-selecionará a combinação mais adequada de configurações para essa opção. Usuários avançados podem querer substituir essas seleções, o que pode ser feito nos vários painéis da guia Configurações.

Balancear velocidade e precisão. Seleciona o melhor número de vizinhos dentro de um pequeno intervalo.

Velocidade. Localiza um número fixo de vizinhos.

Precisão. Seleciona o melhor número de vizinhos dentro de um intervalo maior e utiliza a importância do preditor ao calcular distâncias.

Análise customizada. Escolha esta opção para fazer um ajuste preciso do algoritmo na guia Configurações.

Nota: o tamanho do modelo KNN resultante, ao contrário da maioria dos outros modelos, aumenta linearmente com a quantia de dados de treinamento. Se, ao tentar construir um modelo KNN, for exibido um erro de falta de memória, tente aumentar a memória máxima do sistema utilizada pelo IBM SPSS Modeler. Para isso, escolha

Ferramentas > **Opções** > **Opções do Sistema**

e insira o novo tamanho no campo **Máximo de memória**. As mudanças feitas no diálogo Opções do Sistema não entrarão em vigor até que você reinicie o IBM SPSS Modeler.

Configurações do Nó KNN

A guia Configurações é onde você define as opções que são específicas para a Análise do Vizinho Mais Próximo. A barra lateral à esquerda da tela lista os painéis que são utilizados para especificar as opções.

Modelo

O painel Modelo fornece opções que controlam como o modelo deve ser construído, por exemplo, se deseja utilizar modelos de particionamento ou de divisão, se deseja transformar campos de entrada numéricos para que todos eles estejam dentro do mesmo intervalo e como gerenciar os casos de interesse. Também é possível escolher um nome customizado para o modelo.

Nota: As opções **Usar dados particionados** e **Usar rótulos case** não podem usar o mesmo campo.

Nome do modelo É possível gerar o nome do modelo automaticamente com base no campo de destino ou de ID (ou no tipo de modelo nos casos em que não houver tal campo especificado) ou especificar um nome customizado.

Utilizar dados particionados. Se um campo de partição for definido, essa opção assegurará que apenas os dados da partição de treinamento sejam utilizados para construir o modelo.

Criar modelos de divisão. Constrói um modelo separado para cada valor possível de campos de entrada que são especificados como campos de divisão. Consulte o tópico “Construindo Modelos de Divisão” na página 28 para obter mais informações.

Para selecionar campos manualmente... Por padrão, o nó utiliza as configurações do campo de partição e de divisão (se houver) a partir do nó Tipo, mas é possível substituir essas configurações aqui. Para ativar os campos **Partição** e **Divisões**, selecione a guia **Campos**, escolha **Utilizar Configurações Customizadas** e, em seguida, retorne aqui.

- **Partição.** Este campo permite especificar um campo utilizado para particionar os dados em amostras separadas para os estágios de treinamento, de teste e de validação de construção de modelo. Ao utilizar uma amostra para gerar o modelo e uma amostra diferente para testá-lo, é possível obter uma boa indicação do quão bem o modelo será generalizado para conjuntos de dados maiores que forem semelhantes aos dados atuais. Se diversos campos de partição tiverem sido definidos usando os nós Tipo ou Partição, um campo de partição único deverá ser selecionado na guia Campos em cada nó de modelagem que utiliza particionamento. (Se apenas uma partição estiver presente, ela será utilizada automaticamente sempre que o particionamento estiver ativado). Além disso, observe que para aplicar

a partição selecionada à sua análise, o particionamento também deverá ser ativado na guia Opções de Modelo para o nó. (Desmarcar esta opção permite desativar o particionamento sem alterar as configurações do campo).

- **Divisões.** Para os modelos de divisão, selecione o campo ou campos de divisão. Isso é semelhante a configurar o papel do campo para *Divisão* em um nó Tipo. É possível designar apenas campos do tipo **Flag**, **Nominais** ou **Ordinais** como campos de divisão. Os campos escolhidos como campos de divisão não podem ser utilizados como campos de destino, de entrada, de partição, de frequência ou de ponderação. Consulte o tópico "Construindo Modelos de Divisão" na página 28 para obter mais informações.

Normalizar entradas de intervalo. Selecione esta caixa para normalizar os valores dos campos de entrada contínuos. As variáveis normalizadas possuem o mesmo intervalo de valores, que pode melhorar o desempenho do algoritmo de estimação. A normalização ajustada, [2*(x−min)/(max−min)]−1, é usada. Os valores normalizados ajustados estão entre −1 e 1.

Usar rótulos case. Selecione esta caixa para ativar a lista suspensa, na qual é possível escolher um campo cujos valores serão utilizados como rótulos para identificar os casos de interesse no gráfico de espaço do preditor, no gráfico de peers e no mapa de quadrante no visualizador de modelos. Também é possível escolher qualquer campo com um nível de medição de *Nominal*, *Ordinal* ou *Flag* para utilizar como o campo de rotulagem. Se você não escolher um campo aqui, os registros serão exibidos nos gráficos do visualizador de modelo com os vizinhos mais próximos sendo identificados pelo número da linha nos dados de origem. Se desejar manipular os dados no geral após construir o modelo, use os rótulos de caso para evitar ter que consultar de novo os dados de origem toda vez que identificar os casos na exibição.

Identificar registro focal. Selecione esta caixa para ativar a lista suspensa, que permite marcar um campo de entrada de interesse específico (apenas para campos de flag). Se você especificar um campo aqui, os pontos que representam esse campo serão selecionados inicialmente no visualizador de modelo quando o modelo for construído. Selecionar um registro focal aqui é opcional; qualquer ponto pode temporariamente se tornar um registro focal quando selecionado manualmente no visualizador de modelos.

Vizinhos

O painel Vizinhos possui um conjunto de opções que controlam como o número de vizinhos mais próximos é calculado.

Número de Vizinhos Mais Próximos (k). Especifica o número de vizinhos mais próximos para um caso específico. Observe que utilizar um número maior de vizinhos não resultará necessariamente em um modelo mais preciso.

Se o objetivo for prever uma resposta, haverá duas opções:

- **Especificar k fixo.** Utilize esta opção se desejar especificar um número fixo de vizinhos mais próximos para localizar.
- **Selecionar k automaticamente.** Como alternativa, é possível usar os campos **Mínimo** e **Máximo** para especificar um intervalo de valores e permitir que o procedimento escolha o "melhor" número de vizinhos dentro desse intervalo. O método para determinar o número de vizinhos mais próximos depende se a seleção de variável é solicitada no painel Seleção de Variável:

  Se a seleção de variável estiver em vigor, então a seleção de variável será executada para cada valor de *k* no intervalo solicitado, e o *k*, com o conjunto de variáveis associado e com a menor taxa de erro (ou com a menor soma dos quadrados dos erros se a resposta for contínua), serão selecionados.

  Se a seleção de variável não estiver em vigor, então a validação cruzada da dobra *V* será utilizada para selecionar o "melhor" número de vizinhos. Consulte o painel Validação cruzada para obter controle sobre a designação de dobras.

Cálculo de Distância. Esta é a métrica utilizada para especificar a métrica de distância utilizada para medir a similaridade de casos.

- **Métrica euclidiana.** A distância entre dois casos, x e y, é a raiz quadrada da soma, em todas as dimensões, das diferenças quadráticas entre os valores para os casos.
- **Métrica de City Block.** A distância entre dois casos é a soma, em todas as dimensões, das diferenças absolutas entre os valores para os casos. Essa métrica também é chamada de distância de Manhattan.

Opcionalmente, se o objetivo for prever uma resposta, será possível optar por ponderar as variáveis pela sua importância normalizada ao calcular distâncias. A importância de variável para um preditor é calculada pela razão da taxa de erros ou da soma dos quadrados dos erros do modelo com o preditor removido do modelo, com a taxa de erros ou a soma dos quadrados dos erros do modelo integral. A importância normalizada é calculada ao reponderar os valores de importância de variável para que a soma deles seja 1.

Variáveis ponderadas por importância ao calcular distâncias. (Exibido apenas se o objetivo for prever uma resposta). Selecione esta caixa para fazer a importância do preditor seja utilizada ao calcular as distâncias entre os vizinhos. A importância do preditor será então exibida no nugget do modelo e utilizada nas predições (afetando, assim, a escoragem). Consulte o tópico "Importância do preditor" na página 44 para obter mais informações.

Predições para Resposta de Intervalo. (Exibido apenas se o objetivo for prever uma resposta). Se uma resposta contínua (intervalo numérico) for especificada, isso definirá se o valor predito é calculado com base na média ou no valor mediano de vizinhos mais próximos.

Seleção de Variáveis

Este painel será ativado apenas se o objetivo for prever uma resposta. Ele permite solicitar e especificar opções para a seleção de variável. Por padrão, todas as variáveis são consideradas para seleção de variável, no entanto, é possível, opcionalmente, selecionar um subconjunto de variáveis para forçar no modelo.

Executar seleção de variável. Selecione esta caixa para ativar as opções de seleção de variável.

- **Entrada forçada.** Clique no botão de seletor de campo ao lado desta caixa e escolha uma ou mais variáveis para forçar no modelo.

Critério de Parada. Em cada passo, a variável cuja inclusão no modelo resultar no menor erro (calculado como a taxa de erros para uma variável resposta categórica e como a soma dos erros quadráticos para uma variável resposta contínua) é considerada para inclusão no conjunto de modelo. A seleção Forward continua até que a condição especificada seja atendida.

- **Parar quando o número de varáveis especificado tiver sido selecionado.** O algoritmo inclui um número fixo de variáveis além daquelas forçadas no modelo. Especifique um número inteiro positivo. Diminuir valores do número para seleção cria um modelo mais simples, sob risco de perder variáveis importantes. Aumentar valores do número para seleção capturará todas as variáveis importantes, sob risco de incluir eventualmente variáveis que na realidade aumentam o erro do modelo.
- **Parar quando a mudança na razão de erro absoluto for menor ou igual à mínima.** O algoritmo para quando a mudança na razão de erro absoluto indicar que o modelo não poderá ser melhorado ainda mais ao incluir mais variáveis. Especifique um número positivo. Diminuir valores da mudança mínima tende a incluir mais variáveis, sob risco de incluir variáveis que não agregam muito valor ao modelo. Aumentar o valor da mudança mínima tende a excluir mais variáveis, sob risco de perder variáveis que forem importantes para o modelo. O valor "ideal" da mudança mínima dependerá de seus dados e do aplicativo. Consulte o Log de Erro de Seleção de Variável na saída para ajudar a avaliar quais variáveis são mais importantes. Consulte o tópico "Log de Erros de Seleção de Preditor" na página 338 para obter mais informações.

Validação Cruzada

Este painel será ativado apenas se o objetivo for prever uma resposta. As opções neste painel controlam se a validação cruzada deverá ser usada quando calcular os vizinhos mais próximos.

A validação-cruzada divide a amostra em um número de subamostras, ou **dobras**. Em seguida, os modelos de vizinho mais próximo são gerados, excluindo os dados de cada subamostra por vez. O primeiro modelo baseia-se em todos os casos, exceto aqueles na primeira dobra de amostra, o segundo modelo baseia-se em todos os casos, exceto aqueles na segunda dobra de amostra, e assim por diante. Para cada modelo, o erro é estimado ao aplicar o modelo à subamostra excluída ao gerá-lo. O "melhor" número de vizinhos mais próximos é aquele que produz o menor erro entre as dobras.

Dobras de Validação Cruzada. A validação cruzada de dobra *V* é utilizada para determinar o "melhor" número de vizinhos. Ela não está disponível junto com a seleção de variável por motivos de desempenho.

- **Designar aleatoriamente casos às dobras.** Especifique o número de dobras que devem ser utilizadas para validação cruzada. O procedimento designa aleatoriamente casos às dobras, numerados de 1 a *V*, o número de dobras.
- **Configurar semente aleatória.** Ao estimar a precisão de um modelo com base em uma porcentagem aleatória, esta opção permite duplicar os mesmos resultados em outra sessão. Ao especificar o valor inicial utilizado pelo gerador de número aleatório, é possível assegurar que os mesmos registros sejam designados toda vez que o nó for executado. Insira o valor semente desejado. Se essa opção não estiver selecionada, uma amostra diferente será gerada toda vez que o nó for executado.
- **Usar campo para designar casos.** Especifique um campo numérico que designa cada caso no conjunto de dados ativo a uma dobra. O campo deve ser numérico e utilizar valores de 1 a *V*. Se algum valor nesse intervalo estiver omisso, e em quaisquer campos de divisão se os modelos de divisão estiverem em vigor, isto causará um erro.

Analisar

O painel Analisar será ativado apenas se o objetivo for prever uma resposta. Ele pode ser utilizado para especificar se o modelo deve incluir variáveis adicionais para conter:

- as probabilidades de cada valor de campo de destino possível
- as distâncias entre um caso e seus vizinhos mais próximos
- escores de propensão bruta e ajustada (apenas para respostas de flag)

Incluir todas as probabilidades. Se essa opção for selecionada, as probabilidades para cada valor possível de um campo de destino nominal ou de flag são exibidas para cada registro processado pelo nó. Se essa opção estiver desmarcada, apenas o valor predito e sua probabilidade serão exibidos para campos de destino nominal ou de flag.

Salvar as distâncias entre os casos e os k vizinhos mais próximos. Para cada registro focal, uma variável separada é criada para cada um dos *k* vizinhos mais próximos do registro focal (da amostra de treinamento) e as *k* distâncias mais próximas correspondentes.

Escores de Propensão

Os escores de propensão podem ser ativados no nó de modelagem e na guia Configurações no nugget do modelo. Esta funcionalidade estará disponível apenas quando o destino selecionado for um campo de flag. Consulte o tópico “Escores de Propensão” na página 36 para obter mais informações.

Calcular escores de propensão bruta. Os escores de propensão bruta são derivados do modelo com base apenas nos dados de treinamento. Se o modelo prever o valor *true* (responderá), então a propensão será a mesma que P, em que P é a probabilidade da predição. Se o modelo prever o valor false, então e propensão é calculada como (1-P).

- Se você escolher essa opção ao construir o modelo, os escores de propensão serão ativados no nugget do modelo por padrão. No entanto, sempre é possível optar por ativar os escores de propensão bruta no nugget do modelo independentemente se você selecioná-los no nó de modelagem ou não.
- Ao escorar o modelo, os escores de propensão bruta serão incluídos em um campo com as letras *RP* anexadas ao prefixo padrão. Por exemplo, se as predições estiverem em um campo denominado *$R-churn*, o nome do campo de escore de propensão será *$RRP-churn*.

Calcular escores de propensão ajustada. As propensões brutas baseiam-se puramente nas estimativas fornecidas pelo modelo, que podem ser super ajustadas e gerar estimativas de propensão super otimistas. As propensões ajustadas tentam compensar isso ao examinar como o modelo é executado nas partições de teste ou de validação e ajustar as propensões para fornecer uma estimativa melhor de acordo.

- Essa configuração requer que um campo de partição válido esteja presente no fluxo.
- Diferentemente dos escores de confiança bruta, os escores de propensão ajustada devem ser calculados ao construir o modelo; caso contrário, eles não estarão disponíveis quando escorar o nugget do modelo.
- Ao escorar o modelo, os escores de propensão ajustada serão incluídos em um campo com as letras *AP* anexadas ao prefixo padrão. Por exemplo, se as predições estiverem em um campo denominado *$R-churn*, o nome do campo de escore de propensão será *$RAP-churn*. Os escores de propensão ajustada não estão disponíveis para modelos de regressão logística.
- Ao calcular os escores de propensão ajustada, a partição de teste ou de validação utilizada para o cálculo não deverá ter sido balanceada. Para evitar isso, assegure-se de que a opção **Balancear somente dados de treinamento** esteja selecionada em qualquer nó Balanceamento de envio de dados. Além disso, se uma amostra complexa tiver sido obtida anteriormente, isto invalidará os escores de propensão ajustada.
- Os escores de propensão ajustada não estão disponíveis para modelos de árvore e de conjunto de regras "impulsionados". Consulte o tópico “Modelos do C5.0 Impulsionados” na página 118 para obter mais informações.

Nugget do Modelo KNN

O modelo KNN cria um número de novos campos, conforme mostrado na tabela a seguir. Para ver esses campos e seus valores, inclua um nó Tabela no nugget do modelo KNN e execute o nó de Tabela ou clique no botão Visualizar no nugget.

Tabela 29. Campos do modelo KNN

Novo nome de campo	Descrição
$KNN-*fieldname*	Valor predito do campo de destino.
$KNNP-*fieldname*	Probabilidade do valor predito.
$KNNP-*value*	Probabilidade de cada valor possível de um campo nominal ou de flag. Incluído apenas se **Incluir todas as probabilidades** estiver selecionada na guia Configurações do nugget do modelo.
$KNN-neighbor-*n*	O nome do *n*ésimo vizinho mais próximo do registro focal. Incluído apenas se **Exibir Mais Próximo** na guia Configurações do nugget do modelo estiver configurado com um valor diferente de zero.
$KNN-distance-*n*	A distância relativa do registro focal do *n*ésimo vizinho mais próximo ao registro focal. Incluído apenas se **Exibir Mais Próximo** na guia Configurações do nugget do modelo estiver configurado com um valor diferente de zero.

Visualização de Modelo de Vizinho Mais Próximo

Visualização do Modelo

A visualização do modelo possui uma janela de 2 painéis:

- O primeiro painel exibe uma visão geral do modelo chamado de visualização principal.
- O segundo painel exibe um dos dois tipos de visualizações:

  Uma visualização de modelo auxiliar mostra mais informações sobre o modelo, mas não está focada no próprio modelo.

  Uma visualização vinculada é uma visualização que mostra detalhes sobre uma variável do modelo quando o usuário realiza drill down na parte da visualização principal.

Por padrão, o primeiro painel mostra o espaço do preditor e o segundo painel mostra o gráfico de importância do preditor. Se o gráfico de importância do preditor não estiver disponível, ou seja, quando **Ponderar variáveis por importância** não for selecionada no painel Vizinhos da guia Configurações, a primeira visualização disponível na lista suspensa Visualizar é mostrada.

Quando uma visualização não possui nenhuma informação disponível, ela é omitida do menu suspenso Visualizar.

Espaço de Preditor: O gráfico de espaço do preditor é um gráfico interativo do espaço do preditor (ou um subespaço, se houver mais de 3 preditores). Cada eixo representa um preditor no modelo e o local dos pontos no gráfico mostra os valores destes preditores para casos nas partições de treinamento e de validação.

Chaves. Além dos valores do preditor, os pontos no gráfico transmitem outras informações.

- Forma que indica que a partição à qual um ponto pertence, seja Treinamento ou Validação.
- A cor/sombreamento de um ponto indica o valor da resposta para esse caso, com os valores de cores distintos iguais às categorias de uma variável resposta categórica e os sombreamentos indicando o intervalo de valores de uma variável resposta contínua. O valor indicado para a partição de treinamento é o valor observado e, para a partição de validação, é o valor predito. Se nenhuma resposta for especificada, essa chave não será mostrada.
- Estruturas de tópicos mais pesadas indicam que um caso é focal. Os registros focais são mostrados ligados aos seus *k* vizinhos mais próximos.

Controles e Interatividade. Diversos controles no gráfico permitem explorar o espaço de preditor.

- É possível escolher qual subconjunto de preditores deseja mostrar no gráfico e alterar quais preditores são representados nas dimensões.
- Os "Registros Focais" são simplesmente pontos selecionados no gráfico Espaço de Preditor. Se você especificou uma variável de registro focal, os pontos que representam os registros focais serão selecionados inicialmente. No entanto, qualquer ponto poderá se tornar temporariamente um registro focal se ele for selecionado. Os controles "comuns" para seleção de ponto se aplicam; clicar em um ponto seleciona esse ponto e desmarca todos os outros; Controle - Clicar em um ponto o inclui no conjunto de pontos selecionados. As visualizações vinculadas, como o Gráfico de Peers, serão atualizadas automaticamente com base nos casos selecionados no Espaço de Preditor.
- É possível alterar o número de vizinhos mais próximos (*k*) para exibir os registros focais.
- Passar o mouse sobre um ponto no gráfico exibe uma dica de ferramenta com o valor do rótulo case, ou o número do caso se rótulos case não forem definidos, e os valores de resposta observados e preditos.
- O botão "Reconfigurar" permite retornar o Espaço de Preditor para seu estado original.

Alterando os Eixos no Gráfico de Espaço de Preditor: É possível controlar quais variáveis são exibidas nos eixos do gráfico Espaço de Preditor.

Para alterar as configurações de eixo:

1. Clique no botão Modo de Edição (ícone de pincel) no painel esquerdo para selecionar o modo de Edição para o Espaço de Preditor.
2. Altere a visualização (para nada) no painel direito. O painel **Mostrar zonas** aparece entre os dois painéis principais.
3. Clique na caixa de seleção **Exibir zonas**.
4. Clique em qualquer ponto dos dados no Espaço de Preditor.
5. Para substituir um eixo por um preditor do mesmo tipo de dados:
   - Arraste o novo preditor sobre o rótulo da zona (aquele com o botão X pequeno) que você deseja substituir.

6. Para substituir um eixo com um preditor de um tipo de dados diferente:
   - No rótulo da zona do preditor você deseja substituir, clique no botão X pequeno. O espaço do preditor altera para uma visualização bidimensional.
   - Arraste o novo preditor sobre o rótulo da zona **Incluir dimensão**.
7. Clique no botão Modo de Exploração (ícone de ponta da seta) no painel esquerdo para sair do modo de Edição.

Importância do Preditor: Geralmente você desejará focar seus esforços de modelagem nos campos preditores que forem mais importantes e considerar descartar ou ignorar aqueles que forem menos importantes. O gráfico de importância do preditor ajuda a fazer isso ao indicar a importância relativa de cada preditor na estimativa do modelo. Como os valores são relativos, a soma dos valores para todos os preditores na tela é 1,0. A importância do preditor não tem relação com a precisão do modelo. Ela está relacionada apenas com a importância de cada preditor em fazer uma predição, e não se a predição é precisa ou não.

Distâncias de Vizinho Mais Próximo: Esta tabela exibe os *k* vizinhos mais próximos e as distâncias apenas dos registros focais. Ela estará disponível se um identificador de registro focal for especificado na Nó de modelagem e exibe apenas os registros focais identificados por essa variável.

Cada linha da:

- Coluna **Registro Focal** contém o valor da variável de rótulo case para o registro focal; se rótulos case não estiverem definidos, esta coluna conterá o número do caso do registro focal.
- *n*ésima coluna no grupo **Vizinhos Próximos** contém o valor da variável de rótulo case para o *n*ésimo vizinho mais próximo do registro focal. Se os rótulos case não forem definidos, esta coluna conterá o número do caso do *n*ésimo vizinho mais próximo do registro focal.
- *n*ésima coluna no grupo **Distâncias Mais Próximas** contém a distância do *n*ésimo vizinho mais próximo ao registro focal

Peers: Este gráfico exibe os casos focais e seus *k* vizinhos mais próximos em cada preditor e na resposta. Ele estará disponível se um caso focal for selecionado no Espaço de Preditor.

O gráfico de Peers vincula-se ao Espaço de Preditor de duas maneiras.

- Os casos selecionados (focais) no Espaço de Preditor são exibidos no gráfico de Peers, com seus *k* vizinhos mais próximos.
- O valor de *k* selecionado no Espaço de Preditor é utilizado no gráfico de Peers.

Selecionar Preditores. Permite selecionar os preditores para exibir no gráfico de Peers.

Mapa de Quadrante: Este gráfico exibe os casos focais e seus *k* vizinhos mais próximos em um gráfico de dispersão (ou gráfico de pontos, dependendo do nível de medição da resposta) com a resposta no eixo *y* e o preditor de escala no eixo *x*, agrupados em painéis pelos preditores. Ele estará disponível se houver uma resposta e se um caso focal for selecionado no Espaço de Preditor.

- As linhas de referência são desenhadas para variáveis contínuas, nas médias da variável na partição de treinamento.

Selecionar Preditores. Permite selecionar os preditores para exibir no Mapa de Quadrante.

Log de Erros de Seleção de Preditor: Os pontos no gráfico exibem o erro (a taxa de erros ou a soma dos quadrados dos erros, dependendo do nível de medição da resposta) no eixo *y* para o modelo com o preditor listado no eixo *x* (além de todas as variáveis à esquerda no eixo *x*). Este gráfico estará disponível se houver uma seleção de resposta e de variável em vigor.

Tabela de Classificação: Esta tabela exibe a classificação cruzada de valores observados versus valores preditos da resposta, por partição. Ela estará disponível se houver uma resposta e se for categórica (flag, nominal ou ordinal).

- A linha **(Omisso)** na partição Validação contém casos de validação com valores omissos na resposta. Estes casos contribuem com a Amostra de Validação: Valores de Porcentagem Geral, mas não para valores de Porcentagem Correta.

Sumarização de Erro: Esta tabela estará disponível se houver uma variável de destino. Ela exibe o erro associado ao modelo; a soma dos quadrados para uma variável resposta contínua e a taxa de erro (100% – percentual geral correto) para uma variável resposta categórica.

Configurações do Modelo KNN

A guia Configurações permite especificar campos extras a serem exibidos quando visualizar os resultados (por exemplo, ao executar um nó Tabela anexado ao nugget). É possível ver o efeito de cada uma dessas opções ao selecioná-las e clicar no botão Visualizar -- role para a direita da saída Visualizar para ver os campos extras.

Incluir todas as probabilidades (válido apenas para variáveis resposta categóricas). Se essa opção for selecionada, as probabilidades para cada valor possível de um campo de destino nominal ou de flag são exibidas para cada registro processado pelo nó. Se essa opção estiver desmarcada, apenas o valor predito e sua probabilidade serão exibidos para campos de destino nominal ou de flag.

A configuração padrão dessa caixa de seleção é determinada pela caixa de seleção correspondente no nó de modelagem.

Calcular escores de propensão bruta. Para modelos com uma resposta de flag (que retornam uma predição de sim ou não), é possível solicitar os escores de propensão que indicam a probabilidade do resultado real especificado para o campo de destino. Esses são um complemento dos outros valores de predição e de confiança que podem ser gerados durante a escoragem.

Calcular escores de propensão ajustada. Os escores de propensão bruta baseiam-se apenas nos dados de treinamento e esses podem ser altamente otimistas devido à tendência de muitos modelos a super ajustar desses dados. As propensões ajustadas tentam compensar ao avaliar o desempenho do modelo com relação à partição de teste ou de validação. Essa opção requer que um campo de partição seja definido no fluxo e que os escores de propensão ajustada sejam ativados no nó de modelagem antes de gerar o modelo.

Exibir mais próximo. Se você configurar este valor para *n*, em que *n* é um número inteiro positivo diferente de zero, os *n* vizinhos mais próximos ao registro focal serão incluídos no modelo, com suas distâncias relativas do registro focal.

Avisos

Estas informações foram desenvolvidas para produtos e serviços oferecidos no mundo todo.

É possível que a IBM não ofereça os produtos, serviços ou recursos discutidos nesta publicação em outros países. Consulte um representante IBM local para obter informações sobre produtos e serviços disponíveis atualmente em sua área. Qualquer referência a produtos, programas ou serviços IBM não significa que apenas produtos, programas ou serviços IBM possam ser usados. Qualquer produto, programa ou serviço funcionalmente equivalente, que não infrinja nenhum direito de propriedade intelectual da IBM poderá ser usado em substituição. Entretanto, a avaliação e verificação da operação de qualquer produto, programa ou serviço não IBM são de responsabilidade do Cliente.

A IBM pode ter patentes ou solicitações de patentes relativas a assuntos tratados nesta publicação. O fornecimento desta publicação não lhe garante direito algum sobre tais patentes. Pedidos de licença devem ser enviados, por escrito, para:

Gerência de Relações Comerciais e Industriais da IBM Brasil
IBM Corporation
Av. Pasteur, 138-146, Botafogo
Rio de Janeiro. RJ
CEP 22290-240
Brasil

Para pedidos de licença relacionados a informações de byte duplo (DBCS), entre em contato com o Departamento de Propriedade Intelectual da IBM em seu país ou envie pedidos de licença, por escrito, para:

Licença de Propriedade Intelectual
Lei de Propriedade Intelectual
IBM Japan Ltd.
1623-14, Shimotsuruma, Yamato-shi
Kanagawa 242-8502 Japan

O parágrafo a seguir não se aplica a nenhum país em que tais disposições não estejam de acordo com a legislação local: A INTERNATIONAL BUSINESS MACHINES CORPORATION FORNECE ESTA PUBLICAÇÃO "NO ESTADO EM QUE SE ENCONTRA", SEM GARANTIA DE NENHUM TIPO, SEJA EXPRESSA OU IMPLÍCITA, INCLUINDO, MAS A ELAS NÃO SE LIMITANDO, AS GARANTIAS IMPLÍCITAS DE NÃO-INFRAÇÃO, COMERCIALIZAÇÃO OU ADEQUAÇÃO A UM DETERMINADO PROPÓSITO. Alguns países não permitem a exclusão de garantias expressas ou implícitas em determinadas transações, portanto, essa disposição pode não se aplicar ao Cliente.

Essas informações podem conter imprecisões técnicas ou erros tipográficos. São feitas alterações periódicas nas informações aqui contidas; tais alterações serão incorporadas em futuras edições desta publicação. A IBM pode, a qualquer momento, aperfeiçoar e/ou alterar os produtos e/ou programas descritos nesta publicação, sem aviso prévio.

As referências nestas informações a websites não IBM são fornecidas apenas por conveniência e não representam de forma alguma um endosso a estes websites. Os materiais contidos nesses Web sites não fazem parte dos materiais deste produto IBM e a utilização desses Web sites é de inteira responsabilidade do Cliente.

A IBM pode usar ou distribuir as informações fornecidas da forma que julgar apropriada sem incorrer em qualquer obrigação para com o Cliente.

Licenciados deste programa que desejam obter informações sobre este assunto com objetivo de permitir: (i) a troca de informações entre programas criados independentemente e outros programas (incluindo este) e (ii) a utilização mútua das informações trocadas, devem entrar em contato com:

IBM Software Group
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
CEP 22290-240

Tais informações podem estar disponíveis, sujeitas a termos e condições apropriadas, incluindo em alguns casos o pagamento de uma taxa.

O programa licenciado descrito nesta publicação e todo o material licenciado disponível são fornecidos pela IBM sob os termos do IBM Customer Agreement, Contrato de Licença do Programa Internacional IBM ou qualquer outro contrato equivalente.

Quaisquer dados de desempenho aqui contidos foram determinados em um ambiente controlado. Portanto, os resultados obtidos em outros ambientes operacionais podem variar significativamente. Algumas medidas podem ter sido tomadas em sistemas em nível de desenvolvimento e não há garantia de que estas medidas serão as mesmas em sistemas geralmente disponíveis. Além disso, algumas medidas podem ter sido estimadas através de extrapolação. Os resultados reais podem variar. Os usuários desta publicação devem verificar os dados aplicáveis para seu ambiente específico.

As informações relacionadas a produtos não IBM foram obtidas junto aos fornecedores destes produtos, de seus anúncios publicados ou de outras fontes disponíveis publicamente. A IBM não testou estes produtos e não pode confirmar a precisão de desempenho, compatibilidade nem qualquer outra reivindicação relacionada a produtos não IBM. Perguntas sobre os recursos de produtos não IBM devem ser encaminhadas aos fornecedores desses produtos.

Todas as instruções relativas aos objetivos ou intenção futura da IBM estão sujeitas a alterações ou cancelamento sem aviso prévio e representam apenas metas e objetivos.

Essas informações contêm exemplos de dados e relatórios usados em operações diárias de negócios. Para ilustrá-los da forma mais completa possível, os exemplos incluem nomes de indivíduos, empresas, marcas e produtos. Todos estes nomes são fictícios e qualquer semelhança com nomes e endereços usados por uma empresa real é mera coincidência.

Se essas informações estiverem sendo exibidas em formato eletrônico, as fotografias e ilustrações coloridas poderão não aparecer.

Marcas comerciais

IBM, o logotipo IBM e ibm.com são marcas comerciais ou marcas registradas da International Business Machines Corp., registradas em muitas jurisdições em todo o mundo. Outros nomes de produtos e serviços podem ser marcas comerciais da IBM ou de outras empresas. Uma lista atual de marcas comerciais da IBM trademarks está disponível na web em "Informações de Copyright e marcas comerciais" em www.ibm.com/legal/copytrade.shtml.

Intel, logotipo Intel, Intel Inside, logotipo Intel Inside, Intel Centrino, logotipo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium e Pentium são marcas registradas ou marcas comerciais da Intel Corporation ou suas subsidiárias nos Estados Unidos e outros países.

Linux é uma marca registrada da Linus Torvalds nos Estados Unidos e/ou em outros países.

Microsoft, Windows, Windows NT e o logotipo Windows são marcas comerciais da Microsoft Corporation nos Estados Unidos e/ou em outros países.

UNIX é uma marca registrada do The Open Group nos Estados Unidos e em outros países.

Java e todas as marcas e logotipos baseados em Java são marcas comerciais ou marcas registradas da Oracle e/ou suas afiliadas.

Outros nomes de produtos e serviços podem ser marcas comerciais da IBM ou de outras empresa.

Glossário

A

AICC . Uma medida para selecionar e comparar modelos combinados com base no log da verossimilhança -2 (Restrito). Valores menores indicam melhores modelos. O AICC "corrige" o AIC para tamanhos de amostra pequenos. Conforme o tamanho da amostra aumenta, o AICC converge para o AIC.

B

Critério de Informação Bayesiano (BIC) . Uma medida para selecionar e comparar modelos com base no log da verossimilhança -2. Valores menores indicam melhores modelos. O BIC também "penaliza" modelos sobreparametrizados (por exemplo, modelos complexos com um número grande de entradas), mas é mais rígido do que o AIC.

teste M de Box . Um teste para a igualdade das matrizes de covariâncias de grupo. Para amostras suficientemente grandes, um valor p não significativo representa que não há evidência suficiente de que as matrizes diferem. O teste é sensível a partidas da normalidade multivariada.

C

Casos . Os códigos para o grupo real, o grupo predito, as probabilidades posteriores e os escores discriminantes são exibidos para cada caso.

Resultados da Classificação . O número de casos designados correta e incorretamente para cada um dos grupos com base na análise discriminante. Às vezes é chamado de "Matriz de Confusão".

Gráficos de Grupos Combinados . Cria um gráfico de dispersão de todos os grupos dos dois primeiros valores da função discriminante. Se houver apenas uma função, um histograma será exibido.

Covariance . Uma medida de associação não padronizada entre duas variáveis, igual ao desvio de produto vetorial dividido por N-1.

F

de Fisher . Exibe os coeficientes da função de classificação de Fisher que podem ser utilizados diretamente para classificação. Um conjunto separado de coeficientes de função de classificação é obtido para cada grupo, e um caso é designado ao grupo para o qual ele possui o maior escore discriminante (valor da função de classificação).

H

Gráfico de Risco . Exibe a função de risco acumulativo em uma escala linear.

K

Curtose . Uma medida da extensão até a qual as observações se agrupam em torno de um ponto central. Para a uma distribuição normal, o valor da estatística de curtose é zero. A curtose positiva indica que, com relação a uma distribuição normal, as observações são agrupadas mais ao centro da distribuição e têm rastros mais finos até os valores extremos da distribuição, no ponto em que os rastros da distribuição leptocúrtica são mais espessos com relação a uma distribuição normal. Já a curtose negativa indica que, com relação a uma distribuição normal, as observações são menos agrupadas e possuem rodapés mais espessos até os valores extremos da distribuição, no ponto em que os rastros da distribuição platykurtic são mais finos com relação a uma distribuição normal.

L

Classificação de Exclusão de um Item . Cada caso na análise é classificado pelas funções derivadas de todos os outros casos diferentes desse caso. Também é conhecida como "Método U".

M

MAE . Erro médio absoluto. Medidas do quanto a série varia a partir de seu nível predito pelo modelo. O MAE é relatado nas unidades da série original.

Distância de Mahalanobis . A medida do quanto os valores de um caso nas variáveis independentes diferem da média de todos os casos. Uma distância de Mahalanobis grande identifica um caso como tendo valores extremos em uma ou mais variáveis independentes.

MAPE . Erro Percentual Absoluto Médio. Uma medida do quanto uma série dependente varia de seu nível predito pelo modelo. Ela é independente das unidades utilizadas e pode, portanto, ser utilizada para comparar séries com unidades diferentes.

MaxAE . Erro Máximo Absoluto. O maior erro previsto, expresso nas mesmas unidades que a série dependente. Assim como o MaxAPE, é útil para imaginar o cenário do pior caso para suas previsões. O erro absoluto máximo e o erro percentual absoluto máximo poderão ocorrer em diferentes pontos da série - por exemplo, quando um erro absoluto de um valor de série grande for um pouco maior que o erro absoluto de um valor de série pequeno. Nesse caso, o erro absoluto máximo ocorrerá no valor de série maior e o erro percentual absoluto máximo ocorrerá no valor de série menor.

MaxAPE . Erro Percentual Absoluto Máximo.O maior erro previsto, expresso como uma porcentagem. Esta medida é útil para imaginar um cenário do pior caso para suas previsões.

Maximização do Método de Entrada da Menor Razão F . Um método de seleção de variáveis na análise stepwise com base na maximização de uma razão F calculada a partir da distância de Mahalanobis entre os grupos.

Máximo . O maior valor de uma variável numérica.

Média . Uma medida de tendência central. A média aritmética, a soma dividida pelo número de casos.

Médias . Exibe as médias totais e de grupo, bem como os desvios padrão para as variáveis independentes.

Median . O valor acima e abaixo do qual metade dos casos cai, o 50º percentil. Se houver um número par de casos, a mediana será a média dos dois casos intermediários quando forem classificados em ordem crescente ou decrescente. A mediana é uma medida da tendência central não sensível a valores excessivos (ao contrário da média, que pode ser afetada por alguns valores extremamente altos ou baixos).

Minimizar Lambda de Wilks . Um método de seleção de variáveis para análise discriminante stepwise que escolhe variáveis a serem inseridas na equação com base no quanto elas diminuem o lambda de Wilks. Em cada passo, a variável que minimiza o lambda geral de Wilks é inserida.

Mínimo . O menor valor de uma variável numérica.

Modo . O valor que ocorre mais frequentemente. Se vários valores compartilharem a maior frequência da ocorrência, cada um deles será um modo.

N

BIC normalizado . Critério de Informação Bayesiano Normalizado. Uma medida geral do ajuste geral de um modelo que tenta considerar a complexidade do modelo. É um escore com base no erro quadrático médio e inclui uma penalidade para o número de parâmetros no modelo e no comprimento da série. A penalidade remove a vantagem de modelos com mais parâmetros, tornando a estatística fácil de comparar entre diferentes modelos da mesma série.

O

Sobrevivência de Um Menos . Exibe a função um menos a sobrevivência em uma escala linear.

R

Intervalo . A diferença entre os valores maior e menor de uma variável numérica, o máximo menos o mínimo.

V de Rao (Análise Espectral) . A medida das diferenças entre as médias de grupo. Também chamada de rastreio de Lawley-Hotelling. Em cada passo, a variável que maximiza o aumento no V de Rao é inserida. Após selecionar esta opção, insira o valor mínimo que uma variável deve ter para inserir na análise.

RMSE . Erro Quadrático Médio Raiz. A raiz quadrada do erro quadrático médio. Uma medida de quanto uma série dependente varia a partir de seu nível predito pelo modelo, expressa nas mesmas unidades que a série dependente.

R-quadrado . Medida de qualidade de ajuste de um modelo linear, às vezes chamada de coeficiente de determinação. É a proporção de variação na variável dependente explicada pelo modelo de regressão. O valor varia de 0 a 1. Valores pequenos indicam que o modelo não ajusta bem os dados.

S

Grupos-Separados . As matrizes de covariância grupos-separados são usadas para classificação. Como a classificação é baseada nas funções discriminantes (não com base nas variáveis originais), essa opção nem sempre é equivalente à discriminação quadrática.

Covariância de Grupos-Separados . Exibe matrizes de covariância separadas para cada grupo.

Gráficos de Grupos-Separados . Cria gráficos de dispersão de grupos separados dos dois primeiros valores da função discriminante. Se houver apenas uma função, histogramas serão exibidos.

Bonferroni Sequencial . Este é um procedimento de Bonferroni de rejeição sequencialmente decrescente que tende ser muito menos conservador em termos de rejeição de hipóteses individuais, mas mantém o mesmo nível de significância geral.

Sidak Sequencial . Este é um procedimento de Sidak de rejeição sequencialmente decrescente que tende ser muito menos conservador em termos de rejeição de hipóteses individuais, mas mantém o mesmo nível de significância geral.

Assimetria . Uma medida da assimetria de uma distribuição. A distribuição normal é simétrica e tem um valor de assimetria de 0. Uma distribuição com assimetria positiva significativa tem um longo rodapé direito. Uma distribuição com assimetria negativa significativa tem um longo rodapé esquerdo. Como uma orientação, um valor de assimetria mais de duas vezes seu erro padrão é obtido para indicar uma partida de simetria.

desvio padrão . Uma medida de dispersão em torno da média, igual à raiz quadrada da variância. O desvio padrão é medido nas mesmas unidades que a variável original.

Desvio padrão . Uma medida de dispersão em torno da média. Em uma distribuição normal, 68% dos casos estão dentro de um desvio padrão da média e 95% dos casos estão dentro de dois desvios padrão. Por exemplo, se a idade média for 45, com um desvio padrão de 10, 95% dos casos estariam entre 25 e 65 em uma distribuição normal.

Erro padrão . Uma medida do quanto o valor de uma estatística de teste varia de amostra para amostra. Ela é o desvio padrão da distribuição de amostragem para uma estatística. Por exemplo, o erro padrão da média é o desvio padrão das médias da amostra.

Erro Padrão de Curtose . A razão de curtose com seu erro padrão pode ser utilizada como um teste de normalidade (ou seja, é possível rejeitar a normalidade se a razão for inferior a -2 ou superior a +2). Um valor positivo grande para curtose indica que os rodapés da distribuição são maiores do que aqueles de uma distribuição normal, e um valor negativo para curtose indica rodapés mais curtos (como aqueles de uma distribuição uniforme em forma de caixa).

Erro padrão de média . Uma medida do quanto o valor da média pode variar de amostra para amostra obtida da mesma distribuição. Ela pode ser utilizada para comparar aproximadamente a média observada com um valor hipotético (ou seja, é possível concluir que os dois valores serão diferentes se a razão da diferença com o erro padrão for inferior a -2 ou superior a +2).

Erro Padrão de Assimetria . A razão de assimetria com seu erro padrão pode ser utilizada como um teste de normalidade (ou seja, é possível rejeitar a normalidade se a razão for inferior a -2 ou superior a +2). Um valor positivo grande para assimetria indica um rodapé direito longo, e um valor negativo extremo indica um rodapé esquerdo longo.

R-quadrado estacionário . Uma medida que compara a parte estacionária do modelo com um modelo de média simples. Esta medida é preferível ao R-quadrado ordinário quando houver um padrão de tendência ou sazonal. O R-quadrado estacionário pode ser negativo com um intervalo de infinito negativo para 1. Valores negativos significam que o modelo sob consideração é pior do que o modelo de linha de base. Valores positivos significam que o modelo sob consideração é melhor que o modelo de linha de base.

Sum . A soma ou o total dos valores, em todos os casos com valores não omissos.

Gráfico de Sobrevivência . Exibe a função de sobrevivência acumulativa em uma escala linear.

X

Mapa Territorial . Um gráfico dos limites utilizados para classificar os casos em grupos com base nos valores de função. Os números correspondem aos grupos nos quais os casos são classificados. A média de cada grupo é indicada por um asterisco dentro de seus limites. O mapa não será exibido se houver apenas uma função discriminante.

Covariância Total . Exibe uma matriz de covariâncias a partir de todos os casos como se fossem de uma única amostra.

U

Variância Não Explicada . Em cada passo, a variável que minimiza a soma da variação não explicada entre os grupos é inserida.

unique . Avalia todos os efeitos simultaneamente, ajuste cada efeito a todos os outros efeitos de qualquer tipo.

ANOVAs Univariadas . Executa um teste de análise de variância unidirecional de igualdade de médias de grupo para cada variável independente.

Não padronizado . Exibe os coeficientes de função discriminante não padronizadas.

Usar Valor F . Uma variável será inserida no modelo se seu valor F for maior que o valor de Entrada e será removida se o valor F for menor que o valor de Remoção. A Entrada deve ser maior que Remoção, e ambos os valores devem ser positivos. Para inserir mais variáveis no modelo, diminua o valor de Entrada. Para remover mais variáveis do modelo, aumente o valor de Remoção.

Usar Probabilidade de F . Uma variável será inserida no modelo se o nível de significância de seu valor F for menor que o valor de Entrada e será removida se o nível de significância for maior que o valor de Remoção. A Entrada deve ser menor que Remoção, e ambos os valores devem ser positivos. Para inserir mais variáveis no modelo, aumente o valor de Entrada. Para remover mais variáveis do modelo, diminua o valor de Remoção.

V

Válidos . Casos válidos que não possuem o valor omisso do sistema, nem um valor definido como omisso de usuário.

Variância . Uma medida de dispersão ao redor da média, igual à soma dos desvios quadrados da média dividido por um menor que o número de casos. A variância é medida em unidades que são o quadrado daquelas da própria variável.

D

Dentro de Grupos . A matriz de covariâncias dentro de grupos em conjunto é usada para classificar casos.

Correlação Dentro de Grupos . Exibe uma matriz de correlações dentro de grupos em conjunto que é obtida pela média das matrizes de covariâncias separadas de todos os grupos antes de calcular as correlações.

Covariância Dentro de Grupos . Exibe uma matriz de covariâncias dentro de grupos em conjunto, que pode diferir da matriz de covariâncias totais. A matriz é obtida pela média das matrizes covariâncias separadas para todos os grupos.

Índice Remissivo

E

F

G

H

I

K

L

M

N

O

P

Q

R

S

T

V

Impresso no Brasil