Guide d'applications de IBM SPSS Modeler 15

Remarque : Avant d'utiliser ces informations et le produit qu'elles concernent, lisez les informations générales sous Remarques sur p. .

Cette version s'applique à IBM SPSS Modeler 15 et à toutes les publications et modifications ultérieures jusqu'à mention contraire dans les nouvelles versions.

Les captures d'écran des produits Adobe sont reproduites avec l'autorisation de Adobe Systems Incorporated.

Les captures d'écran des produits Microsoft sont reproduites avec l'autorisation de Microsoft Corporation.

Matériel sous licence - Propriété d'IBM

Préface

IBM® SPSS® Modeler est le puissant utilitaire de Data mining de IBM Corp.. SPSS Modeler aide les entreprises et les organismes à améliorer leurs relations avec les clients et les citoyens grâce à une compréhension approfondie des données. A l'aide des connaissances plus précises obtenues par le biais de SPSS Modeler, les entreprises et les organismes peuvent conserver les clients rentables, identifier les opportunités de vente croisée, attirer de nouveaux clients, détecter les éventuelles fraudes, réduire les risques et améliorer les services gouvernementaux.

L'interface visuelle de SPSS Modeler met à contribution les compétences professionnelles de l'utilisateur, ce qui permet d'obtenir des modèles prédictifs plus efficaces et de trouver des solutions plus rapidement. SPSS Modeler dispose de nombreuses techniques de modélisation, telles que les algorithmes de prévision, de classification, de segmentation et de détection d'association. Une fois les modèles créés, l'utilisateur peut utiliser IBM® SPSS® Modeler Solution Publisher pour les remettre aux responsables, où qu'ils se trouvent dans l'entreprise, ou pour les transférer vers une base de données.

A propos de IBM Business Analytics

Le logiciel IBM Business Analytics fournit des informations complètes, cohérentes et précises que les preneurs de décision utilisent avec confiance pour améliorer la performance du marché. Un portefeuille étendu d'outils de business intelligence, d'analyses prédictives, de performance financière et de gestion de stratégie, et des applications analytiques offre des connaissances claires, immédiates et appliquables pour améliorer l'efficacité actuelle ainsi que la capacité de prévoir les résultats futurs. Combinées avec de riches solutions industrielles, des pratiques éprouvées et des services professionnels, les organisations de toutes tailles peuvent atteindre la productivité la plus élevée, automatiser des décisions en toute tranquilité et fournir de meilleurs résultats.

Dans le cadre de ce portefeuille, le logiciel IBM SPSS Predictive Analytics aide les organisations à prévoir des évènements futurs et à agir en conséquence pour mener à de meilleurs résultats Des clients dans le domaine commercial, gouvernemental et académique à travers le monde font confiance à la technologie IBM SPSS et considèrent qu'elle représente un avantage compétitif pour attirer, retenir et ajouter des clients, tout en réduisant la fraude et en atténuant les risques. En incorporant le logiciel IBM SPSS dans leur opérations quotidiennes, les organisations deviennent des entreprises prédictives – capables de diriger et d'automatiser les décisions pour atteindre les buts qu'ils se sont fixés et obtenir des avantages compétitifs sensibles. Pour informations supplémentaires ou pour joindre un revendeur, visitez le site *http://www.ibm.com/spss*.

Assistance technique

L'assistance technique est à la disposition des clients pour la maintenance des produits. Les clients peuvent contacter l'assistance technique pour obtenir de l'aide concernant l'utilisation des produits IBM Corp. ou l'installation dans l'un des environnements matériels pris en charge. Pour joindre l'assistance technique, consultez le site Web de IBM Corp. à l'adresse *http://www.ibm.com/support*. Lorsque vous contactez l'assistance technique, n'oubliez pas de préparer vos identifiants, le nom de votre société et votre contrat d'assistance.

Contenu

Partie III: Exemples de modélisation

Annexe

Chapitre
1

A propos de IBM SPSS Modeler

À propos de IBM SPSS Modeler

IBM® SPSS® Modeler est un ensemble d'outils de data mining qui vous permet de développer rapidement, grâce à vos compétences professionnelles, des modèles prédictifs et de les déployer dans des applications professionnelles afin de faciliter la prise de décision. Conçu autour d'un modèle confirmé, le modèle CRISP-DM, SPSS Modeler prend en charge l'intégralité du processus de Data mining, des données à l'obtention de meilleurs résultats commerciaux.

SPSS Modeler propose différentes méthodes de modélisation issues des domaines de l'apprentissage automatique, de l'intelligence artificielle et des statistiques. Les méthodes disponibles dans la palette Modélisation vous permettent d'extraire de nouvelles informations de vos données et de développer des modèles prédictifs. Chaque méthode possède ses propres avantages et est donc plus adaptée à certains types de problème spécifiques.

Il est possible d'acquérir SPSS Modeler comme produit autonome ou de l'utiliser en tant que client en combinaison avec SPSS Modeler Server. Plusieurs autres options sont également disponibles, telles que décrites dans les sections suivantes. Pour plus d'informations, consultez *http://www.ibm.com/software/analytics/spss/products/modeler/*.

Produits IBM SPSS Modeler

La famille des produits IBM® SPSS® Modeler et les logiciels associés sont composés des éléments suivants.

- IBM SPSS Modeler
- IBM SPSS Modeler Server
- IBM SPSS Modeler Administration Console
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- IBM SPSS Modeler Server adaptateurs pour IBM SPSS Collaboration and Deployment Services

IBM SPSS Modeler

SPSS Modeler est une version complète du produit que vous installez et exécutez sur votre ordinateur personnel. Pour obtenir de meilleures performances lors du traitement d'ensembles de données volumineux, vous pouvez exécuter SPSS Modeler en mode local, comme produit autonome, ou l'utiliser en mode réparti, en association avec IBM® SPSS® Modeler Server.

Avec SPSS Modeler, vous pouvez créer des modèles prédictifs précis rapidement et de manière intuitive, sans aucune programmation. L'interface visuelle unique vous permet de visualiser facilement le processus de Data mining. Grâce aux analyses avancées intégrées au produit, vous pouvez découvrir des motifs et tendances masqués dans vos données. Vous pouvez modéliser les résultats et comprendre les facteurs qui les influencent, afin d'exploiter les opportunités commerciales et de réduire les risques.

SPSS Modeler est disponible en deux éditions : SPSS Modeler Professional et SPSS Modeler Premium. Pour plus d'informations, reportez-vous à la section Éditions de IBM SPSS Modeler dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.

IBM SPSS Modeler Server

Grâce à une architecture client/serveur, SPSS Modeler adresse les demandes d'opérations très consommatrices de ressources à un logiciel serveur puissant. Il offre ainsi des performances accrues sur des ensembles de données plus volumineux.

SPSS Modeler Server est un produit avec licence distincte qui s'exécute en permanence en mode d'analyse réparti sur un hôte de serveur en combinaison avec une ou plusieurs installations de IBM® SPSS® Modeler. Ainsi, SPSS Modeler Server fournit des performances supérieures sur de grands ensembles de données car les opérations nécessitant beaucoup de mémoire peuvent être effectuées sur le serveur sans télécharger de données sur l'ordinateur client. IBM® SPSS® Modeler Server prend également en charge l'optimisation SQL et propose des fonctionnalités de modélisation dans la base de données pour des performances et une automatisation améliorées.

IBM SPSS Modeler Administration Console

Le Modeler Administration Console est une application graphique permettant de gérer de nombreuses options de SPSS Modeler Server qui peuvent également être configurées au moyen d'un fichier d'options. Cette application offre une interface utilisateur sous forme de console permettant de surveiller et de configurer les installations SPSS Modeler Server ; elle est disponible gratuitement pour les clients actuels de SPSS Modeler Server. L'application ne peut être installée que sur des ordinateurs Windows ; en revanche, elle peut administrer un serveur installé sur n'importe quelle plate-forme prise en charge.

IBM SPSS Modeler Batch

Alors que le Data mining est généralement un processus interactif, il est également possible d'exécuter SPSS Modeler à partir d'une ligne de commande sans recourir à l'interface utilisateur graphique. Par exemple, vous pouvez avoir des tâches longue durée ou répétitives à exécuter sans intervention de l'utilisateur. SPSS Modeler Batch est une version spécifique du produit qui prend en charge toutes les fonctions d'analyse de SPSS Modeler sans avoir besoin d'accéder à l'interface utilisateur standard. Une licence SPSS Modeler Server est nécessaire pour utiliser SPSS Modeler Batch.

IBM SPSS Modeler Solution Publisher

SPSS Modeler Solution Publisher est un outil qui permet de créer une version « packagée » d'un flux SPSS Modeler qui peut être exécutée par un moteur Runtime externe ou intégrée dans une application externe. Ainsi, vous pouvez publier et déployer des flux SPSS Modeler complets dans des environnements où SPSS Modeler n'est pas installé. SPSS Modeler Solution Publisher est fourni avec le service IBM SPSS Collaboration and Deployment Services - Scoring et nécessite une licence distincte. Avec cette licence, vous recevez SPSS Modeler Solution Publisher Runtime qui vous permet d'exécuter les flux publiés.

IBM SPSS Modeler Server Adaptateurs pour IBM SPSS Collaboration and Deployment Services

Différents adaptateurs pour IBM® SPSS® Collaboration and Deployment Services sont disponibles et permettent à SPSS Modeler et SPSS Modeler Server d'interagir avec un référentiel IBM SPSS Collaboration and Deployment Services. Ainsi, un flux SPSS Modeler déployé sur le référentiel peut être partagé par différents utilisateurs ou peut être accessible depuis l'application client léger IBM SPSS Modeler Advantage. Installez l'adaptateur sur le système qui héberge le référentiel.

Éditions de IBM SPSS Modeler

SPSS Modeler est disponible dans les éditions suivantes.

SPSS Modeler Professional

SPSS Modeler Professional offre tous les outils nécessaires à l'utilisation de la plupart des types de données structurées, tels que les comportements et interactions suivis dans les systèmes CRM, les caractéristiques sociodémographiques, les comportements d'achat et les données de vente.

SPSS Modeler Premium

SPSS Modeler Premium est un produit avec licence distincte qui étend le champ d'applications de SPSS Modeler Professional afin de pouvoir traiter des données spécialisées telles que celles utilisées pour les analyses d'entités ou les réseaux sociaux ainsi que des données de texte non structurées. SPSS Modeler Premium comprend les composants suivants :

IBM® SPSS® Modeler Entity Analytics ajoute une dimension entièrement nouvelle aux analyses prédictives IBM® SPSS® Modeler. Alors que les analyses prédictives essaient de prévoir les comportements futurs à partir de données passées, les analyses d'entités se concentrent sur l'amélioration de la cohérence des données actuelles en résolvant les conflits d'identités dans les enregistrements eux-mêmes. Une identité peut être celle d'un individu, d'une organisation, d'un objet ou d'une autre entité pour laquelle une ambiguïté peut exister. La résolution d'identité peut être vitale dans de nombreux domaines, y compris la gestion de la relation client, la détection de la fraude, le blanchiment d'argent et la sécurité nationale et internationale.

IBM SPSS Modeler Social Network Analysis transforme les informations sur les relations en champs qui caractérisent le comportement social des individus et des groupes. Grâce aux données qui décrivent les relations qui sous-tendent les réseaux sociaux, IBM® SPSS® Modeler Social Network Analysis identifie les chefs sociaux qui influencent le comportement des autres individus du réseau. De plus, il est possible de déterminer les individus qui sont le plus influencés par les autres participants du réseau. En combinant ces résultats avec d'autres mesures, il est possible de créer des profils détaillés des individus sur lesquels baser vos modèles prédictifs. Les modèles qui contiennent ces informations sociales seront plus efficaces que les modèles qui en sont dépourvus.

Text Analytics for IBM® SPSS® Modeler utilise des technologies linguistiques avancées et le traitement du langage naturel pour traiter rapidement une large variété de données textuelles non structurées, en extraire les concepts clés et les organiser pour les regrouper dans des catégories. Les concepts extraits et les catégories peuvent ensuite être combinés aux données structurées existantes, telles que les données démographiques, et appliqués à la modélisation grâce à la gamme complète d'outils de Data mining de SPSS Modeler, afin de favoriser une prise de décision précise et efficace.

Documentation de IBM SPSS Modeler

Une documentation au format d'aide en ligne est disponible dans le menu Aide de SPSS Modeler. Vous y trouverez la documentation de SPSS Modeler, SPSS Modeler Server et de SPSS Modeler Solution Publisher, ainsi que le Guide des applications et d'autres documentations utiles.

La documentation complète de chaque produit (y compris les instructions d'installation) au format PDF est disponible dans le dossier *\Documentation* de chaque DVD de produit. Ces documents d'installation peuvent également être téléchargés sur Internet à l'adresse *http://www-01.ibm.com/support/docview.wss?uid=swg27023172*.

La documentation dans les deux formats est également disponible depuis le Centre d'informations SPSS Modeler à l'adresse *http://publib.boulder.ibm.com/infocenter/spssmodl/v15r0m0/*.

Documentation de SPSS Modeler Professional

La suite de documentation SPSS Modeler Professional (à l'exception des instructions d'installation) est la suivante.

- **Guide de l'utilisateur IBM SPSS Modeler.** Introduction générale à SPSS Modeler : création de flux de données, traitement des valeurs manquantes, création d'expressions CLEM, utilisation des projets et des rapports et regroupement des flux pour le déploiement dans IBM SPSS Collaboration and Deployment Services, des applications prédictives ou IBM SPSS Modeler Advantage.
- **Noeuds de Source, d'exécution et de sortie IBM SPSS Modeler.** Descriptions de tous les noeuds utilisés pour lire, traiter et renvoyer les données de sortie dans différents formats. En pratique, cela signifie tous les noeuds autres que les noeuds de modélisation.
- **IBM SPSS Modeler Noeuds de modélisation.** Description de tous les noeuds utilisés pour créer des modèles de Data mining. IBM® SPSS® Modeler propose différentes méthodes de modélisation issues des domaines de l'apprentissage automatique, de l'intelligence artificielle

et des statistiques. Pour plus d'informations, reportez-vous à la section Description des noeuds de modélisation dans le chapitre 3 dans *Noeuds de modélisation de IBM SPSS Modeler 15*.

- **Guide des Algorithmes IBM SPSS Modeler.** Descriptions des fondements mathématiques des méthodes de modélisation utilisées dans SPSS Modeler. Ce guide est disponible au format PDF uniquement.
- **Guide des applications IBM SPSS Modeler.** Les exemples de ce guide fournissent des introductions brèves et ciblées aux méthodes et techniques de modélisation. Un version en ligne de ce guide est également disponible dans le menu Aide. Pour plus d'informations, reportez-vous à la section Exemples d'application dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.
- **Génération de scripts et automatisation IBM SPSS Modeler.** Informations sur l'automatisation du système via la génération de scripts, y compris les propriétés permettant de manipuler les noeuds et les flux.
- **IBM SPSS Modeler Guide de déploiement.** Informations sur l'exécution des scénarios et des flux SPSS Modeler comme étapes des tâches d'exécution sous IBM® SPSS® Collaboration and Deployment Services Deployment Manager.
- **IBM SPSS Modeler CLEF Guide du développeur.** CLEF permet d'intégrer des programmes tiers tels que des programmes de traitement de données ou des algorithmes de modélisation en tant que noeuds dans SPSS Modeler.
- **Guide d'exploration de base de données IBM SPSS Modeler.** Informations sur la manière de tirer parti de la puissance de votre base de données pour améliorer les performances et étendre la gamme des fonctions analytiques via des algorithmes tiers.
- **Guide des performances et d'administration IBM SPSS Modeler Server.** Informations sur le mode de configuration et d'administration de IBM® SPSS® Modeler Server.
- **Guide de l'utilisateur de IBM SPSS Modeler Administration Console.** Informations concernant l'installation et l'utilisation de l'interface utilisateur de la console permettant de surveiller et de configurer SPSS Modeler Server. La console est implémentée en tant que plug-in à l'application Deployment Manager.
- **Guide IBM SPSS Modeler Solution Publisher.** SPSS Modeler Solution Publisher est un module complémentaire qui permet aux entreprises de publier des flux destinés à être utilisés en dehors de l'environnement SPSS Modeler.
- **Guide CRISP-DM IBM SPSS Modeler** Guide détaillé sur l'utilisation de la méthodologie CRISP-DM pour le Data mining avec SPSS Modeler
- **Guide de l'utilisateur IBM SPSS Modeler Batch.** Guide complet sur l'utilisation de IBM SPSS Modeler en mode par lots, avec des détails sur l'exécution en mode par lots et les arguments de ligne de commande. Ce guide est disponible au format PDF uniquement.

Documentation de SPSS Modeler Premium

La suite de documentation SPSS Modeler Premium (à l'exception des instructions d'installation) est la suivante.

- **IBM SPSS Modeler Entity Analytics Guide de l'utilisateur.** Informations sur l'utilisation des analyses d'entités avec SPSS Modeler, notamment l'installation et la configuration du référentiel, les nœuds d'analyses d'entités et les tâches administratives.

- **IBM SPSS Modeler Social Network Analysis Guide de l'utilisateur.** Guide sur l'exécution des analyses de réseaux sociaux avec SPSS Modeler, y compris les analyses de groupe et analyses de diffusion.
- **Text Analytics for SPSS Modeler Guide de l'utilisateur.** Informations sur l'utilisation des analyses de texte avec SPSS Modeler, notamment sur les nœuds de Text Mining, l'espace de travail interactif, les modèles et d'autres ressources.
- Guide de l'utilisateur de **Text Analytics for IBM SPSS Modeler Administration Console.** Informations concernant l'installation et l'utilisation de l'interface utilisateur de la console permettant de surveiller et de configurer IBM® SPSS® Modeler Server pour l'utiliser avec Text Analytics for SPSS Modeler. La console est implémentée en tant que plug-in à l'application Deployment Manager.

Exemples d'application

Tandis que les outils de Data mining de SPSS Modeler peuvent vous aider à résoudre une grande variété de problèmes commerciaux et organisationnels, les exemples d'application fournissent des introductions brèves et ciblées aux méthodes et aux techniques de modélisation. Les ensembles de données utilisés ici sont beaucoup plus petits que les énormes entrepôts de données gérés par certains Data miners, mais les concepts et les méthodes impliqués doivent pouvoir être adaptés à des applications réelles.

Vous pouvez accéder aux exemples en cliquant Exemples d'application dans le menu Aide de SPSS Modeler. Les fichiers de données et les flux d'échantillons sont installés dans le dossier *Demos*, sous le répertoire d'installation du produit. Pour plus d'informations, reportez-vous à la section Dossier Demos dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.

Exemples de modélisation de bases de données. Consultez les exemples dans le *IBM SPSS ModelerGuide d'exploration de base de données*.

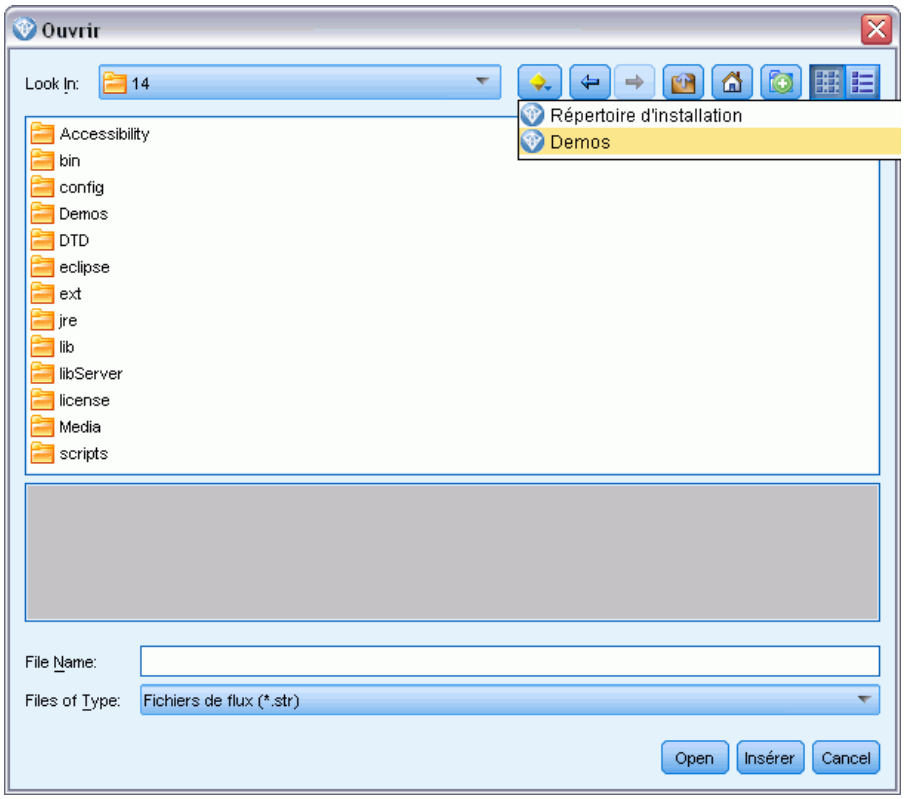
Exemples de génération de scripts. Consultez les exemples dans le *IBM SPSS ModelerGuide de génération de scripts et d'automatisation*.

Dossier Demos

Les fichiers de données et les flux d'échantillons utilisés avec les exemples d'application sont installés dans le dossier *Demos*, sous le répertoire d'installation du produit. Ce dossier est également accessible à partir du groupe de programmes sous IBM SPSS Modeler 15 dans le menu

Démarrer de Windows, ou en cliquant sur *Demos* dans la liste des répertoires récents de la boîte de dialogue Ouverture de fichier.

Figure 1-1
Sélection du dossier Demos dans la liste des répertoires récemment consultés



Partie I:
Introduction et démarrage

Chapitre
2

Présentation de IBM SPSS Modeler

Démarrage

En tant qu'application de Data mining, IBM® SPSS® Modeler constitue une méthode stratégique de recherche de relations utiles dans les grands ensembles de données. Contrairement aux méthodes statistiques plus traditionnelles, il n'est pas indispensable de savoir ce que vous recherchez avant de commencer. Vous pouvez explorer vos données, créer divers modèles et explorer diverses relations, jusqu'à ce que vous trouviez des informations utiles.

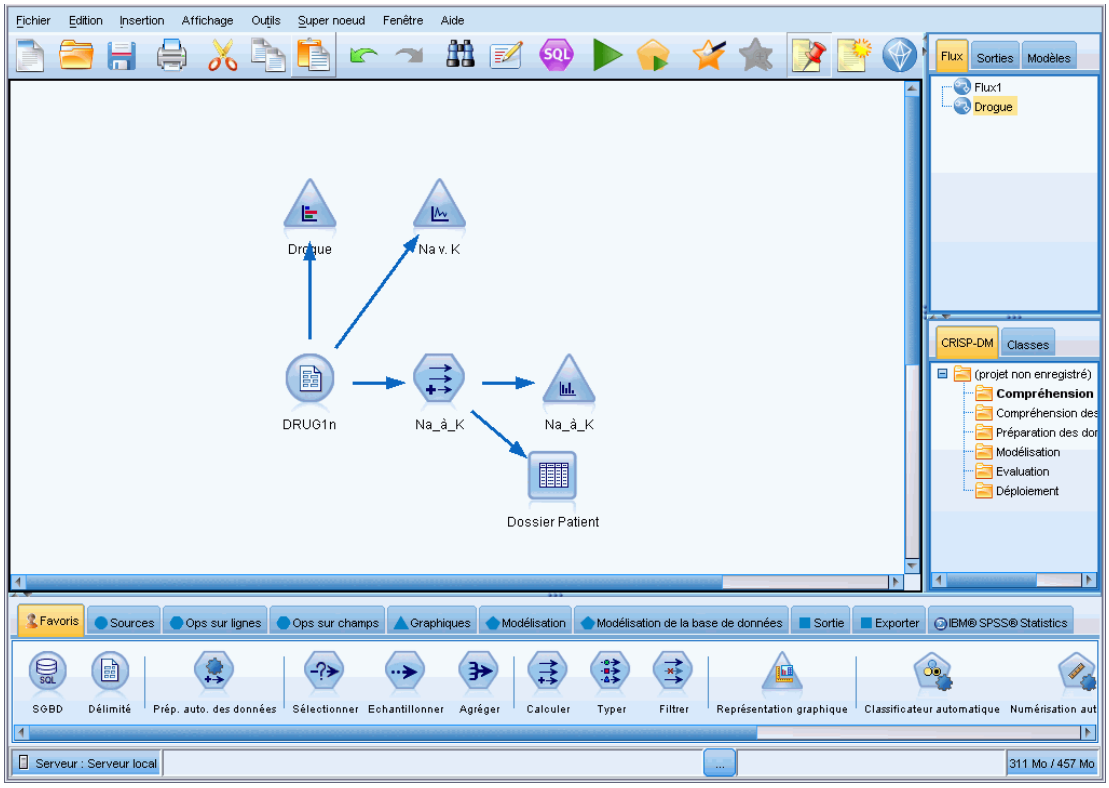
Démarrage de IBM SPSS Modeler

Pour démarrer l'application, cliquez :

Début > Programmes [tout] > IBM SPSS Modeler15 > IBM SPSS Modeler15

La fenêtre principale apparaît après quelques secondes.

Figure 2-1
fenêtre d'application principale de IBM SPSS Modeler



Lancement de l'application à partir de la ligne de commande

Vous pouvez utiliser la ligne de commande de votre système d'exploitation pour lancer IBM® SPSS® Modeler comme suit :

- Dans le cas d'un ordinateur sur lequel est installé IBM® SPSS® Modeler, ouvrez une fenêtre DOS ou une invite de commande.
- Pour lancer l'interface SPSS Modeler en mode interactif, tapez la commande modelerclient suivie des arguments souhaités, par exemple :

```
modelerclient -stream report.str -execute
```

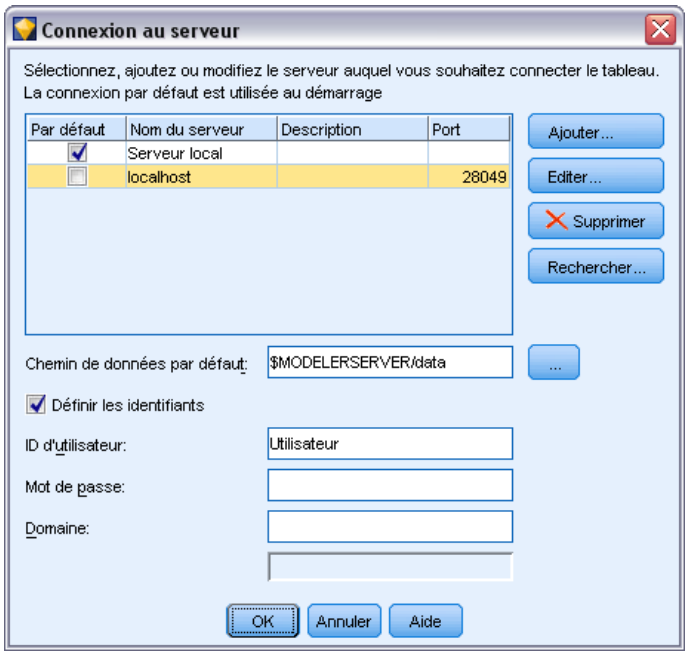
Les arguments disponibles (drapeaux) vous permettent de vous connecter à un serveur, de charger des flux, d'exécuter des scripts, ou d'indiquer les autres paramètres nécessaires.

Connexion au IBM SPSS Modeler Server

Il est possible d'exécuter IBM® SPSS® Modeler comme une application autonome ou un comme un client connecté directement à IBM® SPSS® Modeler Server ou à SPSS Modeler Server or à un groupe de serveurs par le biais du Coordinateur des processus connecté à partir de IBM® SPSS® Collaboration and Deployment Services. L'état de la connexion apparaît en bas à gauche de la fenêtre SPSS Modeler.

Lorsque vous souhaitez vous connecter à un serveur, vous pouvez saisir manuellement son nom ou sélectionner un nom que vous aurez préalablement défini. En revanche, si vous avez IBM SPSS Collaboration and Deployment Services, vous avez la possibilité de chercher dans une liste de serveurs ou de groupes de serveurs à partir de la boîte de dialogue Connexion au serveur. Vous pouvez naviguer via les services Statistics s'exécutant sur un réseau grâce au Coordinateur des processus. Pour plus d'informations, reportez-vous à la section Equilibrage de charge avec classes de serveur dans l'annexe D dans *Guide d'administration et des performances de IBM SPSS Modeler Server 15*.

Figure 2-2
Boîte de dialogue Connexion au serveur



Pour vous connecter à un serveur

▶ Dans le menu Outils, cliquez sur Connexion au serveur. La boîte de dialogue Connexion au serveur s'affiche. Vous pouvez également double-cliquez sur la zone d'état de la connexion dans la fenêtre SPSS Modeler.

▶ Dans la boîte de dialogue, indiquez les options de connexion à l'ordinateur du serveur local ou sélectionnez une connexion dans le tableau.

- Cliquez sur Ajouter ou Modifier pour ajouter ou modifier une connexion. Pour plus d'informations, reportez-vous à la section Ajout et modification d'une connexion à IBM SPSS Modeler Server dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.
- Cliquez sur Rechercher pour accéder au serveur ou à un groupe de serveurs dans le Coordinateur de processus. Pour plus d'informations, reportez-vous à la section Recherche de serveurs dans IBM SPSS Collaboration and Deployment Services dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.

Tableau de serveur. Ce tableau comprend un ensemble de connexions au serveur définies. Il affiche la connexion par défaut, le nom du serveur, sa description et le numéro du port. Vous pouvez ajouter manuellement une nouvelle connexion ainsi que sélectionner ou rechercher une connexion existante. Pour définir un serveur particulier comme connexion par défaut, cochez la case dans la colonne Par défaut du tableau de la connexion.

Chemin de données par défaut. Indiquez le chemin d'accès aux données situées sur l'ordinateur serveur. Cliquez sur le bouton ... pour accéder à l'emplacement requis.

Définir les informations d'identification. Laissez cette case décochée pour permettre à la fonction de **connexion unique** de se connecter au serveur à l'aide de vos informations de nom d'utilisateur et de mot de passe locaux. Si la connexion unique n'est pas disponible, ou si vous cochez la case pour désactiver la connexion unique (par exemple pour vous connecter à un compte administrateur), les champs suivants sont activés et vous permettent d'entrer vos informations d'identification.

Nom d'utilisateur. Entrez le nom d'utilisateur avec lequel effectuer la connexion au serveur.

Mot de passe : Entrez le mot de passe associé au nom d'utilisateur défini.

Domaine. Indiquez le domaine utilisé pour la connexion au serveur. Le nom de domaine n'est requis que si l'ordinateur serveur se trouve dans un autre domaine Windows que l'ordinateur client.

▶ Cliquez sur OK pour terminer la connexion.

Pour se déconnecter d'un serveur

▶ Dans le menu Outils, cliquez sur Connexion au serveur. La boîte de dialogue Connexion au serveur s'affiche. Vous pouvez également double-cliquez sur la zone d'état de la connexion dans la fenêtre SPSS Modeler.

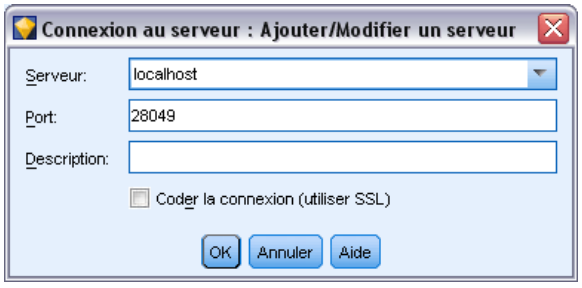
▶ Dans la boîte de dialogue, sélectionnez le serveur local puis cliquez sur OK.

Ajout et modification d'une connexion à IBM SPSS Modeler Server

Dans la boîte de dialogue Connexion au serveur, vous pouvez modifier ou ajouter une connexion au serveur. Cliquez sur Ajouter pour accéder à une boîte de dialogue Ajouter/Modifier un serveur non renseignée, dans laquelle vous pourrez entrer les données de la connexion au serveur. Si vous sélectionnez une connexion existante et cliquez sur Modifier, dans la boîte de dialogue Connexion au serveur, la boîte de dialogue Ajouter/Modifier un serveur s'ouvre, affichant les données de cette connexion, vous permettant ainsi d'apporter toutes les modifications que vous souhaitez.

Remarque : Vous ne pouvez pas modifier une connexion au serveur qui a été ajoutée à partir de IBM® SPSS® Collaboration and Deployment Services, car le nom, le port et d'autres détails sont définis dans IBM SPSS Collaboration and Deployment Services.

Figure 2-3
Boîte de dialogue Ajouter/Modifier un serveur pour la connexion au serveur



Pour ajouter des connexions au serveur

▶ Dans le menu Outils, cliquez sur Connexion au serveur. La boîte de dialogue Connexion au serveur s'affiche.

- Dans la boîte de dialogue, cliquez sur Ajouter. La boîte de dialogue Ajouter/Modifier un serveur pour la connexion au serveur s'ouvre.
- Saisissez les données de connexion au serveur puis cliquez sur OK pour enregistrer la connexion et retourner à la boîte de dialogue Connexion au serveur.
  - **Serveur.** Indiquez un serveur disponible ou sélectionnez-en un dans la liste. L'ordinateur serveur peut être identifié par un nom alphanumérique (par exemple, *monserveur*) ou une adresse IP qui lui est affectée (par exemple, 202.123.456.78).
  - **Port.** Indiquez le numéro de port d'écoute du serveur. Si ce port par défaut ne fonctionne pas, demandez à l'administrateur système le numéro de port correct.
  - **Description.** Saisissez une description optionnelle pour la connexion à ce serveur.
  - **Coder la connexion (utiliser SSL).** Indique si une connexion SSL (**Secure Sockets Layer**) doit être utilisée. Le protocole SSL est fréquemment utilisé pour la sécurisation des données sur un réseau. Pour pouvoir utiliser cette fonction, vous devez activer le protocole SSL sur le serveur hébergeant le IBM® SPSS® Modeler Server. Si nécessaire, contactez votre administrateur local pour plus d'informations.

Pour modifier des connexions au serveur

- Dans le menu Outils, cliquez sur Connexion au serveur. La boîte de dialogue Connexion au serveur s'affiche.
- Dans la boîte de dialogue, sélectionnez la connexion que vous souhaitez modifier puis cliquez sur Modifier. La boîte de dialogue Ajouter/Modifier un serveur pour la connexion au serveur s'ouvre.
- Modifiez les données de connexion au serveur puis cliquez sur OK pour enregistrer les changements et retourner à la boîte de dialogue Connexion au serveur.

Recherche de serveurs dans IBM SPSS Collaboration and Deployment Services

Au lieu d'entrer manuellement une connexion au serveur, vous pouvez sélectionner un serveur ou un groupe de serveurs disponible sur le réseau par le biais du Coordinateur de processus, disponible dans IBM® SPSS® Collaboration and Deployment Services. Un groupe de serveurs contient plusieurs serveurs, et permet au Coordinateur de processus de déterminer le serveur qui répond le mieux à la demande de traitement. Pour plus d'informations, reportez-vous à la section Equilibrage de charge avec classes de serveur dans l'annexe D dans *Guide d'administration et des performances de IBM SPSS Modeler Server 15*.

Bien que vous ne vous puissiez pas ajouter manuellement de serveurs dans la boîte de dialogue Connexion au serveur, la recherche de serveurs disponibles vous permet de vous connecter aux serveurs sans que vous ayez besoin de connaître le nom du serveur et le numéro du port. Ces informations sont fournies automatiquement. Il vous faut néanmoins corriger les données de connexion telles que le nom de l'utilisateur, le domaine et le mot de passe.

Remarque : Si vous n'avez pas accès au Coordinateur de processus, vous pouvez tout de même saisir manuellement le nom du serveur auquel vous souhaitez vous connecter ou sélectionner un nom que vous aurez défini au préalable. Pour plus d'informations, reportez-vous à la section Ajout et modification d'une connexion à IBM SPSS Modeler Server dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.

Figure 2-4
Boîte de dialogue Recherche de serveurs

Pour rechercher des serveurs et des groupes de serveurs

- Dans le menu Outils, cliquez sur Connexion au serveur. La boîte de dialogue Connexion au serveur s'affiche.
- Cliquez sur Rechercher pour ouvrir la boîte de dialogue Recherche de serveurs. Si vous n'êtes plus connecté à IBM SPSS Collaboration and Deployment Services, lors de votre tentative d'accès au Coordinateur de processus, il vous sera demandé de vous reconnecter. Pour plus d'informations, reportez-vous à la section Connexion au référentiel dans le chapitre 9 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.
- Sélectionnez le serveur ou le groupe de serveurs dans la liste.
- Cliquez sur OK pour fermer la boîte de dialogue et ajouter cette connexion au tableau de la boîte de dialogue Connexion au serveur.

Changement de répertoire temporaire

Certaines opérations effectuées par IBM® SPSS® Modeler Server peuvent nécessiter la création de fichiers temporaires. Par défaut, IBM® SPSS® Modeler crée les fichiers temporaires dans le répertoire temporaire du système. Vous pouvez modifier l'emplacement du répertoire temporaire en effectuant les opérations suivantes.

- Créez un répertoire intitulé *spss* et un sous-répertoire intitulé *servertemp*.
- Editez le fichier *options.cfg*, situé dans le répertoire */config* du dossier d'installation de SPSS Modeler. Editez le paramètre temp_directory de ce fichier en saisissant : temp_directory, "C:/spss/servertemp".
- Redémarrez ensuite le service SPSS Modeler Server. Pour ce faire, cliquez sur Services dans les outils d'administration du Panneau de configuration Windows. Il vous suffit d'arrêter le service et de le redémarrer pour appliquer les modifications apportées. Vous pouvez redémarrer l'ordinateur pour redémarrer le service.

  Tous les fichiers temporaires sont désormais écrits dans ce nouveau répertoire.

Remarque : L'erreur la plus courante est de ne pas utiliser le type correct de barre oblique. En raison de l'historique UNIX de SPSS Modeler, les barres obliques sont utilisées.

Démarrage de plusieurs sessions IBM SPSS Modeler

Si vous devez lancer plus d'une session IBM® SPSS® Modeler à la fois, vous devez effectuer certaines modifications de votre IBM® SPSS® Modeler et des paramètres Windows. Par exemple, il vous faudra effectuer ces modifications si vous avez deux licences de serveur distinctes et que vous souhaitez exécuter deux flux pour deux serveurs distincts du même ordinateur client.

Pour activer plusieurs sessions SPSS Modeler :

- Cliquez :
  Début > Programmes [tout] > IBM SPSS Modeler15
- Dans le raccourci de IBM SPSS Modeler15 (celui avec l'icône), cliquez avec le bouton droit de la souris et sélectionnez Propriétés.
- Dans la zone de texte Cible, ajoutez -noshare à la fin de la chaîne.
- Dans Windows Explorer, sélectionnez :
  Outils > Options des dossiers...
- Dans l'onglet Types de fichier, sélectionnez l'option Flux de SPSS Modeler et cliquez sur Avancé.
- Dans la boîte de dialogue Modifier le type de fichier, sélectionnez Ouvrir avec SPSS Modeler et cliquez sur Modifier.
- Dans la zone de texte Application utilisée pour effectuer l'action, ajoutez -noshare avant l'argument -flux.

Interface IBM SPSS Modeler en un clin d'oeil

A chaque étape du processus de Data mining, l'interface simplifiée de IBM® SPSS® Modeler met à contribution vos compétences professionnelles. Les algorithmes de modélisation, comme la prévision, la classification, la segmentation et la détection d'association, permettent l'obtention de modèles performants et précis. Les résultats du modèle peuvent facilement être déployés et lus dans des bases de données, dans IBM® SPSS® Statistics et dans de nombreuses autres applications.

Travailler avec SPSS Modeler est un processus en trois étapes de travail avec les données.

- Pour commencer, lisez les données de SPSS Modeler.
- Ensuite, exécutez les données par une série de manipulations.
- Et pour finir, envoyez les données vers une destination choisie.

Cette séquence d'opérations est appelée **flux de données** car les données circulent, enregistrement par enregistrement, de la source à la destination (modèle ou type de sortie de données), en passant par chaque manipulation.

Figure 2-5
Un flux simple

Espace de travail de flux IBM SPSS Modeler

L'espace de travail de flux est la plus grande zone de la fenêtre IBM® SPSS® Modeler. C'est dans cette zone que vous créez et manipulez les flux de données.

Les flux sont créés en dessinant des diagrammes des opérations de données nécessaires à votre entreprise sur l'espace de travail principal de l'interface. Chaque opération est représentée par une icône ou un **noeud**, et les noeuds sont reliés entre eux dans un **flux** représentant le flux de données passant par chaque opération.

Vous pouvez utiliser plusieurs flux à la fois dans SPSS Modeler, que ce soit dans le même espace de travail de flux ou par l'ouverture d'un nouveau flux. Au cours d'une session, les flux sont stockés dans le gestionnaire de flux, en haut à droite de la fenêtre SPSS Modeler.

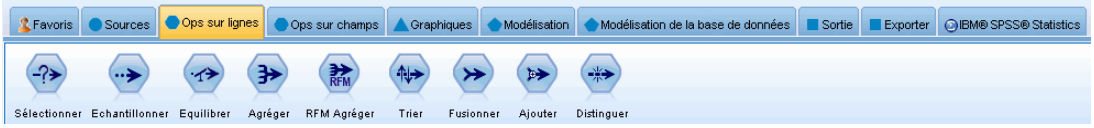
Palette de noeuds

La plupart des données et des outils de modélisation de IBM® SPSS® Modeler se trouvent dans la **Palette de noeuds**, au bas de la fenêtre sous l'espace de travail de flux.

Par exemple, l'onglet de la palette Ops sur lignes contient des noeuds permettant d'effectuer des opérations sur les **lignes**, telles que la sélection, la fusion et l'ajout.

Pour ajouter des noeuds à l'espace de travail, double-cliquez sur les icônes de la palette de noeuds ou faites glisser les icônes vers l'espace de travail. Vous pouvez ensuite les relier afin de créer un **flux** représentant le flux des données.

Figure 2-6
Onglet Ops sur lignes de la palette de noeuds



Chaque onglet de palette contient un ensemble de noeuds connexes employés pour différentes étapes des opérations de flux, comme :

- **Sources.** Les noeuds amènent les données dans SPSS Modeler.

- **Ops sur lignes.** Noeuds utilisés pour les opérations sur les **lignes** de données, comme la sélection, la fusion et l'ajout.
- **Ops sur champs.** Noeuds utilisés pour les opérations sur les **champs** de données, comme le filtrage, le calcul de nouveaux champs et la détermination du niveau de mesure de champs particuliers.
- **Graphiques.** Noeuds utilisés pour visualiser les données avant et après la modélisation. Les graphiques peuvent être des nuages, des histogrammes, des noeuds Relations, ainsi que des graphiques d'évaluation.
- **Modélisation.** Noeuds utilisant les algorithmes de modélisation disponibles dans SPSS Modeler, comme les réseaux de neurones, les arbres décision, les algorithmes de classification et la création de séquences de données.
- **Modélisation de bases de données.** Les noeuds utilisent les algorithmes de modélisation disponibles dans les bases de données Microsoft SQL Server, IBM DB2 et Oracle.
- **Résultats.** Les noeuds produisent diverses sorties pour les données, les diagrammes et les résultats de modèles qui peuvent être affichés dans SPSS Modeler.
- **Exporter.** Les noeuds produisent diverses sorties qui peuvent être affichées dans des applications externes telles que IBM® SPSS® Data Collection ou Excel.
- **SPSS Statistics.** Les noeuds importent des données à partir de, ou exportent des données vers IBM® SPSS® Statistics, et exécutent aussi des procédures SPSS Statistics.

Au fur et à mesure que vous maîtrisez mieux l'application SPSS Modeler, vous pouvez personnaliser le contenu de la palette en fonction de vos besoins. Pour plus d'informations, reportez-vous à la section Personnalisation de la palette Noeuds dans le chapitre 12 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.

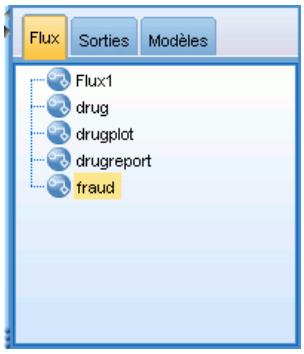
Situé sous la palette de noeuds, un panneau de rapports fournit des informations sur la progression des diverses opérations, telles que la lecture des données dans le flux de données. Egalement situé sous la palette de noeuds, un panneau d'état fournit des informations sur l'activité actuelle de l'application, ainsi que des indications lorsqu'une saisie par l'utilisateur est requise.

Gestionnaires IBM SPSS Modeler

En haut à droite de la fenêtre se trouve le panneau des gestionnaires. Il contient trois onglets qui permettent de gérer les flux, les sorties et les modèles.

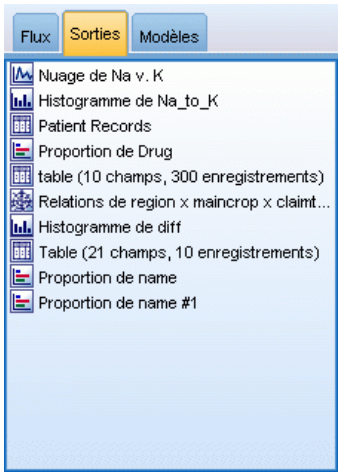
Vous pouvez utiliser l'onglet Flux pour ouvrir, renommer, enregistrer et supprimer les flux créés dans une session.

Figure 2-7
Onglet Flux



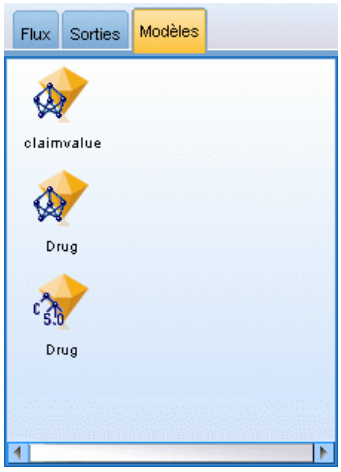
L'onglet Sortie contient différents fichiers, tels que des graphiques et des tableaux, produits par des opérations de flux dans IBM® SPSS® Modeler. Vous pouvez afficher, sauvegarder, renommer et fermer les tableaux, les graphiques et les rapports qui figurent dans cet onglet.

Figure 2-8
Onglet Sorties



L'onglet Modèle est le plus puissant des onglets du gestionnaire. Cet onglet contient tous les **nuggets** de modèle qui contiennent les modèles générés dans SPSS Modeler, pour la session en cours. Vous pouvez accéder à ces modèles directement à partir de l'onglet Modèles ou les ajouter au flux dans l'espace de travail.

Figure 2-9
Onglet Modèles qui contient des nugget de modèles

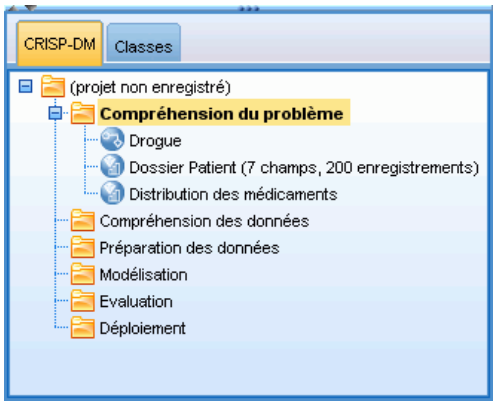


Projets IBM SPSS Modeler

Dans la partie inférieure droite de la fenêtre se trouve le panneau de projet qui permet de créer et de gérer les **projets** de Data mining (groupes de fichiers en rapport avec une tâche de data mining). Vous pouvez afficher les projets créés de deux façons dans IBM® SPSS® Modeler— en mode Classes et en mode CRISP-DM.

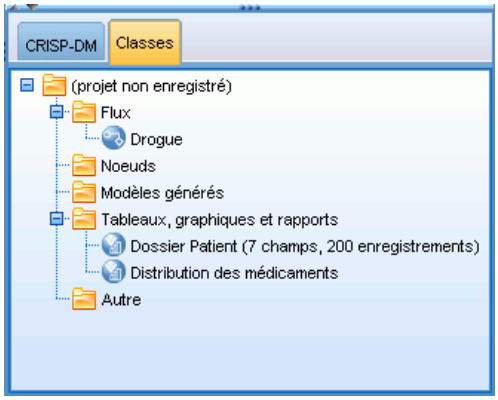
L'onglet CRISP-DM permet d'organiser les projets en fonction de la méthodologie Cross-Industry Standard Process for Data mining commune utilisée dans le domaine. Que vous soyez un utilisateur chevronné ou novice, l'outil CRISP vous aidera à mieux organiser et communiquer vos efforts.

Figure 2-10
Mode CRISP-DM



L'onglet Classes permet d'organiser votre travail dans SPSS Modeler en catégories, selon les types d'objet que vous créez. Ce mode d'affichage est utile lorsque vous effectuez l'inventaire des données, des flux et des modèles.

Figure 2-11
mode Classes



Barre d'outils IBM SPSS Modeler

Une barre d'outils, composée d'icônes fournissant des options très utiles, se trouve en haut de la fenêtre IBM® SPSS® Modeler. Voici les boutons de la barre d'outils et leurs fonctions.


Créer un flux


Permet d'ouvrir un flux


Enregistrer le flux


Imprimer le flux actuel


Déplacer la sélection vers le Presse-papiers


Copier dans le Presse-papiers


Coller la sélection


Annuler l'action précédente


Rétablir


Recherche de noeuds


Editer les propriétés du flux

Aperçu de génération SQL


Exécuter le flux actuel


Exécuter la sélection de flux

	Arrêter le flux (actif uniquement pendant l'exécution du flux)		Ajouter un super noeud
	Zoom avant (super noeuds uniquement)		Zoom arrière (super noeuds uniquement)
	Aucun balisage dans le flux		Insérer un commentaire
	Masquer le balisage de flux (le cas échéant)		Afficher le balisage de flux masqué
	Permet d'ouvrir un flux dans IBM® SPSS® Modeler Advantage		

Le balisage de flux se compose des commentaires de flux, des liens de modèle et des indications de branche de scoring.

Pour plus d'informations sur les commentaires de flux, reportez-vous à la rubrique Ajout de commentaires et d'annotations à des noeuds et à des flux sur p. .

Pour de plus amples informations sur les indications de branche de scoring, reportez-vous à La branche de scoring sur p. .

Les liens de modèle sont décrits dans le guide *Noeuds de modélisation IBM SPSS*.

Personnalisation de la barre d'outils

Vous pouvez modifier plusieurs aspects de la barre d'outils, tels que :

- choisir si elle sera affichée ou non
- Choisir si les icônes comporteront ou non des info-bulles
- Choisir si elle utilisera des petites ou des grandes icônes

activer ou désactiver l'affichage de la barre d'outils :

▶ Dans le menu principal, cliquez sur :
Affichage > Barre d'outils > Afficher

Pour modifier les paramètres des info-bulles ou de la taille des icônes :

▶ Dans le menu principal, cliquez sur :
Affichage > Barre d'outils > Personnaliser

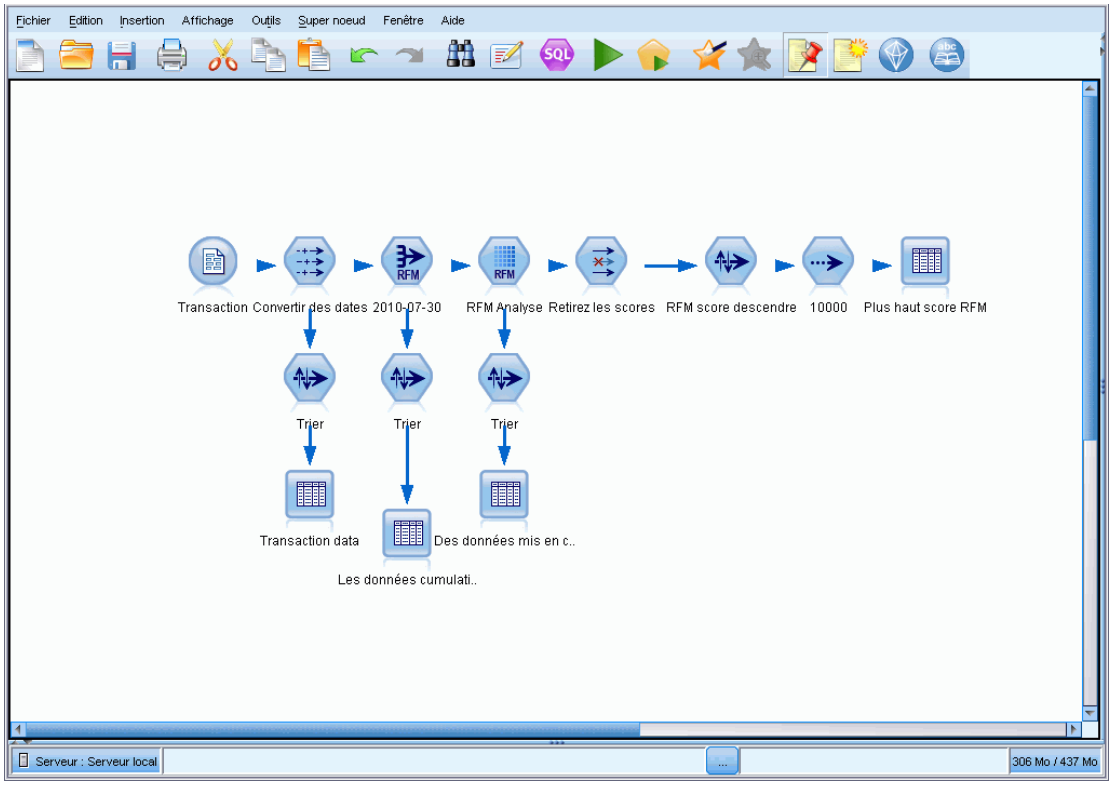
Cliquez sur Afficher les info-bulles ou Gros boutons le cas échéant.

Personnalisation de la fenêtre IBM SPSS Modeler

Vous pouvez utiliser les séparateurs situés entre les différentes zones de l'interface IBM® SPSS® Modeler pour redimensionner ou fermer des outils en fonction de vos besoins. Par exemple, si vous travaillez avec un flux volumineux, vous pouvez utiliser les petites flèches situées sur chaque séparateur pour fermer la palette de noeuds, le panneau des gestionnaires et le panneau des projets. Ainsi, vous agrandissez l'espace de travail de flux et libérez suffisamment d'espace pour les flux volumineux ou multiples.

À partir du menu Affichage, vous pouvez aussi cliquer sur Palette de noeuds, Gestionnaires ou Projet pour activer ou désactiver l'affichage de ces éléments.

Figure 2-12
Espace de travail de flux agrandi



Vous pouvez également garder ouvertes la palette des noeuds, et les panneaux des gestionnaires et des projets, et utiliser les barres de défilement de l'espace de travail de flux pour vous déplacer dans cet espace ; ces barres sont situées sur le côté et en bas de la fenêtre SPSS Modeler.

Vous pouvez aussi commander l'affichage du balisage de l'écran, lequel se compose des commentaires de flux, des liens de modèles et des indications de branche de scoring. Pour activer ou désactiver cet affichage, cliquez sur :

Affichage > Balisage de flux

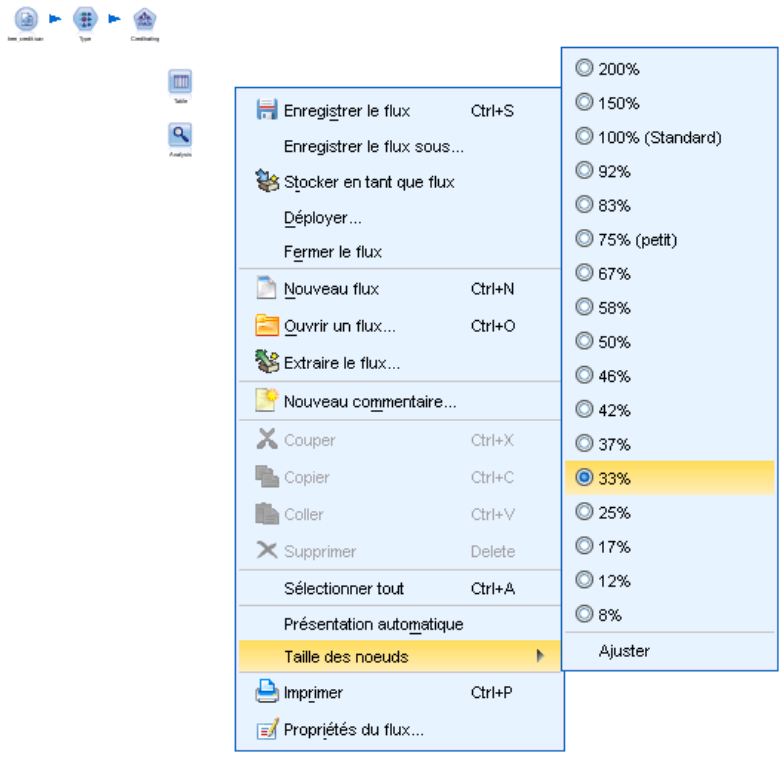
Modification de la taille des icônes d'un flux

Vous pouvez changer la taille des icônes de flux par l'une des méthodes suivantes.

- A l'aide d'un paramètre de propriété de flux
- A l'aide d'un menu contextuel dans le flux
- A l'aide du clavier

Vous pouvez redimensionner la vue entière du flux à une taille comprise entre 8 % et 200 % de la taille d'icône standard.

Figure 2-13
Modification de la taille d'icône



Pour redimensionner le flux entier (méthode des propriétés du flux)

▶ Dans le menu principal, sélectionnez :
Outils > Propriétés du flux > Options > Présentation.

▶ Sélectionnez la taille souhaitée dans le menu Taille d'icône.

▶ Cliquez sur Appliquer pour afficher les résultats.

▶ Cliquez sur OK pour enregistrer cette modification.

Pour redimensionner le flux entier (méthode du menu)

- Cliquez avec le bouton droit de la souris sur l'arrière-plan du flux dans l'espace de travail.
- Sélectionnez l'option Taille d'icône puis la taille souhaitée.

Pour redimensionner le flux entier (méthode du clavier)

- Appuyez sur Ctrl + [-] sur le clavier pour effectuer un zoom arrière et réduire la vue d'une taille.
- Appuyez sur Ctrl + Shift + [+] sur le clavier pour effectuer un zoom avant et agrandir la vue d'une taille.

Cette fonction est particulièrement utile pour obtenir une vue globale d'un flux complexe. Vous pouvez aussi l'utiliser pour réduire le nombre de pages nécessaires à l'impression d'un flux.

Utilisation de la souris dans IBM SPSS Modeler

Dans IBM® SPSS® Modeler, les utilisations les plus courantes de la souris sont les suivantes :

- **Clic simple.** Utilisez le bouton droit ou le bouton gauche de la souris pour sélectionner des options dans les menus, ouvrir des menus contextuels, ou accéder à diverses autres commandes et options standard. Cliquez avec la souris et maintenez le bouton de la souris enfoncé pour faire glisser des noeuds.
- **Double-clic.** Double-cliquez avec le bouton gauche de la souris pour placer des noeuds dans l'espace de travail de flux et éditer des noeuds existants.
- **Clic à l'aide du bouton central.** Cliquez avec le bouton central de la souris et faites glisser le curseur pour connecter des noeuds dans l'espace de travail de flux. Double-cliquez avec le bouton central de la souris pour déconnecter un noeud. Si vous ne possédez pas de souris à trois boutons, vous pouvez simuler cette fonction en appuyant sur la touche Alt tout en cliquant avec la souris et en la faisant glisser.

Utilisation de touches de raccourci

Dans IBM® SPSS® Modeler, de nombreuses opérations de programmation visuelle sont associées à des touches de raccourci. Par exemple, vous pouvez supprimer un noeud en cliquant dessus et en appuyant sur la touche Suppr de votre clavier. De la même façon, vous pouvez enregistrer rapidement un flux en appuyant sur la touche S tout en maintenant la touche Ctrl enfoncée. Les commandes de ce type sont indiquées par Ctrl et une autre touche (par exemple, Ctrl+S).

De nombreuses touches de raccourci sont utilisées dans les opérations Windows standard, telles que Ctrl+X pour couper un élément. Ces raccourcis sont pris en charge dans SPSS Modeler, parallèlement à ceux présentés ci-après, propres à l'application.

Remarque : dans certains cas, les anciennes touches de raccourci utilisées dans SPSS Modeler sont en conflit avec les touches de raccourci Windows standard. Pour que ces anciennes touches de raccourci fonctionnent, il faut utiliser la touche Alt. Par exemple, Ctrl+Alt+C peut activer ou désactiver la mise en cache.

Table 2-1
Touches de raccourci prises en charge

Touche de raccourci	Fonction
Ctrl+A	Tout sélectionner
Ctrl+X	Couper
Ctrl+N	Permet de créer un flux
Ctrl+O	Permet d'ouvrir un flux
Ctrl+P	Imprimer
Ctrl+C	Copier
Ctrl+V	Coller
Ctrl+Z	Annuler
Ctrl+Q	Permet de sélectionner tous les noeuds situés en aval du noeud sélectionné.
Ctrl+W	Permet de désélectionner tous les noeuds en aval (bascule du raccourci Ctrl+Q)
Ctrl+E	Exécuter à partir d'un noeud sélectionné
Ctrl+S	Permet d'enregistrer le flux en cours
Alt+flèches	Permettent de déplacer les noeuds sélectionnés dans l'espace de travail de flux dans le sens indiqué par la flèche utilisée
Maj+F10	Permet d'ouvrir le menu contextuel du noeud sélectionné

Table 2-2
Touches de raccourci prises en charge pour les anciennes touches d'accès rapide

Touche de raccourci	Fonction
Ctrl+Alt+D	Permet de dupliquer un noeud
Ctrl+Alt+L	Permet de charger un noeud
Ctrl+Alt+R	Permet de renommer un noeud
Ctrl+Alt+U	Permet de créer un noeud Utilisateur
Ctrl+Alt+C	Permet d'activer et de désactiver le cache
Ctrl+Alt+F	Permet de vider le cache
Ctrl+Alt+X	Développer le super noeud
Ctrl+Alt+Z	Permet d'effectuer un zoom avant/arrière
Suppr	Permet de supprimer un noeud ou une connexion

Impression

Les objets suivants peuvent être imprimés dans IBM® SPSS® Modeler :

- Diagrammes de flux
- Graphiques
- Tableaux
- Rapports (à partir du noeud Rapport et des rapports de projet)
- Scripts (à partir des boîtes de dialogue Propriétés du flux, Script autonome ou Script Super noeud)

- Modèles (navigateurs de modèle, onglets de boîte de dialogue avec élément en cours, afficheurs d'arbres)
- Annotations (à partir de l'onglet Annotations de la sortie)

Pour imprimer un objet :

- Pour imprimer sans afficher d'aperçu, cliquez sur le bouton Imprimer de la barre d'outils.
- Pour définir la mise en page avant d'imprimer, sélectionnez Mise en page dans le menu Fichier.
- Pour afficher un aperçu avant d'imprimer, sélectionnez Aperçu avant impression dans le menu Fichier.
- Pour afficher la boîte de dialogue d'impression standard vous permettant de sélectionner les imprimantes et de définir des options d'aspect, sélectionnez Imprimer dans le menu Fichier.

Automatisation de IBM SPSS Modeler

Etant donné que le Data mining avancé peut être complexe et parfois long, IBM® SPSS® Modeler comprend plusieurs types d'assistance au codage et à l'automatisation.

- **Control Language for Expression Manipulation** (CLEM) est un langage permettant d'analyser et de manipuler les données circulant au sein des flux de SPSS Modeler. Les data miners utilisent beaucoup le langage CLEM dans les opérations de flux pour exécuter des tâches aussi simples que le calcul du profit à partir des données de coûts et de revenus, ou aussi complexes que la transformation de données du log Web en un ensemble de champs et d'enregistrements contenant des informations utilisables. Pour plus d'informations, reportez-vous à la section A propos de CLEM dans le chapitre 7 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.
- **La génération de scripts** est un outil performant pour automatiser les processus dans l'interface utilisateur. Les scripts effectuent des opérations semblables à celles qui peuvent être exécutées à la souris ou au clavier. Vous pouvez définir des options pour les noeuds et effectuer des dérivations en utilisant un sous-ensemble du CLEM. Vous pouvez également définir une sortie et manipuler des modèles générés. Pour plus d'informations, reportez-vous à la section Génération de scripts - Présentation dans le chapitre 2 dans *Guide de génération de scripts et d'automatisation de IBM SPSS Modeler 15*.

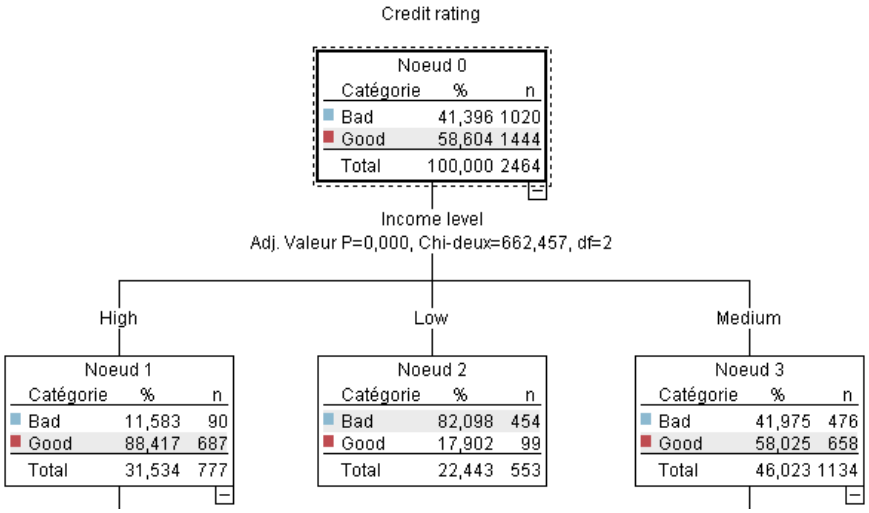
Chapitre
3

Introduction à la modélisation

Un modèle est un ensemble de règles, de formules, ou d'équations pouvant être utilisées pour prédire un résultat en fonction d'un ensemble de champs ou de variables d'entrée. Par exemple, une institution financière peut utiliser un modèle pour prédire si les emprunteurs représentent un risque important ou peu de risque, en fonction des informations déjà connues sur le passé de ces emprunteurs.

La capacité à prédire un résultat est l'objectif central de l'analyse prédictive, et la compréhension du processus de modélisation est essentielle pour l'utilisation de IBM® SPSS® Modeler.

Figure 3-1
Modèle d'arbre décision simple



Cet exemple utilise un modèle d'**arbre décision** qui classifie les enregistrements (et prédit une réponse) à l'aide d'une série de règles de décisions, par exemple :

SI revenu = Moyen
ET cartes <5
ALORS -> 'Bon'

Bien que cet exemple utilise un modèle CHAID (Chi-Squared Automatic Interaction Detection), il est destiné à fournir une introduction générale, et la plupart des concepts s'appliquent globalement aux autres types de modélisation dans SPSS Modeler.

Pour comprendre tous les modèles, vous devez d'abord comprendre les données qu'ils contiennent. Les données de cet exemple contiennent des informations sur les clients d'une banque. Les champs suivants sont utilisés :

Nom de champ	Description
Conditions_crédit	Conditions de crédit : 0=Mauvaises, 1=Bonnes, 9=valeurs manquantes
Age	Age en années
Revenu	Niveau de revenu : 1=bas, 2=moyen, 3=élevé
Cartes_crédit	Nombre de cartes de crédit possédées : 1=Moins de cinq, 2=Cinq ou plus
Éducation	Niveau d'éducation : 1=Université, 2=Lycée
Prêts_voiture	Nombre de prêts voiture en cours : 1=aucun ou un, 2=Plus de deux

La banque gère une base de données contenant des informations sur les clients qui ont contracté un prêt, notamment sur le respect de leur engagement de remboursement (conditions de crédit = bonnes) ou le non-respect de leur engagement (conditions de crédit = mauvaises). À l'aide de ces données, la banque peut créer un modèle qui lui permettra de prédire les probabilités de remboursement des futurs emprunteurs.

À partir d'un modèle d'arbre de décision, vous pouvez analyser les caractéristiques de deux groupes de clients et prédire les risques de non-remboursement.

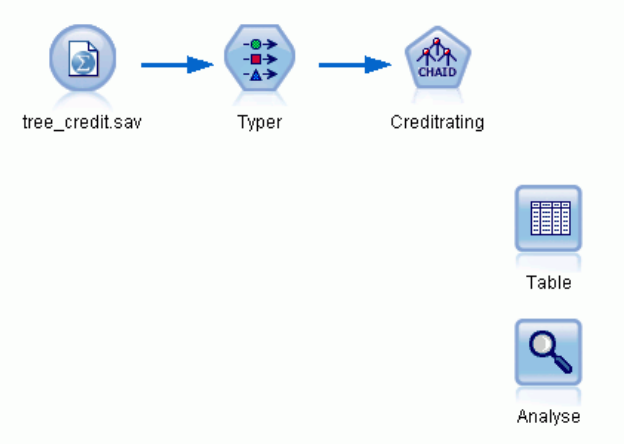
Cet exemple utilise le flux nommé *modelingintro.str*, disponible dans le dossier *Demos* du sous-dossier des *flux*. Le fichier de données est *tree_credit.sav*. Pour plus d'informations, reportez-vous à la section Dossier Demos dans le chapitre 1 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.

Regardons le flux de plus près.

- Dans le menu principal, sélectionnez les options suivantes :
  Fichier > Ouvrir un flux
- Cliquez sur l'icône de la pépite d'or dans la barre d'outils de la boîte de dialogue Ouvrir et choisissez le dossier Demos.
- Double-cliquez sur le dossier des *flux*.
- Double-cliquez sur le fichier *modelingintro.str*.

Création du flux

Figure 3-2
Modélisation du flux



Pour construire un flux qui va créer un modèle, vous avez besoin d'au moins trois éléments :

- Un noeud source qui lit les données issues d'une source externe, dans ce cas un fichier de données IBM® SPSS® Statistics.
- Un noeud source ou Typer qui spécifie les propriétés des champs, telles que le niveau de mesure (le type de données contenues dans le champ) et le rôle de chaque champ en tant que cible ou entrée dans la modélisation.
- Un noeud de modélisation qui génère un nugget de modèle lors de l'exécution du flux.

Dans cet exemple, nous utilisons un noeud de modélisation CHAID. CHAID, ou Chi-Squared Automatic Interaction Detection, est une méthode de classification qui crée des arbres de décision à l'aide d'un type de statistiques spécifique connu sous le nom de statistiques du Khi-deux et qui permet de définir les meilleurs endroit auxquels opérer le découpage dans l'arbre de décision.

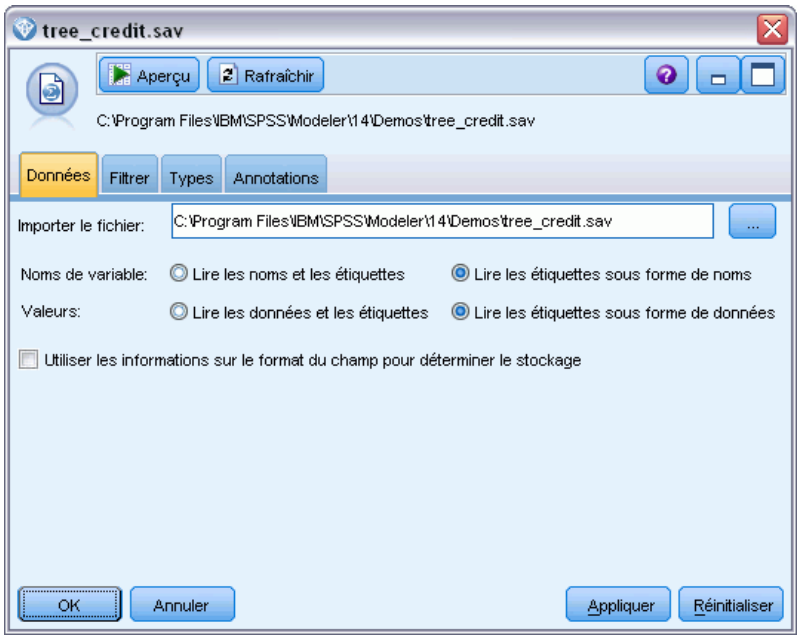
Si les niveaux de mesure sont spécifiés dans le noeud source, le noeud Typer distinct peut être éliminé. D'un point de vue fonctionnel, le résultat est le même.

Ce flux comporte également des noeuds Table et Analyse qui seront utilisés pour afficher les résultats de scoring après la création du nugget de modèle et son ajout au flux.

Le noeud Statistics lit les données au format SPSS Statistics à partir du fichier de données *tree_credit.sav*, qui est installé dans le dossier *Demos*. (Une variable spéciale nommée *$CLEO_DEMOS* est utilisée pour faire référence à ce dossier sous l'installation IBM® SPSS®

Modeler actuelle. Ainsi, le chemin sera toujours valide, quelque soit le dossier d'installation actuel ou la version.)

Figure 3-3
Lecture des données avec un noeud source Statistics

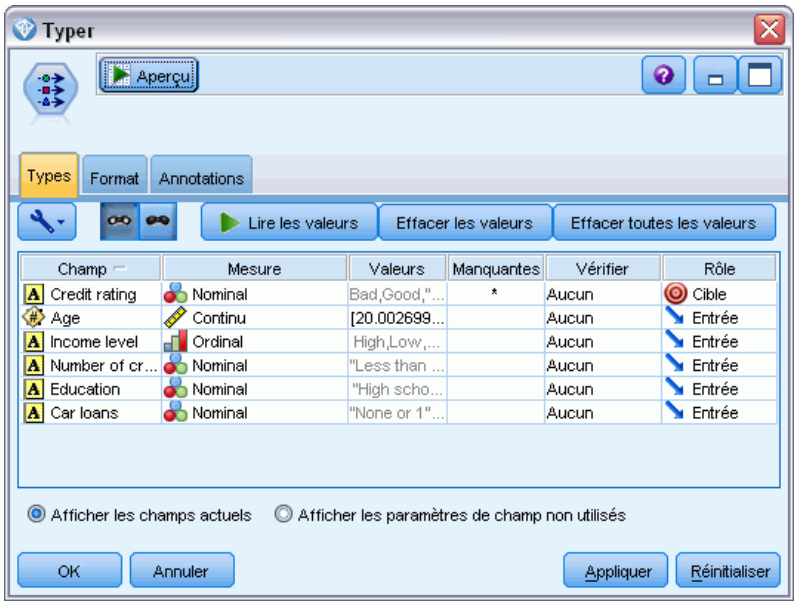


Le noeud Typer définit le **niveau de mesure** pour chaque champ. Le niveau de mesure est une catégorie qui indique le type de données du champ. Notre fichier de données source utilise trois niveaux de mesure différents.

Un champ **Continu** (comme le champ *Age*) contient des valeurs numériques continues, alors qu'un champ **Nominal** (comme le champ *Conditions de crédit*) contient deux valeurs distinctes minimum, par exemple *Mauvaises*, *Bonnes*, ou *Pas d'antécédents de crédit*. Un champ **Ordinal**

(comme le champ *Niveau de revenu*) décrit les données avec différentes valeurs distinctes ayant un ordre inhérent—dans ce cas *Bas*, *Moyen* et *Elevé*.

Figure 3-4
Définition des champs cibles et des champs d'entrées avec le noeud Typer



Pour chaque champ, le noeud Typer spécifie également un **rôle**, qui indique le rôle que joue chaque champ dans la modélisation. Le rôle est défini sur *Cible* pour le champ *Conditions de crédit*, qui indique si un client donné a remboursé ou non son prêt. Il s'agit de la **cible**, soit le champ dont vous souhaitez prédire la valeur.

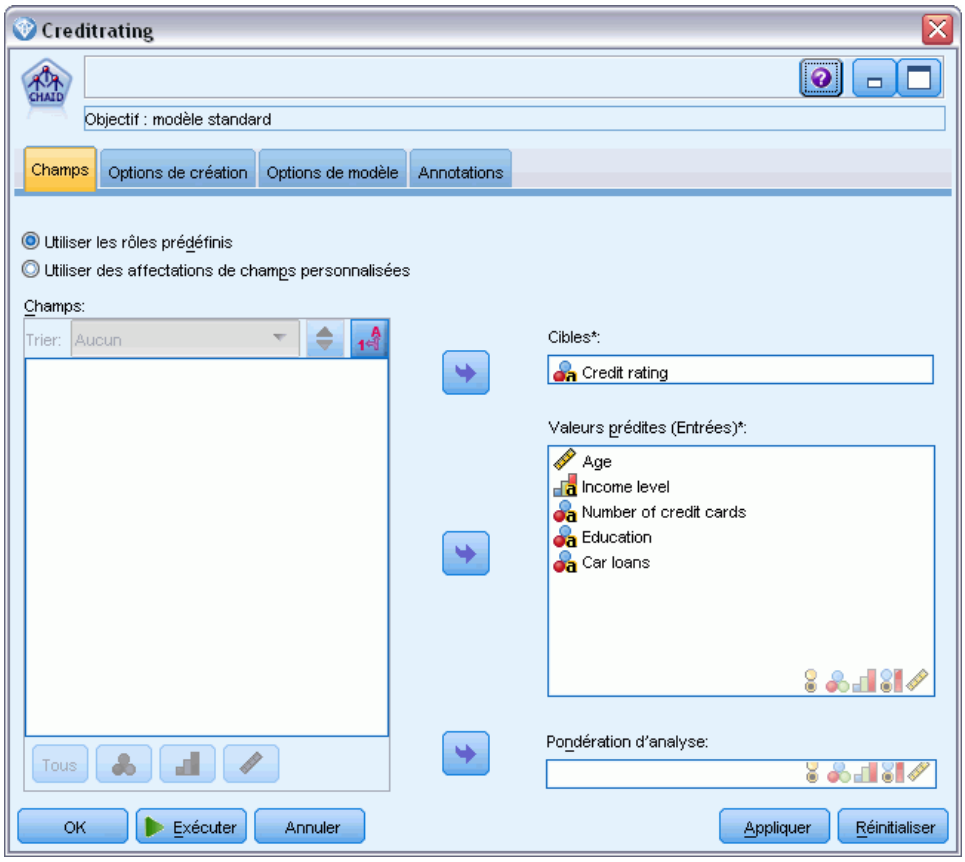
Le rôle est défini sur *Entrée* pour les autres champs. Les champs d'entrée sont quelquefois désignés sous le nom de **variables indépendantes**, ou champs dont les valeurs sont utilisées par l'algorithme de modélisation afin de prévoir la valeur du champ cible.

Le noeud de modélisation CHAID génère le modèle.

Sur l'onglet Champs du noeud de modélisation, l'option Utiliser les rôles prédéfinis est sélectionnée, ce qui signifie que la cible et les entrées indiquées dans le noeud Typer seront utilisées. Vous pouvez modifier les rôles de champ à ce stade, mais pour cet exemple, nous les utiliserons tels quels.

▶ Cliquez sur l'onglet Options de création.

Figure 3-5
Noeud de modélisation CHAID - Onglet Champs



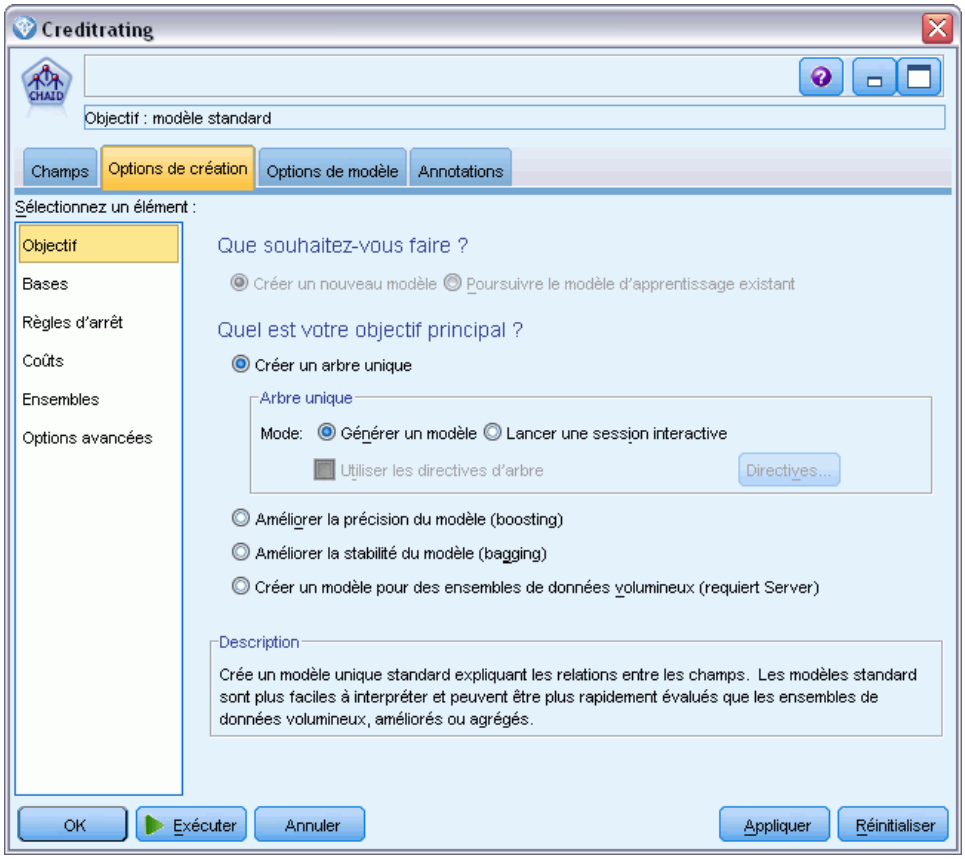
Plusieurs options sont disponibles ici dans lesquelles vous pouvez spécifier le type de modèle que vous voulez créer.

Nous voulons un tout nouveau modèle, donc nous utiliserons l'option par défaut Créer un nouveau modèle.

Nous voulons également un seul modèle d'arbre décision standard sans aucune amélioration, donc nous conserverons l'option d'objectif par défaut Créer un seul arbre.

Bien que vous puissiez lancer une session de modélisation interactive qui vous permet d'ajuster le modèle, cet exemple génère simplement un modèle à l'aide du paramètre de mode par défaut Générer le modèle.

Figure 3-6
Noeud de modélisation CHAID - Onglet Options de création

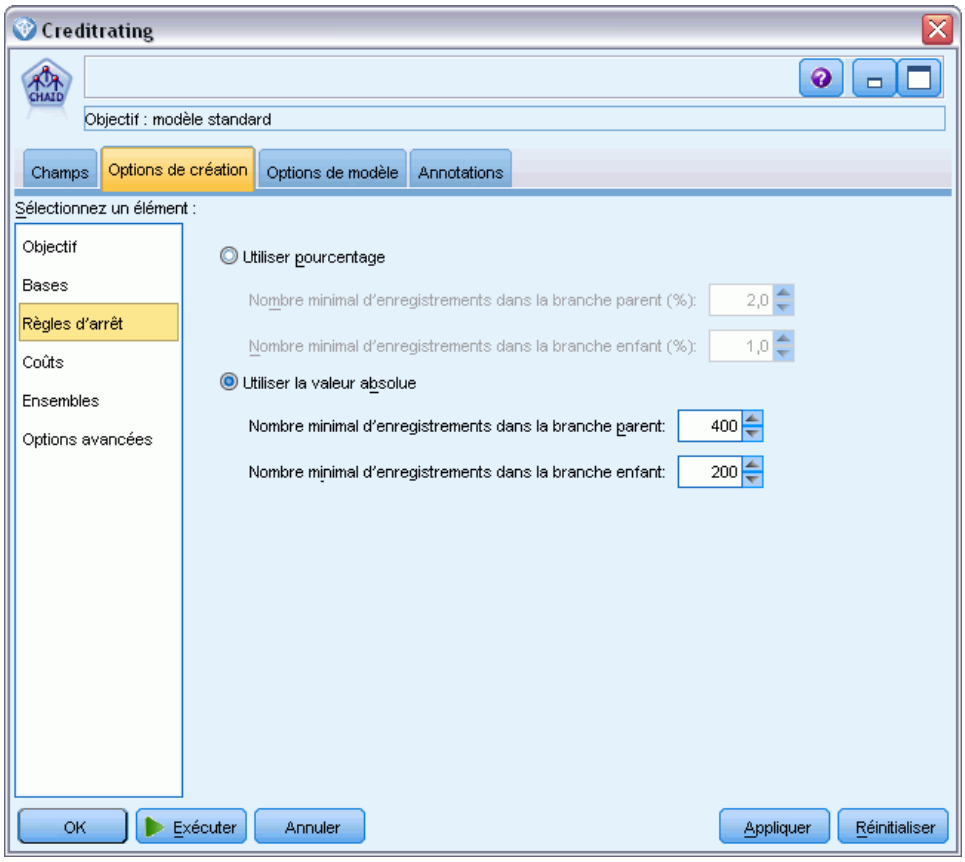


Dans cet exemple, pour que l'arbre reste simple, nous limiterons sa croissance en augmentant le nombre minimum d'observations pour les noeuds parents et enfants.

- Dans l'onglet Options de création, sélectionnez Règles d'arrêt dans le panneau de gauche du navigateur.
- Sélectionnez l'option Utiliser la valeur absolue.
- Définissez Enregistrements minimum dans la branche parent sur 400.

▶ Définissez Enregistrements minimum dans la branche enfant sur 200.

Figure 3-7
Définition des critères d'arrêt pour la création d'un arbre de décision



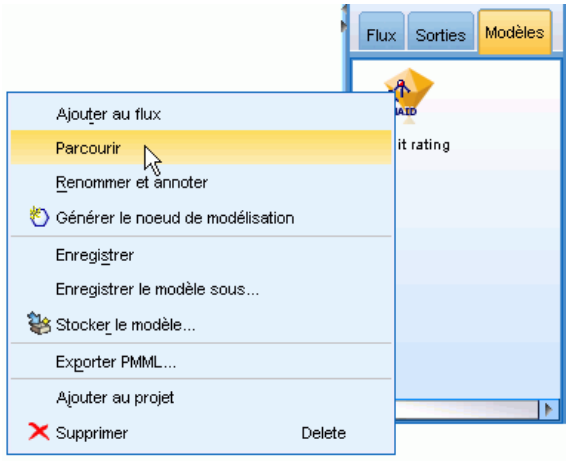
Nous pouvons utiliser toutes les autres options par défaut pour cet exemple, par conséquent, cliquez sur Exécuter pour créer le modèle. (Vous pouvez également cliquer avec le bouton droit de la souris sur le noeud et choisir Exécuter dans le menu contextuel ou sélectionner le noeud et choisir Exécuter dans le menu Outils.)

Navigation dans le modèle

Lorsque l'exécution se termine, le nugget de modèle est ajouté à la palette Modèles dans le coin supérieur droit de la fenêtre de l'application, et est aussi placé dans l'espace de travail du flux avec un lien vers le noeud de modélisation à partir duquel il a été créé. Pour consulter les détails du

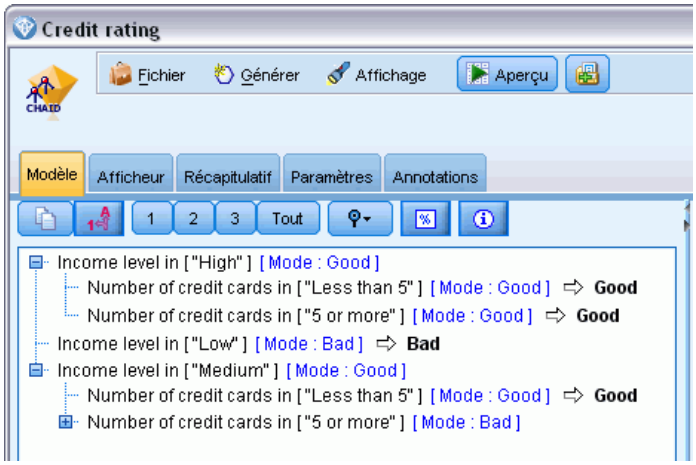
modèle, cliquez avec le bouton droit de la souris sur le nugget de modèle Parcourir (dans la palette des modèles) ou Modifier (dans l'espace de travail).

Figure 3-8
Palette Modèles



Dans le cas du nugget CHAID, l'onglet Modèle affiche les détails sous la forme d'un ensemble de règles. Il s'agit essentiellement d'une série de règles pouvant être utilisées pour affecter des enregistrements individuels à des noeuds enfant, en fonction des valeurs des différents champs d'entrée.

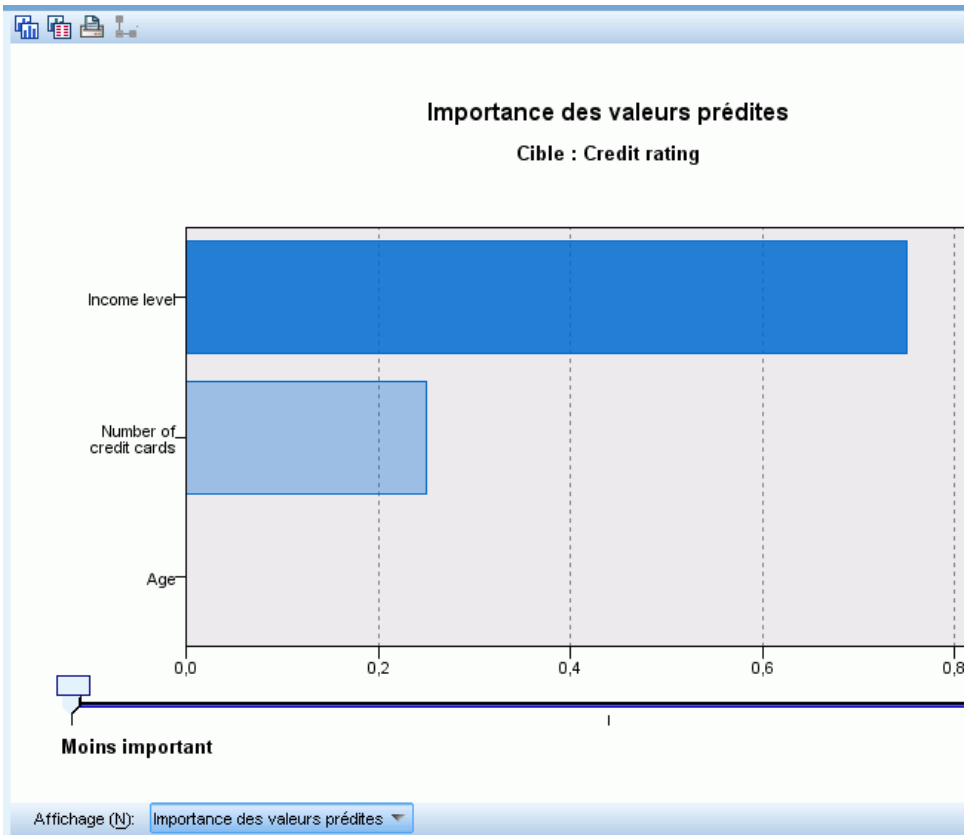
Figure 3-9
Nugget de modèle CHAID, ensemble de règles



Pour chaque noeud de terminal d'arbre de décision, c'est-à-dire ces noeuds Arbre qui ne sont pas plus divisés, une prévision de *Bon* ou *Mauvais*est renvoyée. Dans chaque cas la prédiction est déterminée par le **noeud**, ou par la réponse la plus courante pour les enregistrements qui sont compris dans ce noeud.

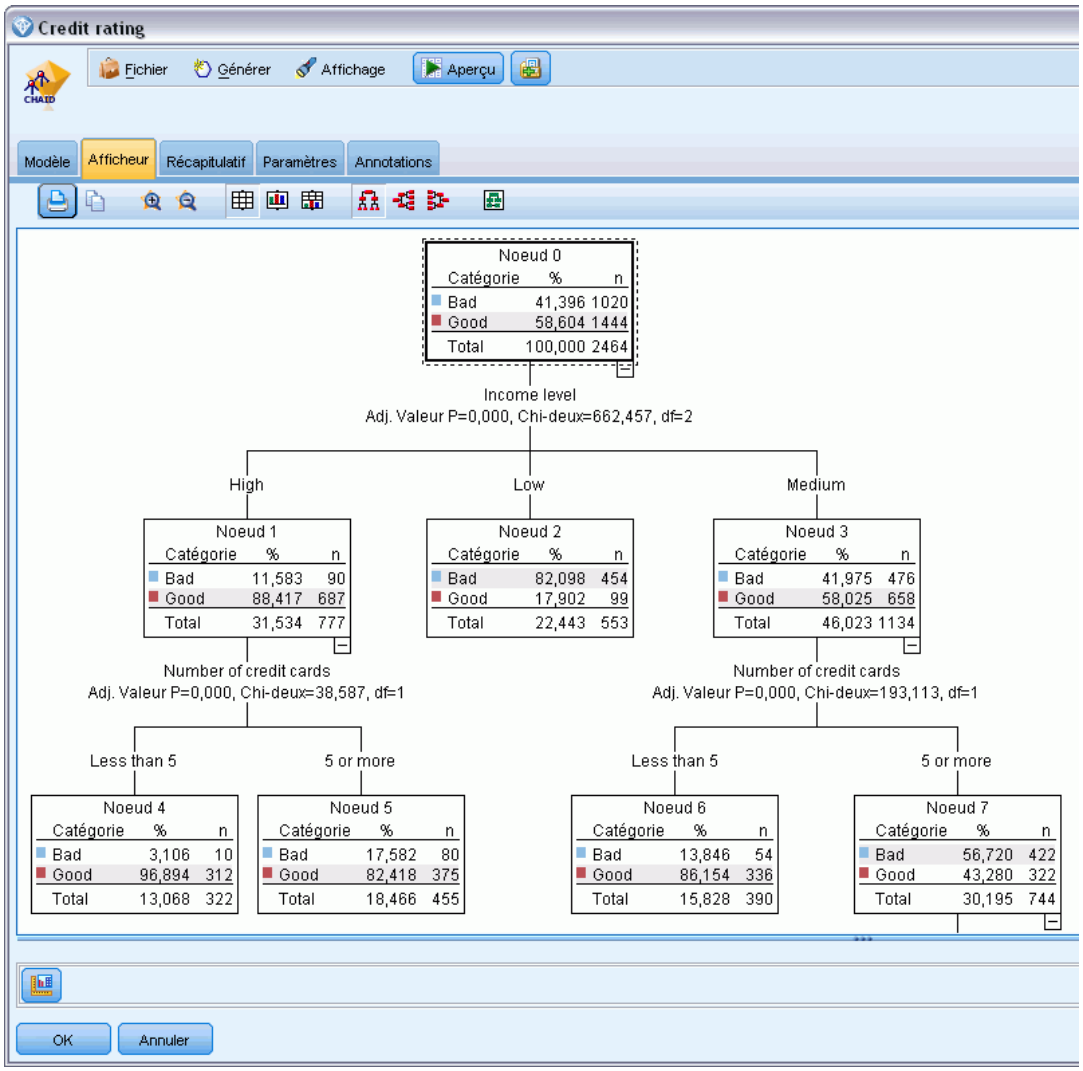
À droite de l'ensemble de règles, l'onglet Modèle affiche le tableau d'importance des variables indépendantes qui montre l'importance relative de chaque variable indépendante dans l'estimation du modèle. Nous pouvons observer que le *niveau de revenu* est le critère plus important dans ce cas et que le seul autre facteur intéressant est le *Nombre de cartes de crédit*.

Figure 3-10
Graphique de l'importance des variables indépendantes



L'onglet Afficheur dans le nugget de modèle affiche le même modèle sous la forme d'un arbre, avec un noeud à chaque point de décision. Utilisez les commandes du Zoom sur la barre d'outils pour effectuer un zoom avant sur un noeud spécifique ou un zoom arrière pour afficher une plus grande partie de l'arbre.

Figure 3-11
Onglet Afficheur dans le nugget de modèle, avec zoom arrière sélctionné



Si l'on regarde la partie supérieure de l'arbre, le premier noeud (Noeud 0) propose un récapitulatif de tous les enregistrements dans l'ensemble de données. Un peu plus de 40 % des observations de cet ensemble de données sont classées comme risquées. Il s'agit d'une proportion élevée. Voyons si l'arbre peut nous donner des informations sur les facteurs responsables.

Nous pouvons observer que la première division se situe au niveau du *Niveau de revenu*. Les enregistrements dans lesquels le niveau de revenu se trouve dans la catégorie *Bas* sont affectés au Noeud 2 et il n'est pas surprenant de voir que cette catégorie contient le plus fort pourcentage

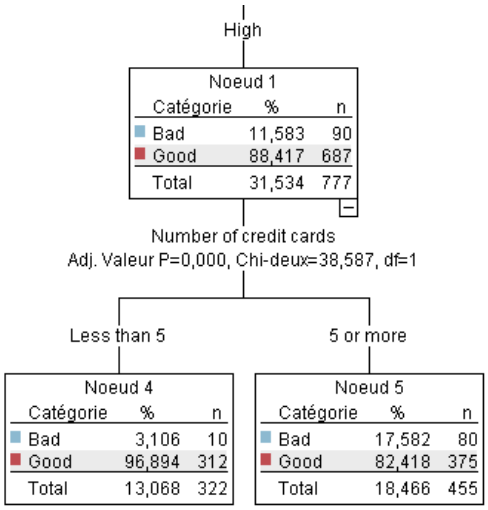
de non-reboursement de prêts. Il est évident qu'accorder un prêt aux clients de cette catégorie présente un risque élevé.

Cependant,16% des clients de cette catégorie ont, en réalité; remboursé leur prêt. Par conséquent, cette prévision n'est pas toujours exacte. Aucun modèle ne peut réellement prédire toutes les réponses, mais un bon modèle doit vous permettre de prédire la réponse *la plus problable* pour chaque enregistrement, sur la base des données disponibles.

De la même façon, si l'on observe les clients avec un revenu élevé (Noeud 1), on s'aperçoit que la grande majorité (89 %) présente un risque peu élevé. Mais plus de 1 clients sur 10 n'a pas remboursé son prêt. Est-il possible d'affiner nos critères de prêt pour diminuer le risque ?

Veuillez noter que le modèle a divisé ces clients en deux sous-catégories (noeuds 4 et 5), en fonction du nombre de cartes de crédit possédées. Pour les clients à revenu élevé, si nous prêtons uniquement à ceux possédant moins de 5 cartes de crédit, nous pouvons faire passer notre taux de succès de 89% à 97%, soit un résultat encore plus satisfaisant.

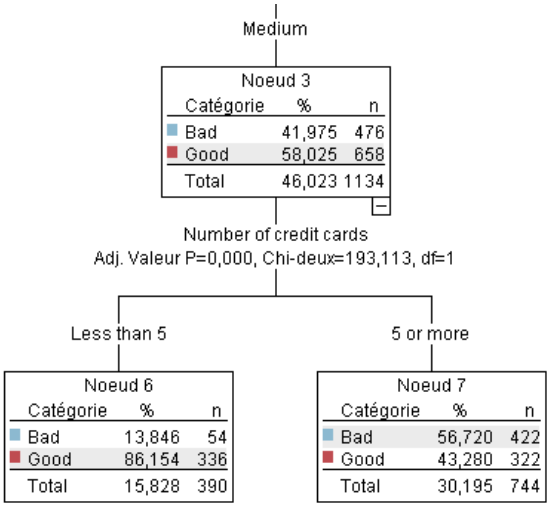
Figure 3-12
Affichage sous forme d'arbre des clients à revenu élevé



Mais qu'en est-il des clients appartenant à la catégorie Revenu moyen (Noeud 3) ? Ils sont encore plus fortement divisés entre les conditions Bonnes et Mauvaises.

De nouveau, les sous-catégories (Noeuds 6 et 7 dans ce cas) peuvent nous aider. Cette fois, prêter uniquement aux clients avec des revenus moyens et possédant moins de 5 cartes de crédit fait passer le pourcentage de conditions Bonnes de 58% à 85%, soit une augmentation importante.

Figure 3-13
Affichage sous forme d'arbre des clients à revenu moyen



Nous avons appris que chaque enregistrement contenu dans ce modèle sera attribué à un noeud spécifique et recevra une prévision *Bonne* ou *Mauvaise* en fonction des réponses les plus courantes de ce noeud.

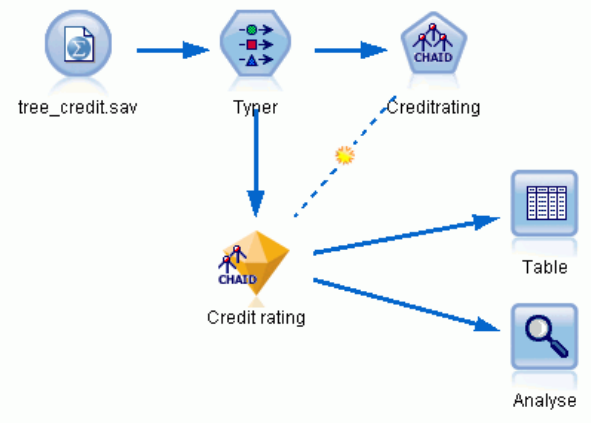
Ce processus consistant à affecter des prédictions à des enregistrements individuels s'appelle le **scoring**. En effectuant le scoring des mêmes enregistrements utilisés pour estimer le modèle, il est possible d'évaluer sa précision sur les données d'apprentissage, données dont nous connaissons le résultat. Examinons comment effectuer cette opération.

Evaluation du modèle

Nous avons parcouru le modèle pour comprendre le fonctionnement du scoring. Mais pour évaluer sa *précision*, nous devons déterminer le score de certains enregistrements et comparer les réponses prédites par le modèle aux résultats réels. Nous allons déterminer le score des mêmes

enregistrements qui ont été utilisés pour estimer le modèle, ce qui nous permet de comparer les réponses observées et les réponses prédites.

Figure 3-14
Liez le nugget de modèle au noeuds de sortie pour l'évaluation du modèle.



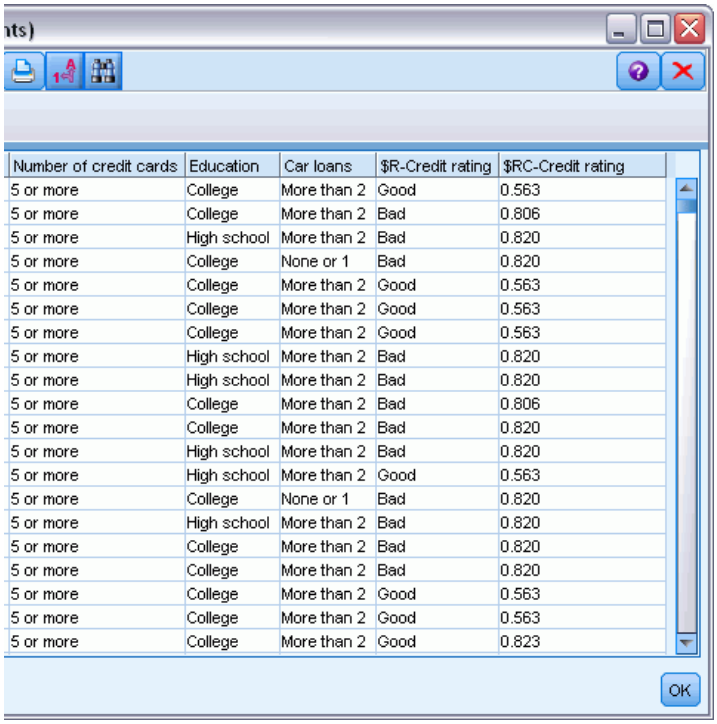
- Pour voir les scores ou les prédictions, attachez le noeud Table au nugget de modèle, double-cliquez sur le noeud Table et cliquez sur Exécutez.

  La table affiche les scores prédits dans un champ nommé *$R-Credit rating*, qui a été créé par le modèle. Nous pouvons comparer ces valeurs au champ *Conditions de crédit* d'origine qui contient les réponses réelles.

  Par convention, les noms des champs générés au cours du scoring sont déterminés en fonction du champ cible, mais avec un préfixe standard tel que *$R-* pour les prédictions ou *$RC-* pour les valeurs de confiance. Différents types de modèles utilisent différents ensembles de préfixes.

Une **valeur de confiance** est la propre estimation du modèle, sur une échelle de 0,0 à 1,0, de la précision de chaque valeur prédite.

Figure 3-15
Table affichant les scores générés et les valeurs de confiance



Number of credit cards	Education	Car loans	$R-Credit rating	$RC-Credit rating
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Bad	0.806
5 or more	High school	More than 2	Bad	0.820
5 or more	College	None or 1	Bad	0.820
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.563
5 or more	High school	More than 2	Bad	0.820
5 or more	High school	More than 2	Bad	0.820
5 or more	College	More than 2	Bad	0.806
5 or more	College	More than 2	Bad	0.820
5 or more	High school	More than 2	Bad	0.820
5 or more	High school	More than 2	Good	0.563
5 or more	College	None or 1	Bad	0.820
5 or more	High school	More than 2	Bad	0.820
5 or more	College	More than 2	Bad	0.820
5 or more	College	More than 2	Bad	0.820
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.823

Comme prévu, la valeur prédite correspond aux réponses réelles pour de nombreux enregistrements mais pas pour tous. La raison à cela est que chaque noeud terminal CHAID comporte un ensemble de réponses. La prédiction correspond à la réponse la *plus courante*, mais elle sera fausse pour toutes les autres réponses de ce noeud. (Pensez à la minorité de 16% de clients à faible revenu qui ont remboursé leur prêt).

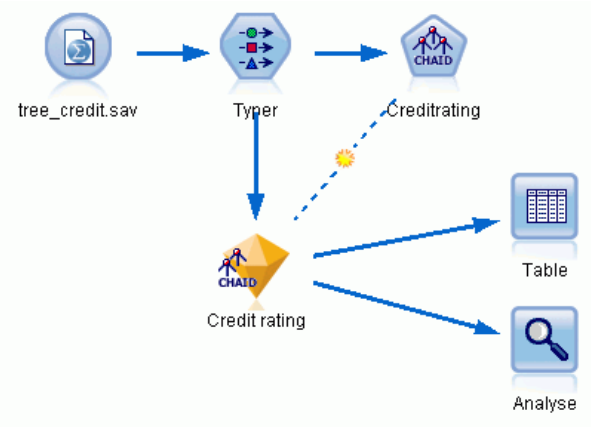
Pour éviter ceci, nous pouvons continuer à diviser l'arbre en branches de plus en plus petites, jusqu'à ce que chaque noeud soit pur à 100%, autrement dit qu'il ne comporte que des *Bonnes* ou des *Mauvaises* sans réponses mixtes. Mais un tel modèle serait extrêmement compliqué et serait probablement difficile à étendre à d'autres ensembles de données.

Pour connaître précisément le nombre de prévisions correctes, nous pouvons lire la table et compter le nombre d'enregistrements où la valeur du champ prédit *$R-Credit rating* correspond à la valeur des *Conditions de crédit*. Heureusement, il y a beaucoup plus simple : nous pouvons utiliser le noeud Analyse, qui effectue automatiquement cette opération.

- Connectez le nugget de modèle au noeud Analyse.

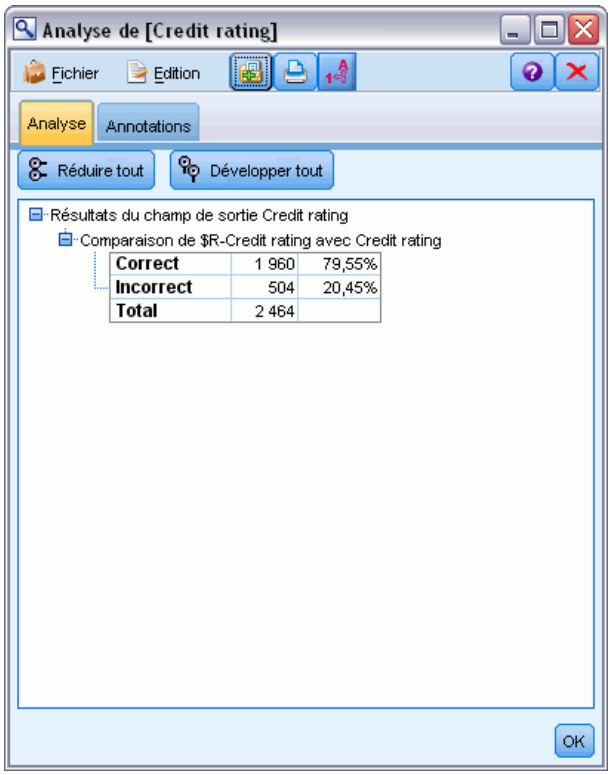
▶ Double-cliquez sur le noeud Analyse, puis cliquez sur Exécuter.

Figure 3-16
Ajout d'un noeud Analyse



L'analyse montre que pour 1899 enregistrements sur 2464 - un peu plus de 77% - la valeur prédite par le modèle correspondait à la réponse réelle.

Figure 3-17
Résultats d'analyse comparant les réponses observées et les réponses prédites



Ce resultat est limité parce que les enregistrements auxquels un score est donné sont les mêmes que ceux utilisés pour évaluer le modèle. Dans la réalité, vous pourriez utiliser un noeud Partitionner pour diviser les données en échantillons distincts pour l'apprentissage et l'évaluation.

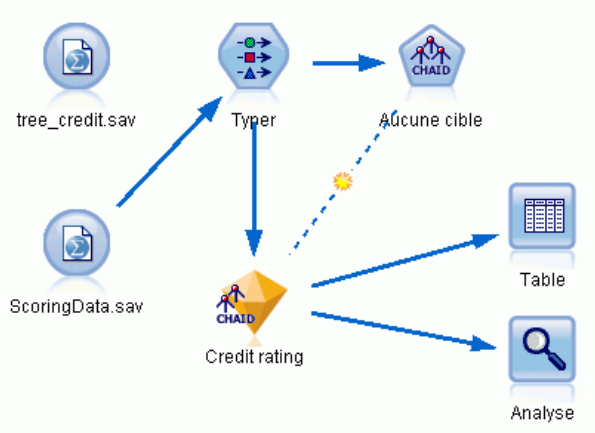
L'utilisation d'un échantillon de partition pour la génération du modèle et d'un autre échantillon pour le tester vous permet d'avoir une bien meilleure indication de la manière dont il peut s'étendre à d'autres ensembles de données.

Le noeud Analyse nous permet de tester le modèle sur les enregistrements pour lesquels nous connaissons déjà le résultat réel. L'étape suivante illustre la façon dont nous pouvons utiliser le modèle pour évaluer les enregistrements dont nous ne connaissons pas le résultat. Par exemple, cela peut comprendre les gens qui ne sont pas des clients de la banque, mais qui sont des cibles potentielles pour un publipostage promotionnel.

Scoring des enregistrements

Auparavant, nous avons évalué les mêmes enregistrements utilisés pour estimer le modèle afin de connaître la précision du modèle. À présent, nous allons voir comme évaluer un ensemble d'enregistrements différent de ceux utilisés pour créer le modèle. Il s'agit de l'objectif de la modélisation avec un champ cible : étudier les enregistrements pour lesquels vous connaissez le résultat pour identifier des schémas qui vous permettront de prédire les résultats que vous ne connaissez pas encore.

Figure 3-18
Association de nouvelles données pour le scoring



Vous pouvez mettre à jour le noeud source Statistics pour qu'il pointe vers un fichier de données différent ou vous pouvez ajouter un nouveau noeud source qui lit dans les données que vous voulez évaluer. Dans les deux méthodes, le nouvel ensemble de données doit contenir les mêmes champs d'entrée utilisés par le modèle (*Age*, *Niveau de revenu*, *Education*, etc.) mais pas le champ cible *Conditions de crédit*.

Vous pouvez également ajouter le nugget de modèle à tout flux contenant les champs d'entrée attendus. Qu'il soit lu à partir d'un fichier ou d'une base de données, le type de source n'importe pas du moment que les noms et les types des champs correspondent à ceux utilisés par le modèle.

Vous pouvez également enregistrer le nugget de modèle en tant que fichier distinct, exporter le modèle au format PMML pour une utilisation avec d'autres applications qui prennent en charge ce format ou stocker le modèle dans un répertoire IBM® SPSS® Collaboration and Deployment Services, ce qui permet le déploiement, le scoring et la gestion des modèles à l'échelle de l'entreprise.

Quelque soit l'infrastructure utilisée, le modèle proprement dit fonctionne de la même manière.

Récapitulatif

Cet exemple décrit la procédure standard de création, d'évaluation et de scoring d'un modèle.

- Le noeud de modélisation estime le modèle en étudiant les enregistrements pour lesquels le résultat est connu et crée un nugget de modèle. On parle parfois d'apprentissage du modèle.
- Le nugget de modèle peut être ajouté à n'importe quel flux contenant les champs attendus pour évaluer les enregistrements. En effectuant le scoring des enregistrements pour lesquels vous connaissez déjà le résultat (les clients existants par exemple), vous pouvez évaluer la performance du modèle.
- Une fois que vous êtes satisfait de la performance du modèle, vous pouvez effectuer un scoring de nouvelles données (des clients potentiels par exemple) pour prédire leur réponse.
- Les données utilisées pour l'apprentissage ou l'estimation du modèle peuvent être appelées données analytiques ou historiques; les données de scoring peuvent également être appelées données opérationnelles.

Chapitre
4

Modélisation automatisée d'une cible de type booléen

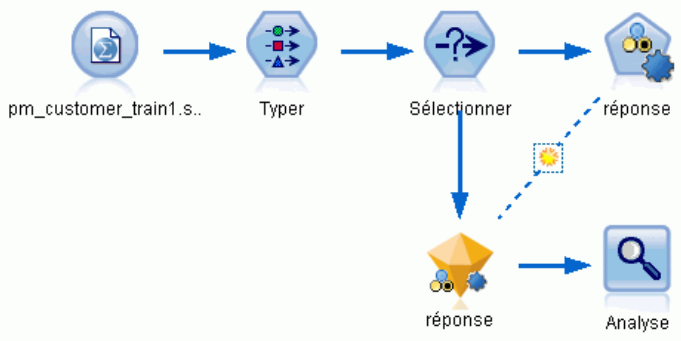
Modélisation de la réponse client (Classificateur automatique)

Le noeud Classificateur automatique vous permet de créer et de comparer automatiquement différents modèles pour les cibles de type booléen (comme la probabilité selon laquelle un client donné est susceptible ou non de rembourser une échéance de prêt ou de répondre à une offre spécifique) ou les cibles (d'ensemble) nominales . Dans cet exemple, nous allons rechercher un résultat booléen (oui ou non). Dans un flux relativement simple, le noeud génère et classe un ensemble de modèles candidats, choisit les meilleurs et les combine en un modèle (combiné) agrégé unique. Cette approche conjugue la facilité de l'automatisation aux avantages de combiner plusieurs modèles ce qui permet généralement des prédictions plus précises que celles de tout autre modèle.

Cet exemple repose sur une société fictive qui souhaite obtenir des résultats plus rentables en présentant à chaque client une offre adaptée.

Cette approche souligne les avantages de l'automatisation. Pour un autre exemple qui utilise une cible continue (d'intervalle numérique), consultez la rubrique le chapitre 5 sur p. 57.

Figure 4-1
Exemple de flux Classificateur automatique



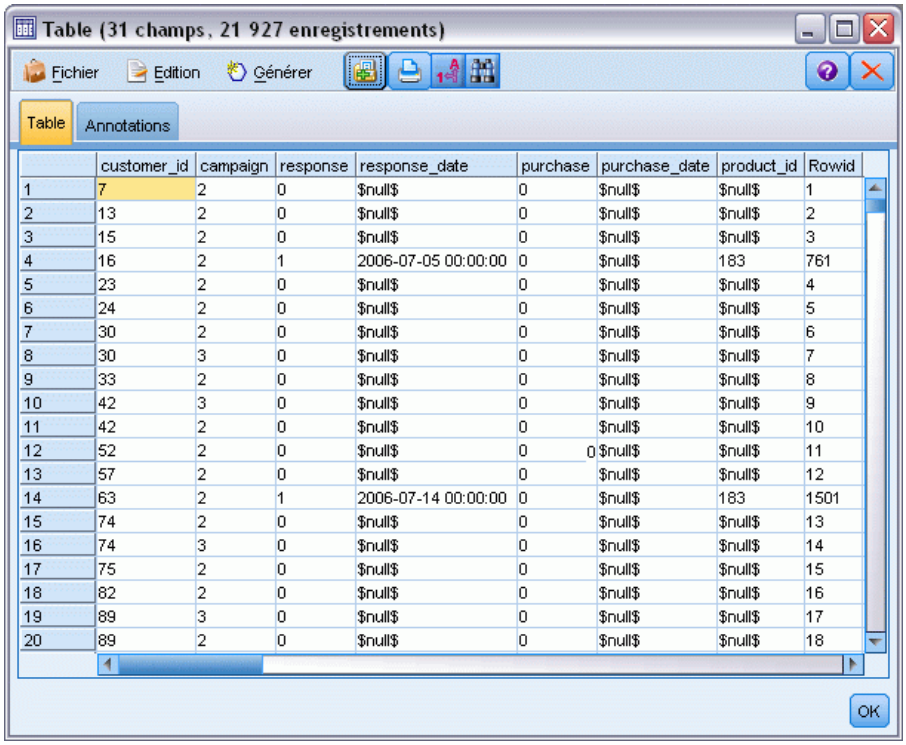
Cet exemple utilise le flux *pm_binaryclassifier.str*, installé dans le dossier Démo dans le répertoire des *flux*. Le fichier de données est *pm_customer_train1.sav*. Pour plus d'informations, reportez-vous à la section Dossier Demos dans le chapitre 1 sur p. 6.

Données historiques

Le fichier *pm_customer_train1.sav* comporte des données historiques suivant les offres faites à des clients spécifiques au cours de campagnes passées, comme l'indique la valeur du champ *campaign*. Le plus grand nombre d'enregistrements se trouve dans la campagne *Premium account*.

Les valeurs du champ *campaign* sont en fait codées comme des entiers dans les données (par exemple *2 = Premium account*). Plus tard, vous définirez les étiquettes de ces valeurs qui vous permettront d'obtenir des résultats plus probants.

Figure 4-2
Données sur les anciennes promotions



	customer_id	campaign	response	response_date	purchase	purchase_date	product_id	Rowid
1	7	2	0	$null$	0	$null$	$null$	1
2	13	2	0	$null$	0	$null$	$null$	2
3	15	2	0	$null$	0	$null$	$null$	3
4	16	2	1	2006-07-05 00:00:00	0	$null$	183	761
5	23	2	0	$null$	0	$null$	$null$	4
6	24	2	0	$null$	0	$null$	$null$	5
7	30	2	0	$null$	0	$null$	$null$	6
8	30	3	0	$null$	0	$null$	$null$	7
9	33	2	0	$null$	0	$null$	$null$	8
10	42	3	0	$null$	0	$null$	$null$	9
11	42	2	0	$null$	0	$null$	$null$	10
12	52	2	0	$null$	0	$null$	$null$	11
13	57	2	0	$null$	0	$null$	$null$	12
14	63	2	1	2006-07-14 00:00:00	0	$null$	183	1501
15	74	2	0	$null$	0	$null$	$null$	13
16	74	3	0	$null$	0	$null$	$null$	14
17	75	2	0	$null$	0	$null$	$null$	15
18	82	2	0	$null$	0	$null$	$null$	16
19	89	3	0	$null$	0	$null$	$null$	17
20	89	2	0	$null$	0	$null$	$null$	18

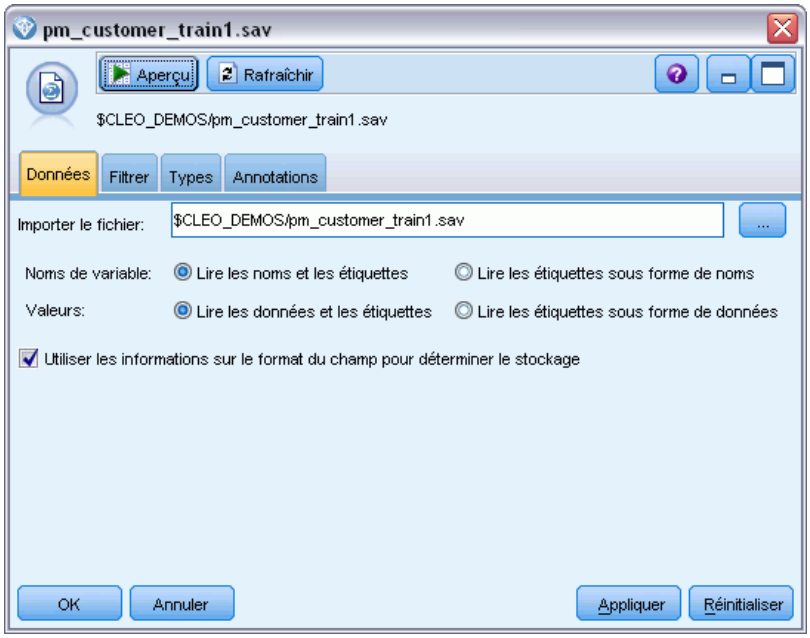
Ce fichier comprend également un champ *réponse* qui indique si l'offre a été acceptée (0 = *non*, et 1 = *oui*). Il s'agit du **champ cible**, ou la valeur, que vous souhaitez prédire. Plusieurs champs contenant des informations démographiques et financières sur chaque client ont également été ajoutés. Ces champs peuvent permettre de créer ou de "former" un modèle qui prédit les taux de réponse des individus ou des groupes en fonction de caractéristiques telles que le revenu, l'âge ou le nombre de transactions mensuelles.

Création du flux

- Ajoutez un noeud source Statistics qui pointe sur *pm_customer_train1.sav*, dans le dossier *Demos* du répertoire d'installation de IBM® SPSS® Modeler. (Vous pouvez saisir `$CLEO_DEMOS/` dans le chemin d'accès comme raccourci permettant de référencer ce dossier. Veuillez noter qu'une

barre oblique (/) —plutôt qu'une barre oblique inverse (\)— doit être utilisée dans le chemin d'accès, comme indiqué.)

Figure 4-3
Lecture de données



- Ajoutez un noeud Typer, puis sélectionnez *Réponse* en tant que champ cible (Rôle = Cible). Paramétrez l'option Mesure de ce champ sur Booléen.

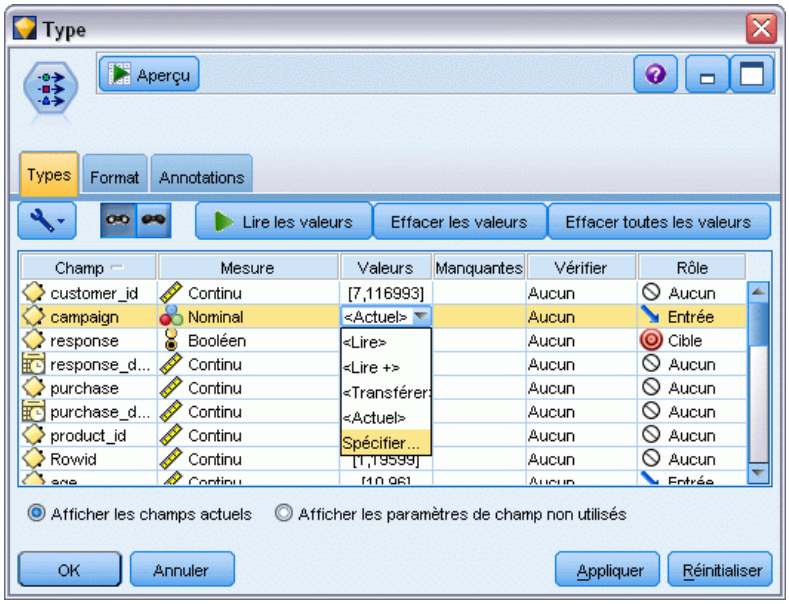
Figure 4-4
Configuration du niveau de mesure et du rôle

- Paramétrez le rôle sur Aucun pour les champs suivants : *customer_id*, *campaign*, *response_date*, *purchase*, *purchase_date*, *product_id*, *Rowid* et *X_random*. Ces champs seront ignorés lors de la création du modèle.

- Cliquez sur le bouton Lire les valeurs dans le noeud Typer pour vérifier que les valeurs sont instanciées.

  Comme nous l'avons vu auparavant, nos données source contiennent des informations sur quatre campagnes différentes, chacune visant un type de compte client différent. Ces campagnes sont codées comme entiers dans les données, et pour se rappeler plus facilement quel type de compte chaque entier représente, définissons les étiquettes de chacun d'eux.

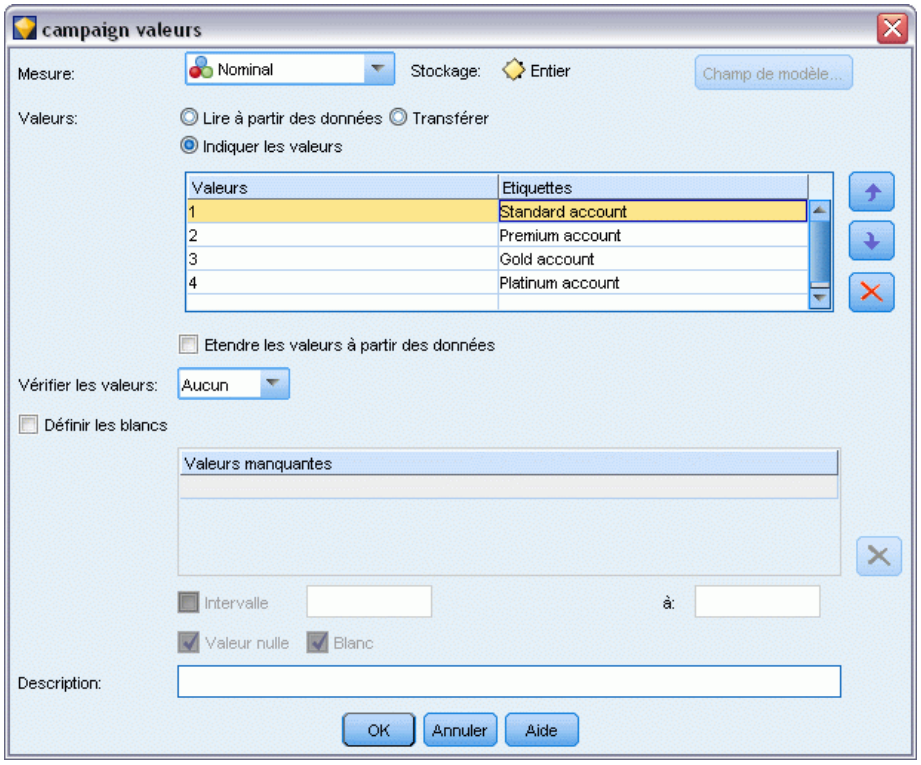
Figure 4-5
Choix de spécification des valeurs d'un champ



- Sur la ligne du champ campaign, cliquez sur l'entrée dans la colonne Valeurs.

- Sélectionnez Spécifier dans la liste déroulante.

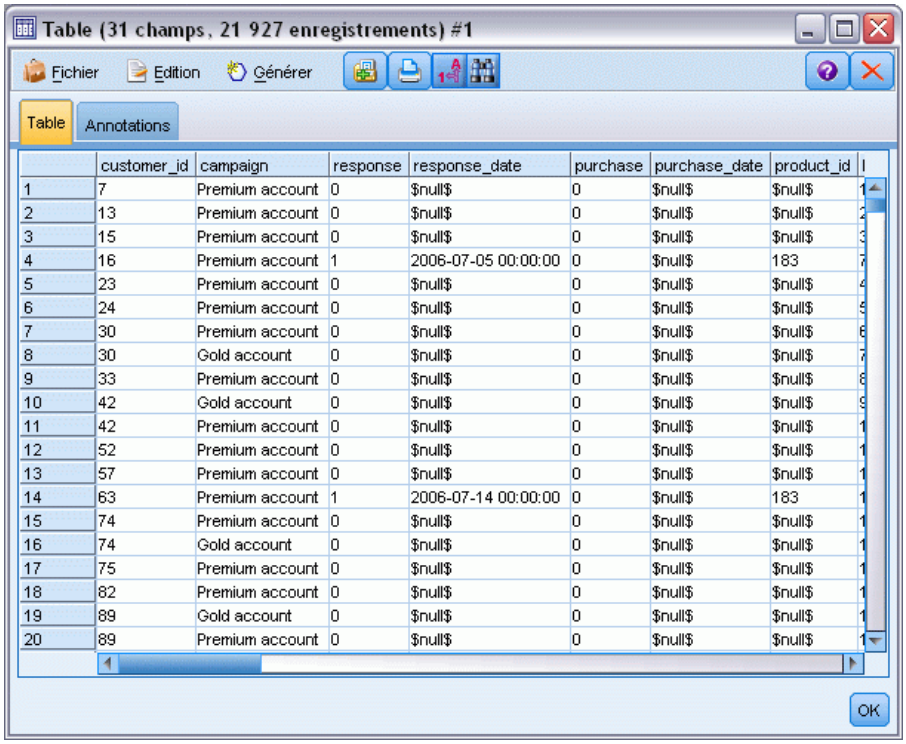
Figure 4-6
Définition des étiquettes pour les valeurs de champ



- Dans la colonne Etiquettes, saisissez les étiquettes comme indiqué pour chacune des quatre valeurs du champ campaign.
- Cliquez sur OK.

Vous pouvez maintenant afficher les étiquettes dans les fenêtres de sortie plutôt que les entiers.

Figure 4-7
Affichage des étiquettes de valeur de champ



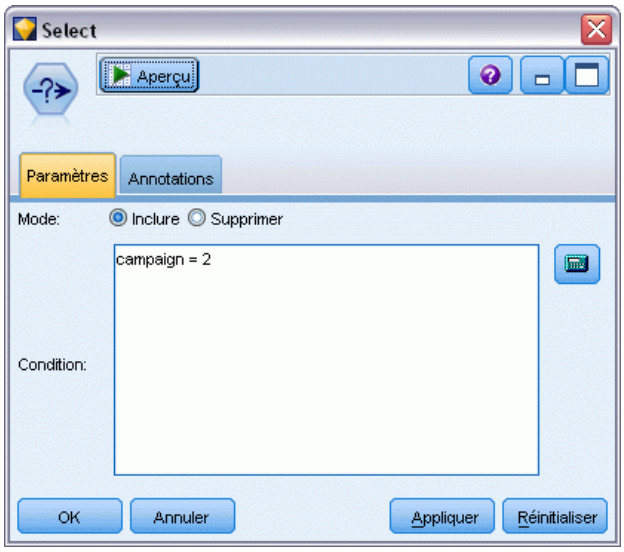
	customer_id	campaign	response	response_date	purchase	purchase_date	product_id	l
1	7	Premium account	0	$null$	0	$null$	$null$	1
2	13	Premium account	0	$null$	0	$null$	$null$	2
3	15	Premium account	0	$null$	0	$null$	$null$	3
4	16	Premium account	1	2006-07-05 00:00:00	0	$null$	183	7
5	23	Premium account	0	$null$	0	$null$	$null$	4
6	24	Premium account	0	$null$	0	$null$	$null$	5
7	30	Premium account	0	$null$	0	$null$	$null$	6
8	30	Gold account	0	$null$	0	$null$	$null$	7
9	33	Premium account	0	$null$	0	$null$	$null$	8
10	42	Gold account	0	$null$	0	$null$	$null$	9
11	42	Premium account	0	$null$	0	$null$	$null$	1
12	52	Premium account	0	$null$	0	$null$	$null$	1
13	57	Premium account	0	$null$	0	$null$	$null$	1
14	63	Premium account	1	2006-07-14 00:00:00	0	$null$	183	1
15	74	Premium account	0	$null$	0	$null$	$null$	1
16	74	Gold account	0	$null$	0	$null$	$null$	1
17	75	Premium account	0	$null$	0	$null$	$null$	1
18	82	Premium account	0	$null$	0	$null$	$null$	1
19	89	Gold account	0	$null$	0	$null$	$null$	1
20	89	Premium account	0	$null$	0	$null$	$null$	1

- Reliez un noeud Table au noeud Typer.
- Ouvrez le noeud Table, puis cliquez sur Exécuter.
- Dans la fenêtre de sortie, cliquez sur le bouton de la barre d'outils Afficher les étiquettes de champ et de valeur pour afficher les étiquettes.
- Cliquez sur OK pour fermer la fenêtre de sortie.

Les données incluent des informations sur quatre campagnes différentes, mais vous vous concentrerez sur l'analyse d'une seule campagne à la fois. Comme le plus grand nombre d'enregistrements se trouve dans la campagne de compte Premium (codée *campaign=2* dans les

données), vous pouvez utiliser un noeud Sélectionner pour n'inclure que ces enregistrements dans le flux.

Figure 4-8
Sélection d'enregistrements pour une seule campagne

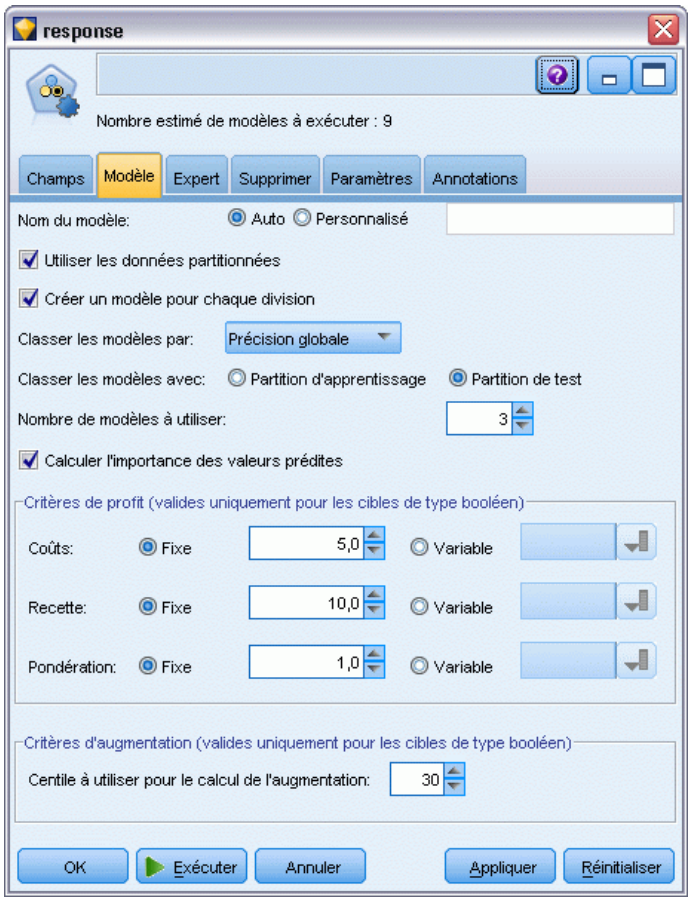


Génération et comparaison de modèles

- Liez un noeud Classificateur automatique et sélectionnez Précision globale comme système métrique utilisé pour classer les modèles.

- Définissez le Nombre de modèles à utiliser sur 3. Cela signifie que les trois meilleurs modèles seront créés lorsque vous exécuterez le noeud.

Figure 4-9
Noeud Classificateur automatique - Onglet Modèle

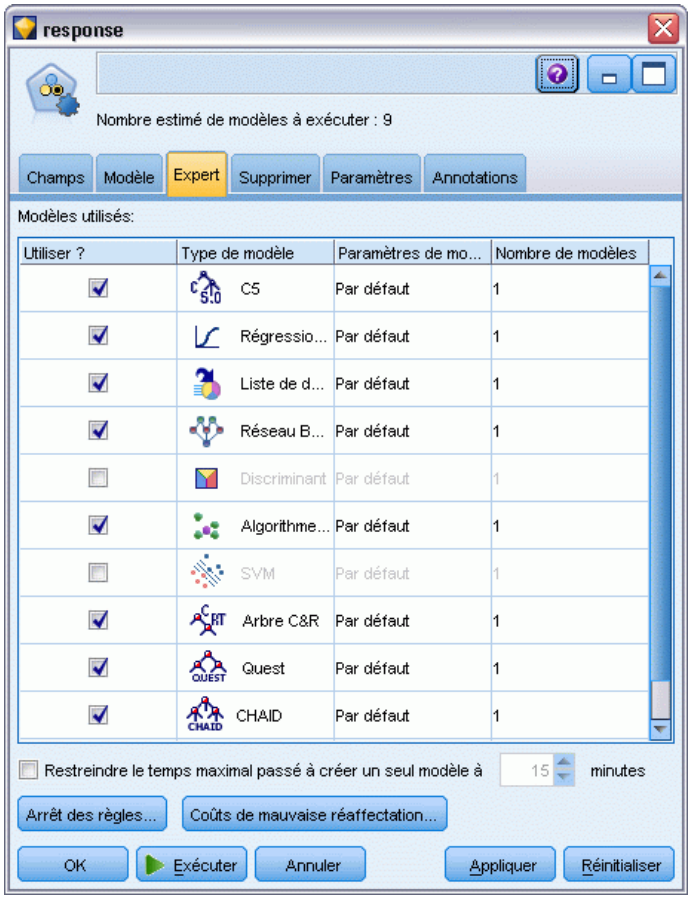


Dans l'onglet Expert, vous pouvez choisir jusqu'à 11 algorithmes de modèle différents.

- Désélectionnez les types de modèle Discriminant et SVM. (Ces modèles prennent plus longtemps à se former à partir de ces données et les désélectionner acccélèrera l'exemple. Mais si patienter ne vous dérange pas, n'hésitez pas à les laisser sélectionnés).

Comme vous avez défini le Nombre de modèles à utiliser sur 3 dans l'onglet Modèle, le noeud calcule la précision des neuf algorithmes restants et crée un nugget de modèle unique contenant les trois plus précis.

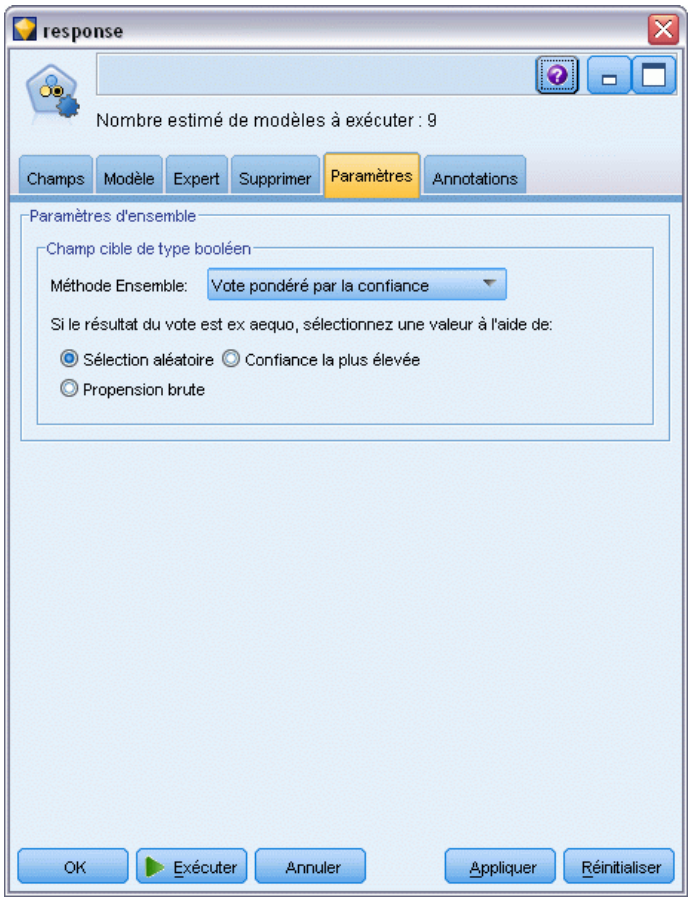
Figure 4-10
Noeud Classificateur automatique - Onglet Expert



- Dans l'onglet Paramètres, pour la méthode d'ensemble, sélectionnez Vote pondéré par la confiance. Cela détermine la façon dont un score agrégé unique est produit pour chaque enregistrement.

Avec le vote simple, si deux modèles sur trois prédisent *oui*, alors*oui* l'emporte par un vote de 2 contre 1. Dans le cas de vote pondéré par la confiance, les votes sont pondérés en fonction de la valeur de confiance de chaque prévision. Par conséquent, si un modèle prévoit *non* avec un niveau de confiance plus élevé que les deux prévisions *oui* combinées, alors*non* l'emporte.

Figure 4-11
Noeud Classificateur automatique : onglet Paramètres



► Cliquez sur Exécuter.

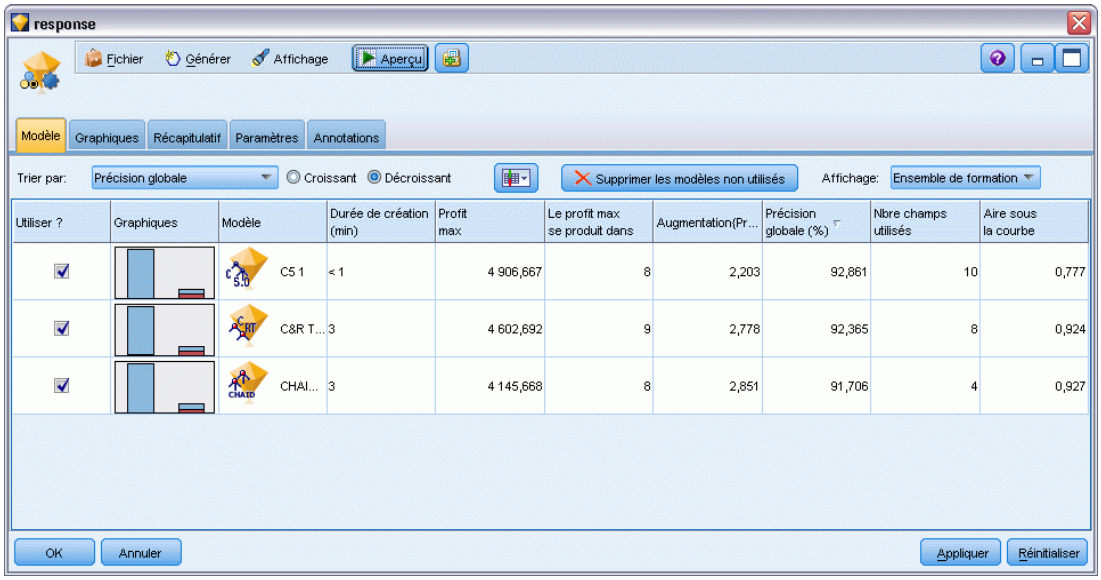
Après quelques minutes, le nugget de modèle généré est créé et placé sur l'espace de travail et dans la palette Modèles en haut à droite de la fenêtre. Vous pouvez parcourir le nugget de modèle ou l'enregistrer ou le déployer de plusieurs façons.

Ouvrez le nugget de modèle ; il répertorie les détails concernant chacun des modèles créés au cours de l'exécution. (En situation réelle, lorsque des centaines de modèles peuvent être créés à partir d'un grand nombre de données, cette opération peut prendre plusieurs heures.) Consultez Figure 4-1 sur p. 45.

Si vous souhaitez analyser plus en détail l'un des modèles individuels, vous pouvez double-cliquer sur un nugget de modèles dans la colonne Modèle pour la faire défiler et parcourir les résultats du modèle individuel ; à partir de là, vous pouvez générer des noeuds de modélisation, des nuggets de

modèle ou des graphiques d'évaluation. Dans la colonne Graphique, vous pouvez double-cliquer sur une miniature pour générer un graphique en grandeur nature.

Figure 4-12
Résultats du Classificateur automatique



Par défaut, les modèles sont classés en fonction de la précision globale, cette mesure ayant été sélectionnée dans l'onglet Modèle du noeud Classificateur automatique. Le modèle C51 se classe en meilleure position selon cette mesure, mais les modèles Arbre C&R et CHAID sont presque aussi précis.

Vous pouvez effectuer le tri sur une autre colonne en cliquant sur l'en-tête de cette colonne ou vous pouvez choisir la mesure désirée dans la liste déroulante Trier par de la barre d'outils.

En fonction de ses résultats, vous pouvez décider d'utiliser les trois modèles les plus précis. En combinant les prévisions à partir de plusieurs modèles, il est possible d'éviter les limitations dans les modèles individuels. Ce qui entraîne une plus grande précision globale.

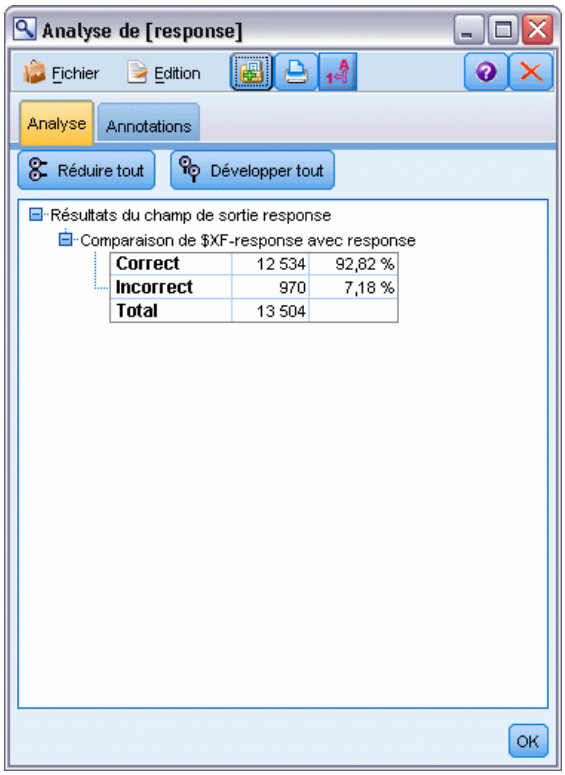
Dans la colonne Utiliser ?, sélectionnez les modèles C51, Arbre C&R et CHAID.

Liez un noeud Analyse (palette Sortie) après le nugget de modèle. Cliquez avec le bouton droit de la souris sur le noeud Analyse et sélectionnez Exécuter pour exécuter le flux.

Le score agrégé généré par le modèle combiné est affiché dans un champ nommé *$XF-response*. Lorsque les valeurs prédites sont mesurées en fonction des données d'apprentissage, elles correspondent à la réponse réelle (comme enregistrées dans le champ *réponse* d'origine) avec une précision globale de 92,82%.

Bien que ce modèle ne soit pas aussi précis que le meilleur des trois modèles individuels (92,86 % pour C51), la différence est trop minime pour être significative. Généralement, un modèle combiné sera plus performant lorsqu'il sera appliqué à des ensembles de données autres que les données d'apprentissage.

Figure 4-13
Analyse des trois modèles combinés



Récapitulatif

Pour résumer, vous avez utilisé le noeud Classificateur automatique pour comparer plusieurs modèles différents, vous avez utilisé les trois modèles les plus précis et vous les avez ajoutés au flux dans un nugget de modèle Classificateur automatique combiné.

- Concernant la précision globale, les modèles C51, Arbre C&R et CHAID sont plus performants avec les données d'apprentissage.
- Le modèle combiné a presque été aussi performant que le meilleur des modèles individuels et peut être aussi efficace lorsqu'il est appliqué à d'autres ensembles de données. Si votre objectif est d'automatiser autant que possible le processus, cette approche vous permet d'obtenir un modèle fiable dans la plupart des circonstances sans avoir à creuser trop dans les spécificités des modèles.

Chapitre 5

Modélisation automatisée d'une cible continue

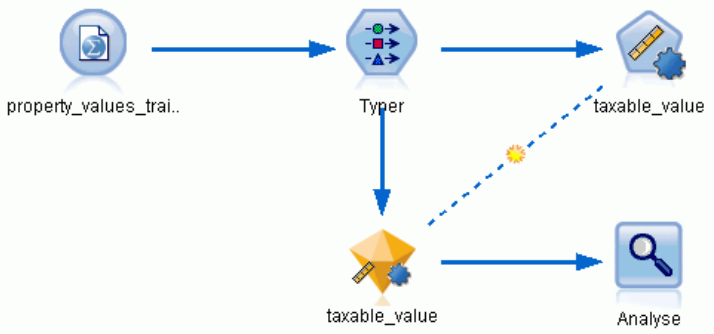
Valeurs de propriété (Numérisation automatique)

Le noeud Numérisation automatique vous permet de créer et de comparer automatiquement différents modèles pour des résultats continus (intervalle numérique), tels que la prévision de la valeur imposable d'une propriété. Avec un seul noeud, vous pouvez estimer et comparer un ensemble de modèles candidats et générer un sous-ensemble de modèles pour des analyses ultérieures. Ce noeud fonctionne de la même manière que le noeud Classificateur automatique mais pour les cibles continues plutôt que pour les cibles booléennes ou les cibles nominales.

Le noeud combine le meilleur des modèles candidats dans un nugget de modèle agrégé (d'ensemble) unique. Cette approche conjugue la facilité de l'automatisation aux avantages de combiner plusieurs modèles ce qui permet généralement des prédictions plus précises que celles de tout autre modèle.

Cet exemple se concentre sur un responsable de municipalité fictif qui ajuste et estime les taxes foncières. Pour obtenir une plus grande précision, il va construire un modèle qui prédit les valeurs immobilières en fonction du type de bâtiment, du voisinage, de la taille et d'autres facteurs connus.

Figure 5-1
Exemple de flux Numérisation automatique



Cet exemple utilise le flux *property_values_numericpredictor.str*, installé dans le dossier Démos dans le répertoire des *flux*. Le fichier de données utilisé est *property_values_train.sav*. Pour plus d'informations, reportez-vous à la section Dossier Demos dans le chapitre 1 sur p. 6.

Données d'apprentissage

Le fichier de données comprend un champ nommé *taxable_value*, qui est le **champ cible**, ou la valeur à prédire. Les autres champs contiennent des informations telles que le voisinage, le type de bâtiment et le volume intérieur et peuvent être utilisés comme variables indépendantes.

Nom de champ	Etiquette
property_id	ID propriété
voisinage	Zone à l'intérieur de la ville
building_type	Type de bâtiment
year_built	Année de construction
volume_interior	Volume intérieur
volume_other	Volume du garage et des bâtiments supplémentaires
lot_size	Taille du lot
taxable_value	Valeur imposable

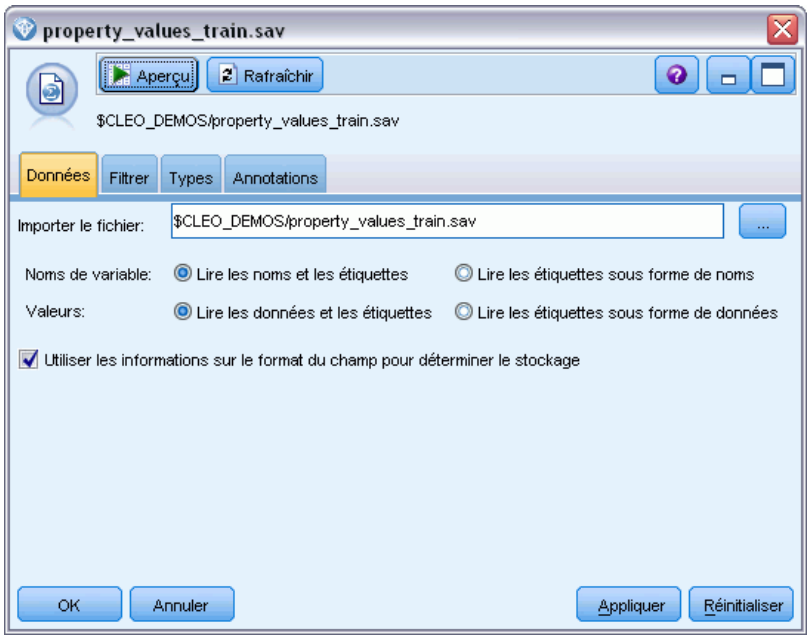
Le dossier Démos contient également un fichier de données de scoring nommé *property_values_score.sav*. Ce fichier contient les mêmes champs mais sans le champ *taxable_value* . Après la formation des modèles à l'aide des ensembles de données où la valeur imposable est connue, vous pouvez évaluer des enregistrements où cette valeur ne l'est pas.

Création du flux

- Ajoutez un noeud source Statistics qui pointe sur *property_values_train.sav*, dans le dossier *Demos* du répertoire d'installation de IBM® SPSS® Modeler. (Vous pouvez saisir `$CLEO_DEMOS/` dans le chemin d'accès comme raccourci permettant de référencer ce dossier. Veuillez noter qu'une

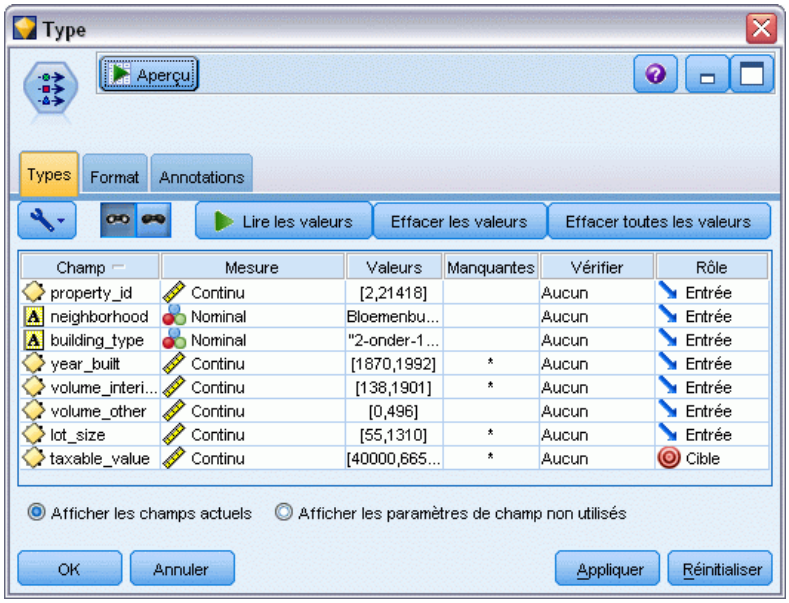
barre oblique (/) —plutôt qu'une barre oblique inverse (\)— doit être utilisée dans le chemin d'accès, comme indiqué.)

Figure 5-2
Lecture de données



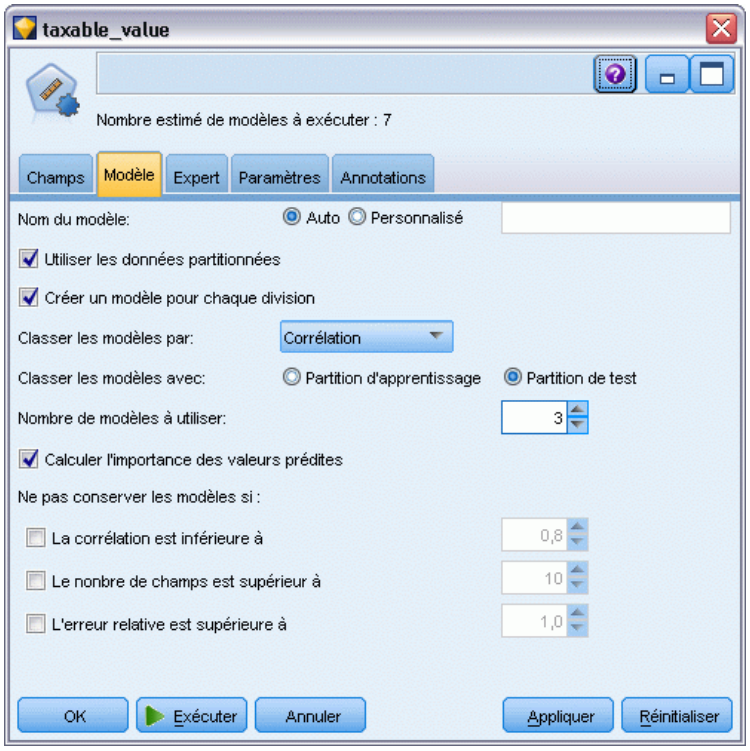
- Ajoutez un noeud Typer, puis sélectionnez *taxable_value* en tant que champ cible (Rôle = Cible). Le rôle doit être défini sur Entrée pour tous les autres champs, indiquant ainsi qu'ils seront utilisés comme variables indépendantes.

Figure 5-3
Définition du champ cible



- Liez un noeud Numérisation automatique et sélectionnez Corrélation comme mesure utilisée pour classer les modèles.
- Définissez le Nombre de modèles à utiliser sur 3. Cela signifie que les trois meilleurs modèles seront créés lorsque vous exécuterez le noeud.

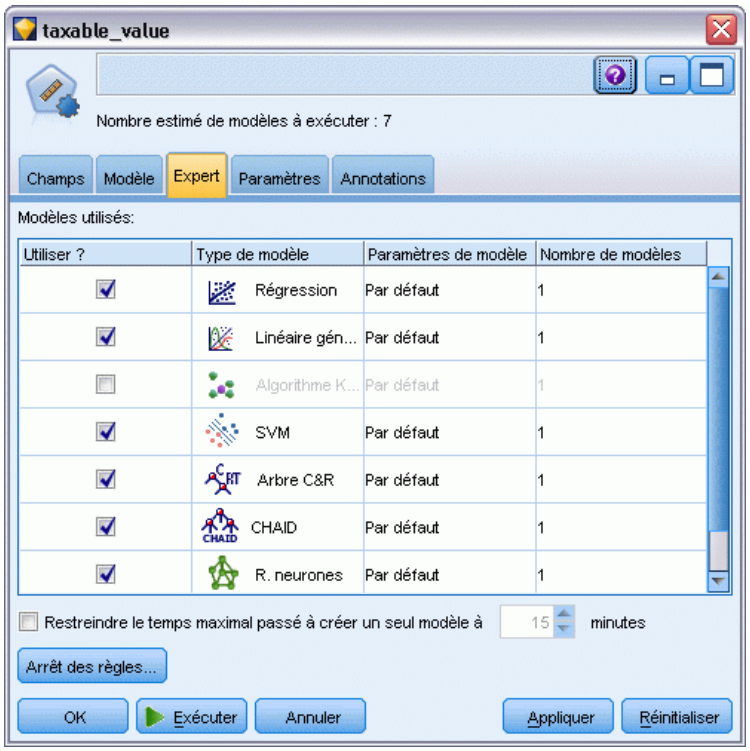
Figure 5-4
Noeud Numérisation automatique - Onglet Modèle



- Dans l'onglet Expert, laissez les paramètres par défaut ; le noeud estime un modèle unique pour chaque algorithme, pour un total de sept modèles. (Vous pouvez également modifier ces paramètres pour comparer plusieurs variantes pour chaque type de modèle.)

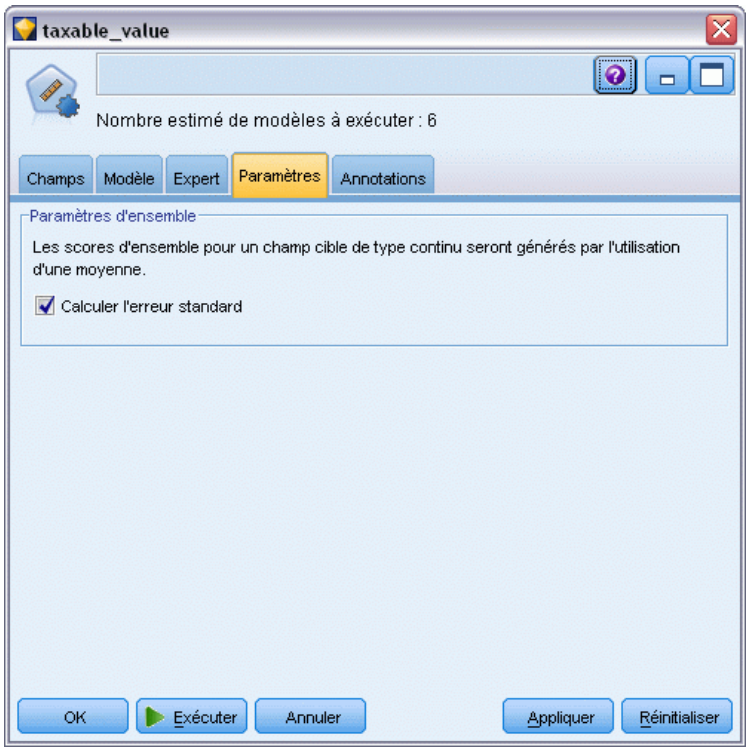
Comme vous avez défini le Nombre de modèles à utiliser sur 3 dans l'onglet Modèle, le noeud calcule la précision des sept algorithmes restants et crée un nugget de modèle simple contenant les trois plus précis.

Figure 5-5
Noeud Numérisation automatique - Onglet Expert



▶ Dans l'onglet Paramètres, laissez les paramètres par défaut tels quels. Parce qu'il s'agit d'une cible continue, le score d'ensemble est généré en effectuant la moyenne de ces scores pour les modèles individuels.

Figure 5-6
Noeud Numérisation automatique - Onglet Paramètres



Comparaison des modèles

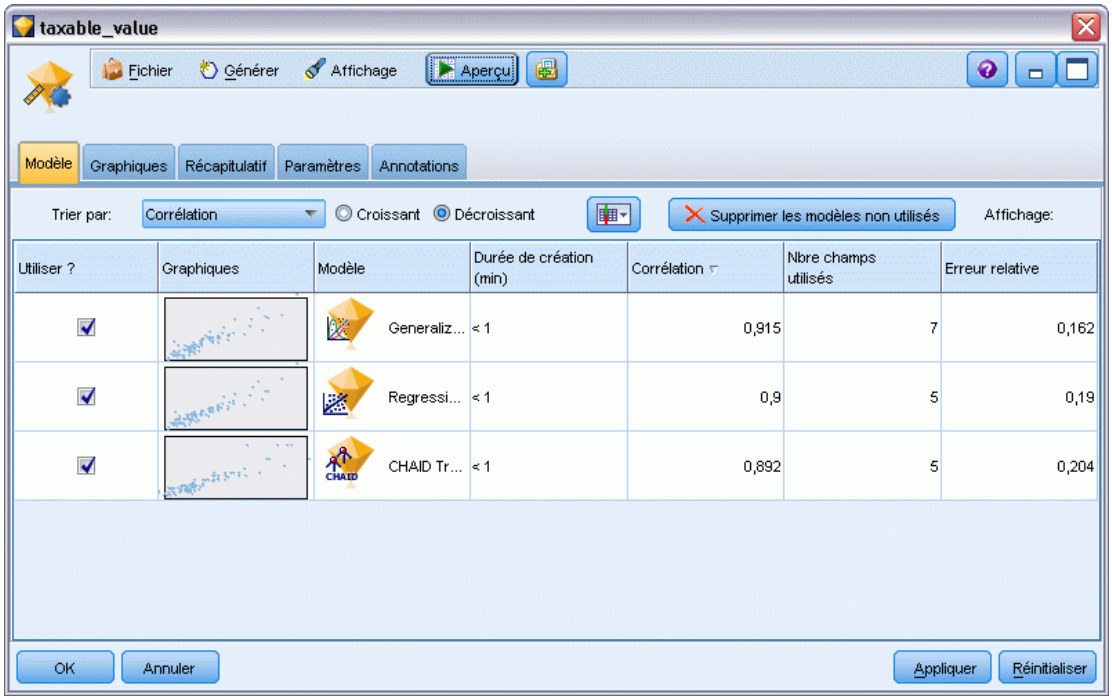
▶ Cliquez sur le bouton Exécuter.

Le nugget de modèle est créé et placé sur l'espace de travail et dans la palette Modèles en haut à droite de la fenêtre. Vous pouvez parcourir le nugget ou l'enregistrer ou le déployer de plusieurs façons.

Ouvrez le nugget de modèle ; il répertorie les détails concernant chacun des modèles créés au cours de l'exécution. (En situation réelle, lorsque des centaines de modèles sont estimés à partir d'un grand nombre de données, cette opération peut prendre plusieurs heures.) Consultez Figure 5-1 sur p. 57.

Si vous souhaitez analyser plus en détail l'un des modèles individuels, vous pouvez double-cliquer sur un nugget de modèles dans la colonne Modèle pour la faire défiler et parcourir les résultats du modèle individuel ; à partir de là, vous pouvez générer des noeuds de modélisation, des nuggets de modèle ou des graphiques d'évaluation.

Figure 5-7
Résultats de la numérisation automatique



Par défaut, les modèles sont classés en fonction de la corrélation, cette mesure ayant été sélectionnée dans le noeud Numérisation automatique. Pour faciliter le classement, la valeur absolue de la corrélation est utilisée, avec les valeurs les plus proches de 1 indiquant une relation très forte. Le modèle linéaire généralisé est classé comme étant le meilleur en fonction de cette mesure, mais plusieurs autres sont presque aussi précis. Ce modèle linéaire généralisé a également l'erreur relative la plus basse.

Vous pouvez effectuer le tri sur une autre colonne en cliquant sur l'en-tête de cette colonne ou vous pouvez choisir la mesure désirée dans la liste Trier par de la barre d'outils.

Chaque graphique présente un nuage de valeurs observées par rapport aux valeurs prédites pour le modèle et fournit ainsi une indication visuelle rapide de leurs corrélation. Pour un modèle performant, les points doivent être regroupés le long de la diagonale, ce qui est vrai pour tous les modèles de cet exemple.

Dans la colonne Graphique, vous pouvez double-cliquer sur une miniature pour générer un graphique en grandeur nature.

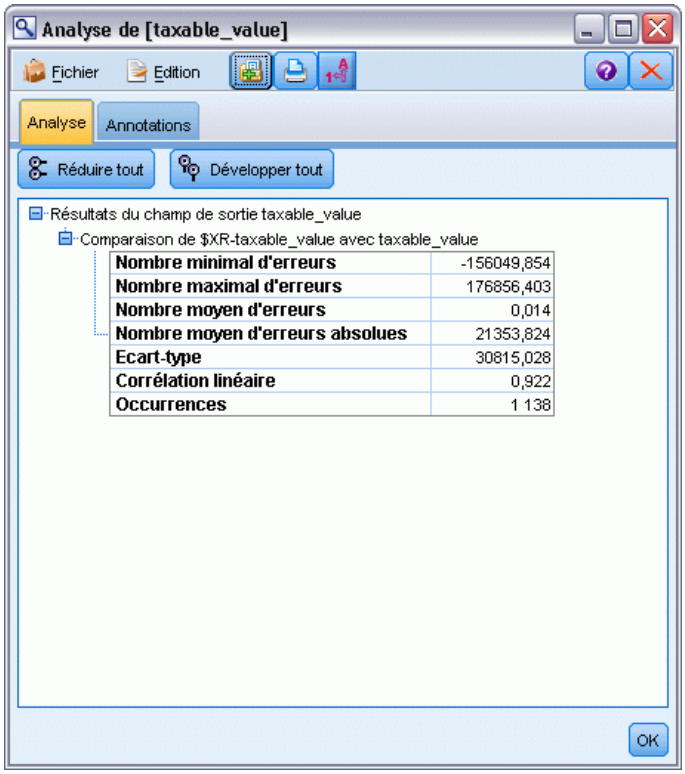
En fonction de ses résultats, vous pouvez décider d'utiliser les trois modèles les plus précis. En combinant les prévisions à partir de plusieurs modèles, il est possible d'éviter les limitations dans les modèles individuels. Ce qui entraîne une plus grande précision globale.

Dans la colonne Utiliser ?, vérifiez que les trois modèles sont sélectionnés.

Liez un noeud Analyse (palette Sortie) après le nugget de modèle. Cliquez avec le bouton droit de la souris sur le noeud Analyse et sélectionnez Exécuter pour exécuter le flux.

La moyenne du score généré par le modèle d'ensemble est ajoutée à un champ *$XR-taxable_value*, avec une corrélation de 0,922, ce qui est supérieur à ceux des trois modèles individuels. Les scores d'ensemble affichent également une faible erreur moyenne absolue et peuvent être plus efficaces que tous les modèles individuels lorsqu'ils sont appliqués à d'autres ensembles de données.

Figure 5-8
Exemple de flux Numérisation automatique



Récapitulatif

Pour résumer, vous avez utilisé le noeud Numérisation automatique pour comparer plusieurs modèles différents, vous avez sélectionné les trois modèles les plus précis et vous les avez ajoutés au flux dans un nugget de modèle Numérisation automatique combiné.

- Concernant la précision globale, les modèles linéaires généralisés, de Régression et CHAID sont plus performants avec les données d'apprentissage.
- Le modèle d'ensemble a été plus performant que deux des trois modèles individuels et peut être aussi efficace lorsqu'il est appliqué à d'autres ensembles de données. Si votre objectif est d'automatiser autant que possible le processus, cette approche vous permet d'obtenir un modèle fiable dans la plupart des circonstances sans avoir à creuser trop dans les spécificités des modèles.

Partie II:
Exemples de préparation des données

Chapitre 6

Préparation automatique de données (ADP)

La préparation des données pour l'analyse est l'une des étapes les plus importantes dans tout projet de Data mining et généralement l'une des plus longues. Le noeud de préparation automatique de données (ADP) gère cette tâche pour vous en analysant vos données et en identifiant des corrections, en filtrant des champs problématiques et peu susceptibles d'être utiles et en créant de nouveaux attributs le cas échéant, et enfin en améliorant la performance au moyen de techniques de filtrage intelligentes. Vous pouvez utiliser le noeud de manière totalement automatisée, en laissant le noeud choisir et appliquer les corrections, ou vous pouvez prévisualiser les modifications avant qu'elles ne soient effectuées et les accepter ou les refuser au choix.

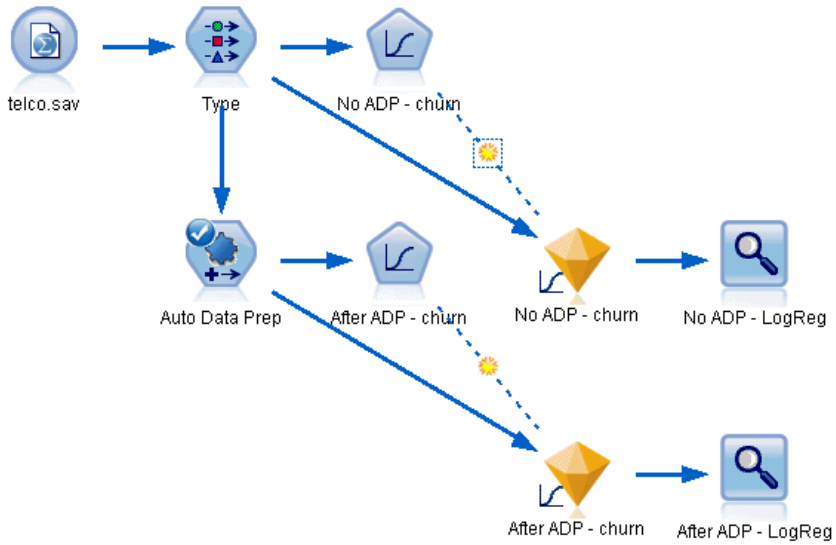
Le noeud ADP vous permet de préparer rapidement et facilement les données pour le Data Mining sans connaissance préalable des concepts statistiques impliqués. Si vous exécutez le noeud avec les paramètres par défaut, les modèles auront tendance à être créé et à réaliser des évaluations plus rapidement.

Cet exemple utilise le flux nommé *ADP_basic_demo.str*, qui se rapporte à un fichier de données nommé *telco.sav* pour expliquer la précision accrue dont vous pouvez bénéficier en utilisant les paramètres par défaut du noeud ADP lors de la création de modèles. Ces fichiers sont disponibles dans le répertoire *Demos* de n'importe quelle installation de IBM® SPSS® Modeler. Ce dossier est accessible à partir du groupe de programmes IBM® SPSS® Modeler dans le menu Démarrer de Windows. Le fichier *ADP_basic_demo.str* se trouve dans le répertoire des *flux*.

Création du flux

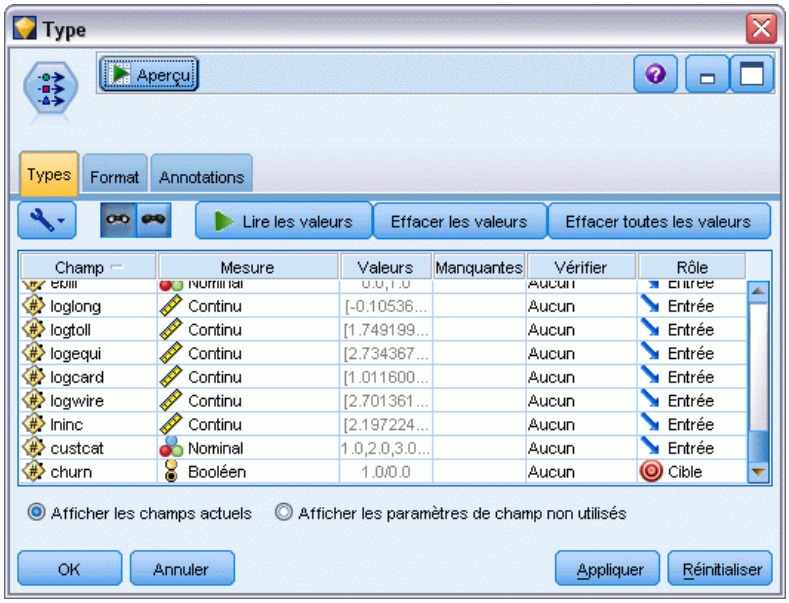
- Pour créer le flux, ajoutez un noeud source Statistics qui pointe sur *telco.sav*, dans le répertoire *Demos* du dossier d'installation de IBM® SPSS® Modeler.

Figure 6-1
Création du flux



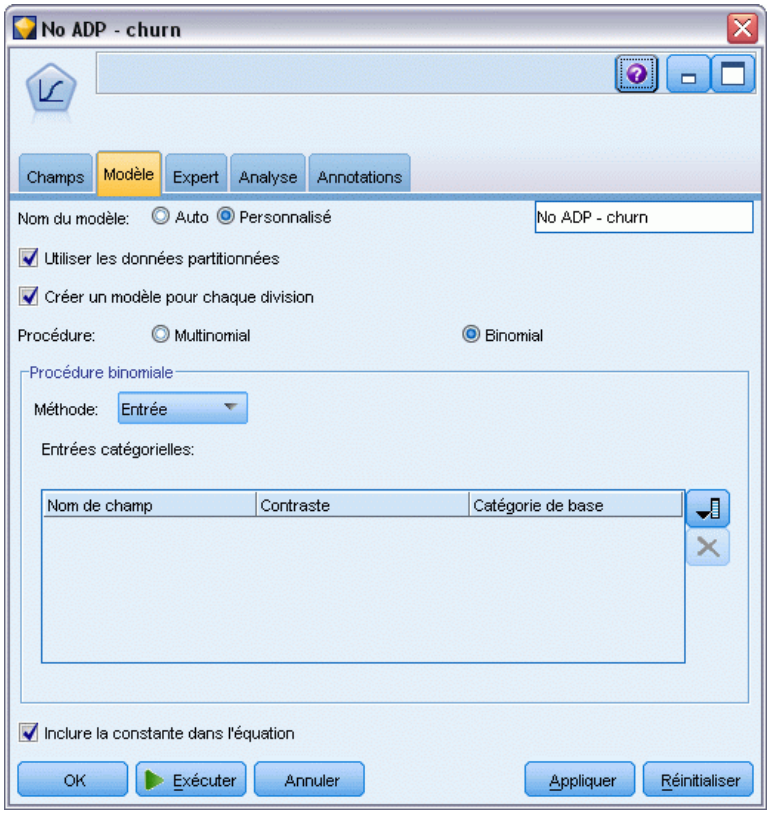
- Attachez un noeud Typer au noeud source, définissez le niveau de mesure du champ *attrition* sur Booléen et le rôle sur Cible. Le rôle de tous les autres champs doit être défini sur Entrée.

Figure 6-2
Sélection de la cible



- Reliez un noeud Logistique au noeud Typer.
- Dans le noeud Logistique, cliquez sur l'onglet Modèle et sélectionnez la procédure Binomial. Dans le champ *Nom de modèle*, sélectionnez Personnalisé et saisissez Pas de ADP - attrition.

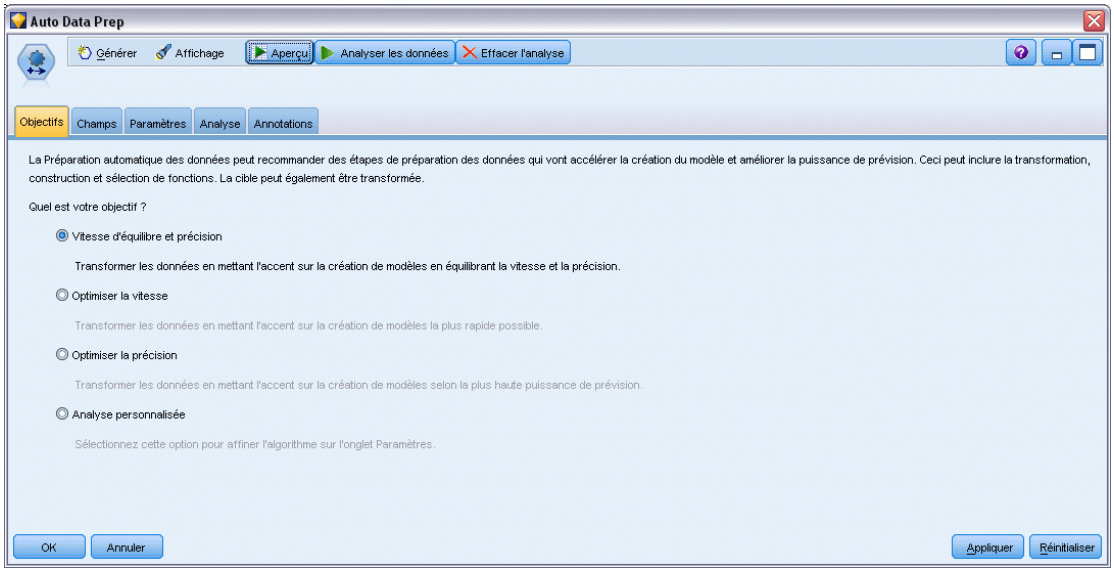
Figure 6-3
Choix des options de modèle



- Reliez un noeud ADP au noeud Typer. Dans l'onglet Objectifs, conservez les paramètres par défaut afin d'analyser et de préparer vos données en équilibrant la vitesse et la précision.
- En haut de l'onglet Objectifs, cliquez sur Analyser les données afin d'analyser et de traiter vos données.

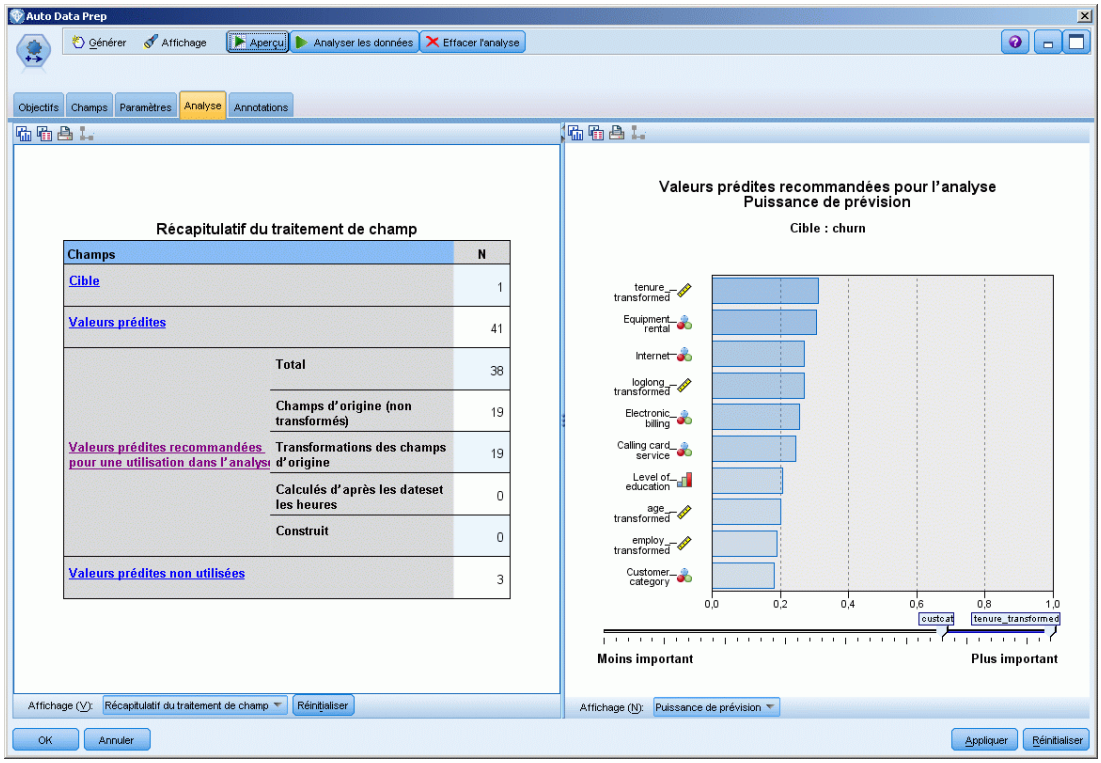
D’autres options du noeud ADP vous permettent de spécifier si vous souhaitez vous concentrer davantage sur la précision, sur la vitesse de traitement ou affiner les nombreuses étapes de traitement de la préparation des données.

Figure 6-4
Objectifs ADP par défaut



Les résultats du traitement des données sont affichés dans l'onglet Analyse. Le Récapitulatif de traitement des champs montre que parmi les 41 éléments de données que propose le noeud ADP, 19 ont été transformés afin d'améliorer le traitement et 3 ont été abandonnés car ils ne sont pas utilisés.

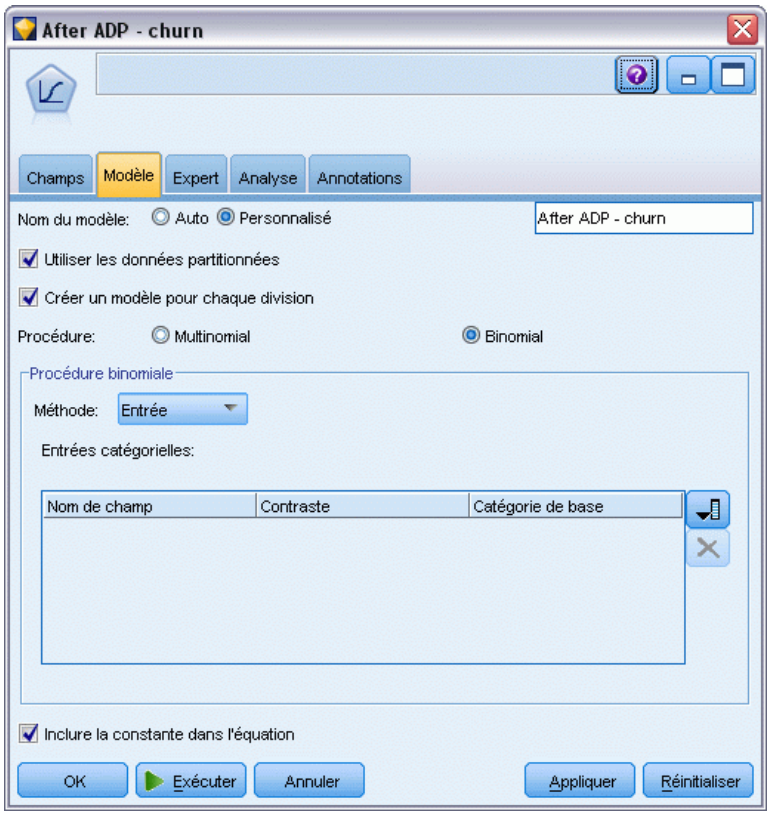
Figure 6-5
Récapitulatif du traitement des données



- Reliez un noeud Logistique au noeud ADP.

- Dans le noeud Logistique, cliquez sur l'onglet Modèle et sélectionnez la procédure Binomial. Dans le champ *Nom de modélisation*, sélectionnez Personnalisé et saisissez Après ADP - attrition.

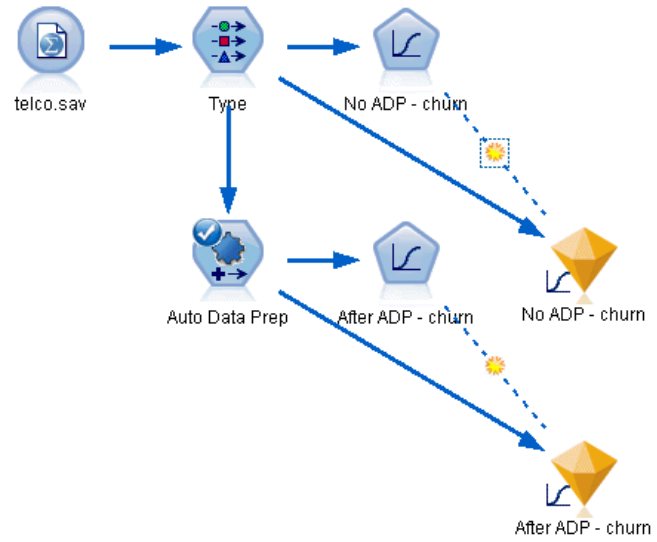
Figure 6-6
Choix des options de modèle



Comparaison de la précision des modèles

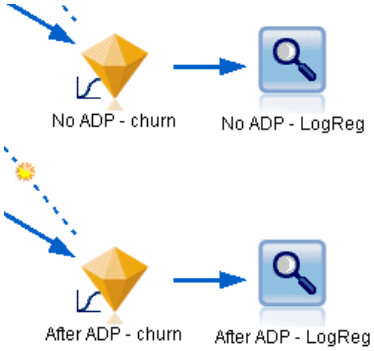
- Exécutez les deux noeuds Logistique pour créer les nuggets de modèle, qui sont ajoutés au flux et à la palette Modèles dans l'angle supérieur droit.

Figure 6-7
Relier les nuggets de modèle



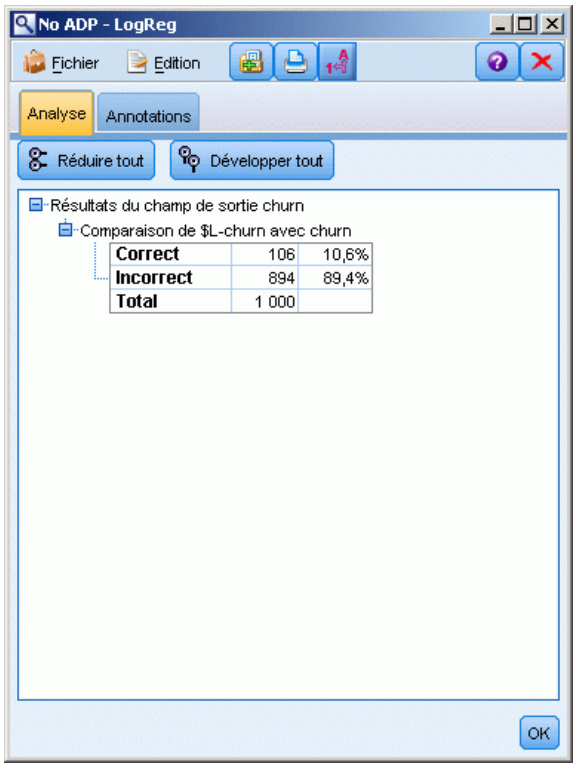
- Reliez les noeuds Analyse aux nuggets de modèle et exécutez des noeuds Analyse avec leurs paramètres par défaut.

Figure 6-8
Relier les noeuds d'analyse



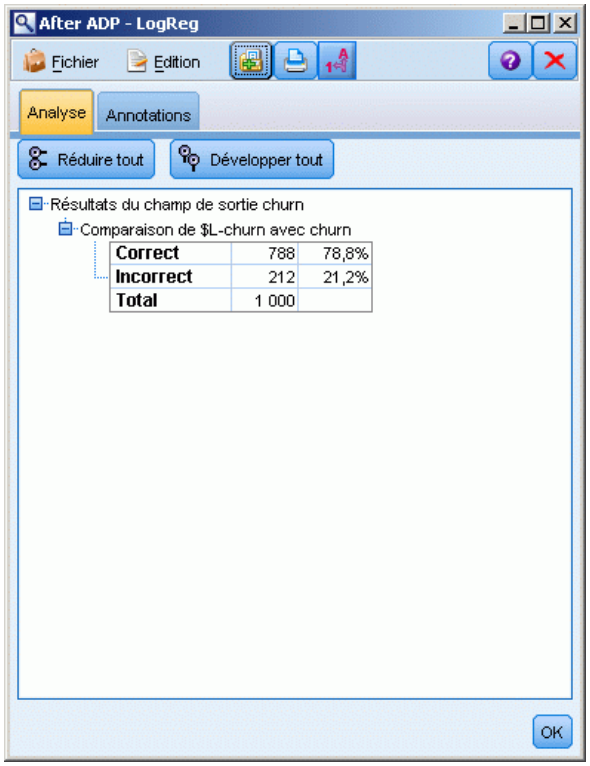
L’analyse du modèle non dérivé ADP montre que la seule exécution des données dans le noeud Régression logistique avec ces paramètres par défaut fournit un modèle de faible précision - seulement 10,6 %.

Figure 6-9
Résultat d’un modèle non dérivé de l’ADP



L'analyse du modèle dérivé de l'ADP montre que dans le cadre de l'exécution des données avec les paramètres ADP par défaut, vous avez construit un modèle beaucoup plus précis, exact à 78.8%.

Figure 6-10
Résultat d'un modèle dérivé de l'ADP



Dans le récapitulatif, en exécutant uniquement le noeud ADP pour affiner le traitement de vos données, vous avez été en mesure de construire un modèle plus précis avec peu de manipulation directe des données.

Bien sûr, si votre objectif est de prouver ou non la validité d'une certaine théorie, ou si vous souhaitez construire des modèles spécifiques, il peut être préférable d'utiliser directement les paramètres de modèle. Cependant, pour les personnes disposant de peu de temps ou si vous avez de grandes quantités de données à préparer, le noeud ADP peut représenter un avantage.

Des explications sur les fondements mathématiques des méthodes de modélisation utilisées dans IBM® SPSS® Modeler sont présentées dans le *Guide des algorithmes de SPSS Modeler*, disponible dans le répertoire *\Documentation* du disque d'installation.

Sachez également que les résultats de cet exemple sont établis uniquement d'après les données d'apprentissage. Pour évaluer comment les modèles peuvent se généraliser à d'autres données dans le monde réel, vous devez utiliser un noeud Partitionner destiné à contenir un sous-ensemble d'enregistrements à des fins de test et de validation. Pour plus d'informations, reportez-vous à la section Noeud Partitionner dans le chapitre 4 dans *Noeuds source, exécution et de sortie de IBM SPSS Modeler 15*.

Chapitre
7

Préparation des données pour l'analyse (Audit données)

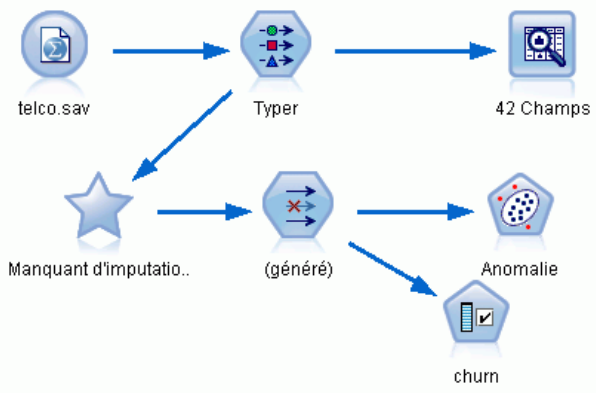
Le noeud Audit données fournit un premier aperçu complet des données importées dans IBM® SPSS® Modeler. Souvent utilisé lors de l'exploration initiale des données, le rapport d'audit des données affiche des statistiques récapitulatives, ainsi que les histogrammes et les graphiques Proportion pour chaque champ de données. Il vous permet en outre d'indiquer comment traiter les valeurs manquantes, les valeurs éloignées et les valeurs extrêmes.

Cet exemple utilise le flux *telco_dataaudit.str*, qui fait référence au fichier de données *telco.sav*. Ces fichiers sont disponibles dans le répertoire *Demos* de n'importe quelle installation de IBM® SPSS® Modeler. Ce dossier est accessible à partir du groupe de programmes SPSS Modeler dans le menu Démarrer de Windows. Le fichier *telco_dataaudit.str* se trouve dans le répertoire des *flux* .

Création du flux

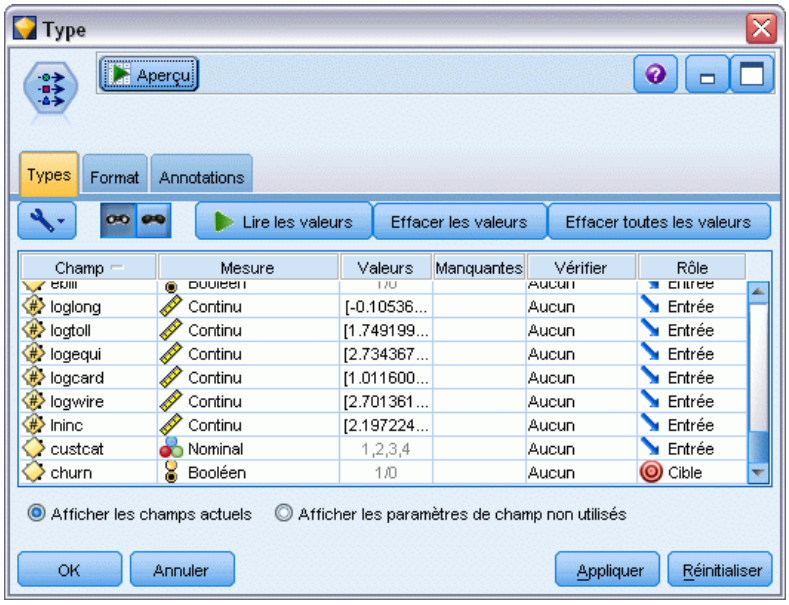
- Pour créer le flux, ajoutez un noeud source Statistics qui pointe sur *telco.sav*, dans le répertoire *Demos* du dossier d'installation de IBM® SPSS® Modeler.

Figure 7-1
Création du flux



- Ajoutez un noeud Typer pour définir des champs, puis désignez *attrition* comme champ cible (Rôle = Cible). Le rôle doit avoir la valeur Entrée pour tous les autres champs pour que cette cible soit la seule cible.

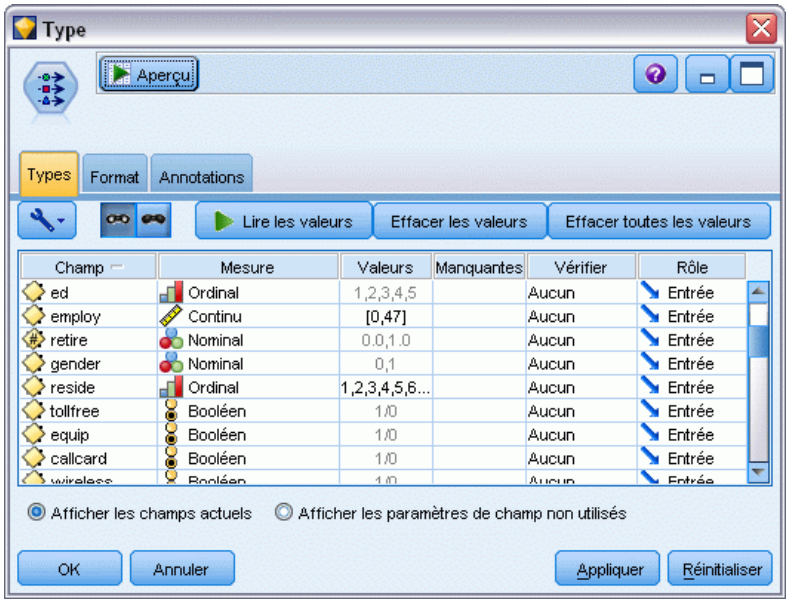
Figure 7-2
Définition de la cible



- Vérifiez que les niveaux de mesure de champ sont correctement définis. Par exemple, la plupart des champs dont les valeurs sont 0 et 1 peuvent être considérés comme des champs booléens.

Cependant, certains champs, tels que celui indiquant le genre, doivent être considérés comme des champs nominaux à deux valeurs.

Figure 7-3
Définition des niveaux de mesure

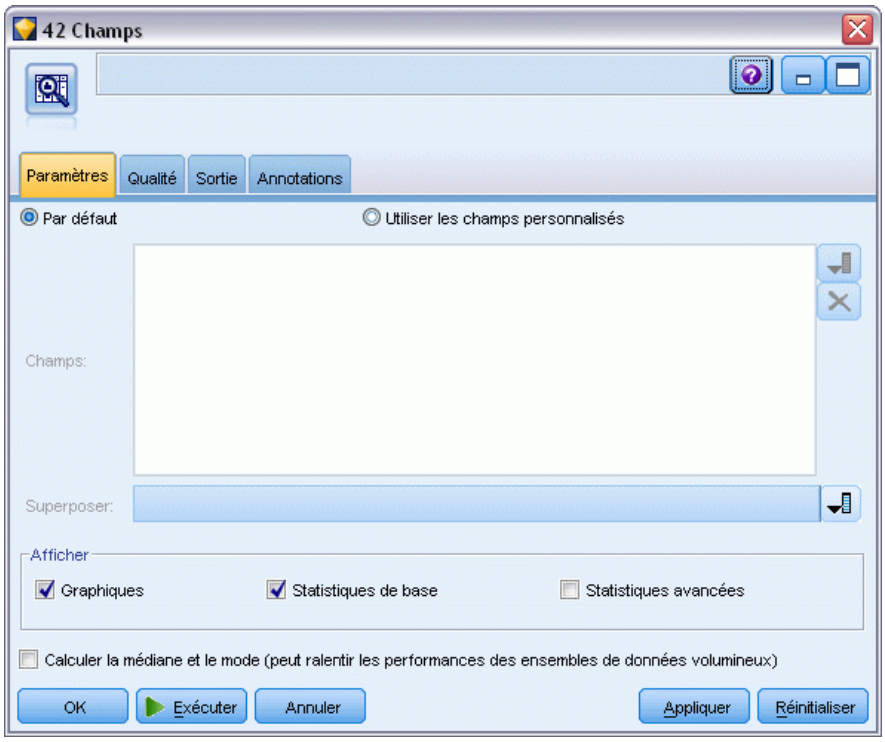


Astuce : Pour modifier les propriétés de plusieurs champs contenant des valeurs similaires (telles que 0/1), cliquez sur l'en-tête de colonne *Valeurs* afin de trier les champs en fonction de cette colonne. Utilisez la touche Maj pour sélectionner tous les champs à modifier. Cliquez ensuite sur la sélection avec le bouton droit de la souris pour modifier le niveau de mesure ou les autres attributs de tous les champs sélectionnés.

- Connectez un noeud Audit données au flux. Dans l'onglet Paramètres, conservez les paramètres par défaut pour que tous les champs soient inclus dans le rapport. Etant donné que *attrition* est

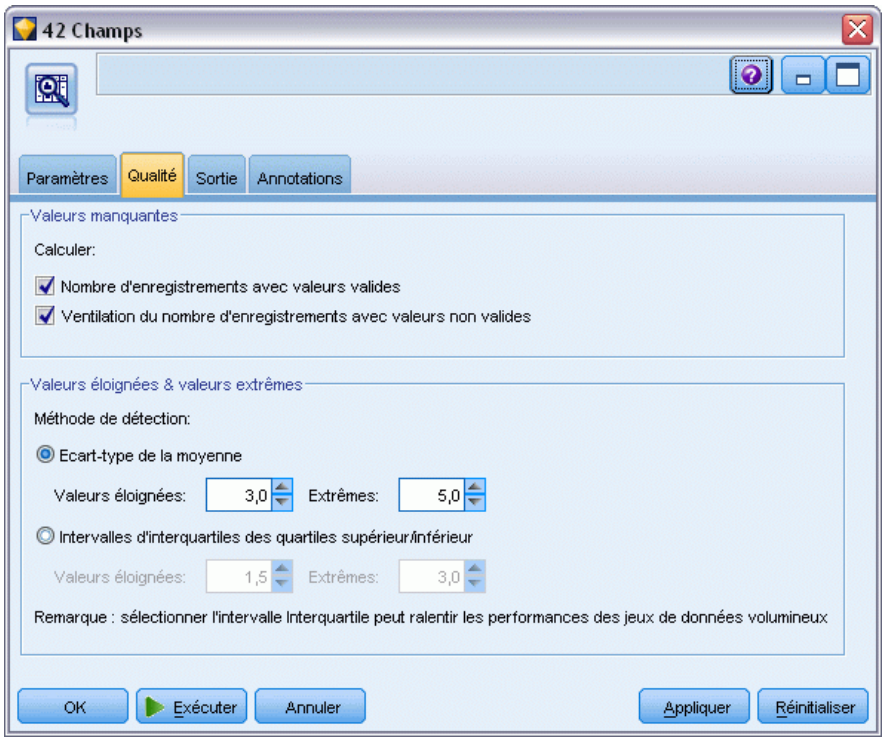
le seul champ cible défini dans le noeud Typer, ce champ est automatiquement utilisé comme champ de superposition.

Figure 7-4
Noeud Audit données - Onglet Paramètres



Dans l'onglet Qualité, conservez les paramètres par défaut de détection des valeurs manquantes, éloignées et extrêmes, puis cliquez sur Exécuter.

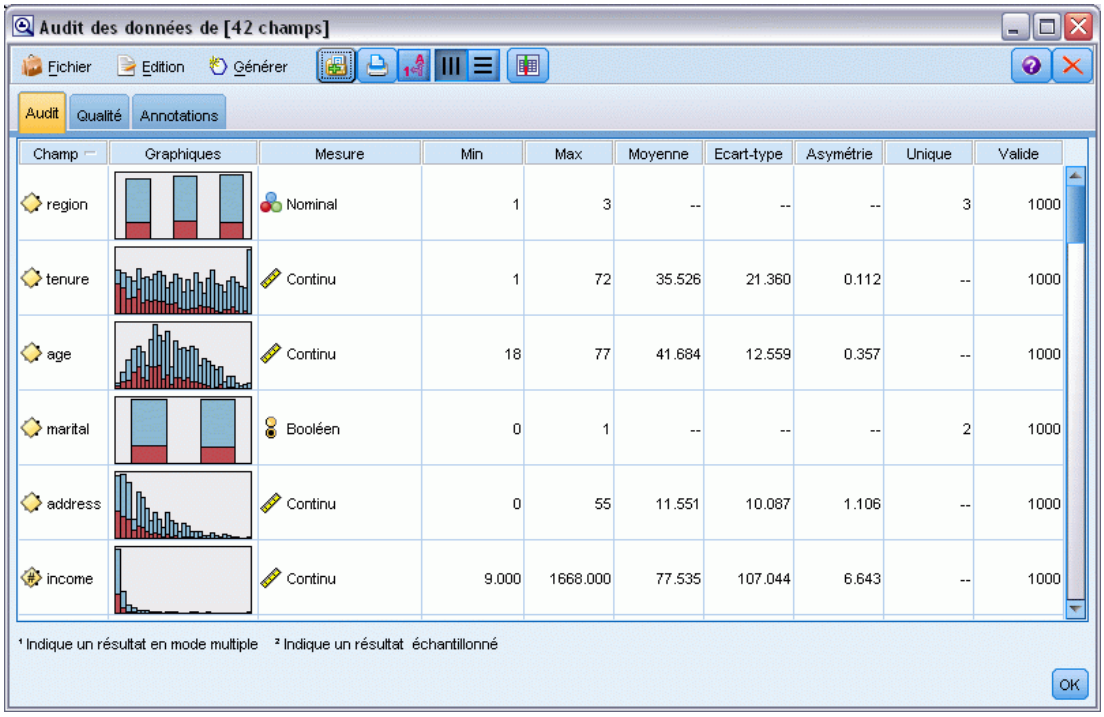
Figure 7-5
Noeud Audit données - Onglet Qualité



Navigation dans les statistiques et les graphiques

Le navigateur Audit données est affiché avec des graphiques en miniature et des statistiques descriptives pour chaque champ.

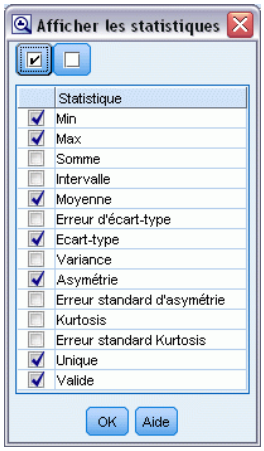
Figure 7-6
Navigateur Audit données



A l'aide de la barre d'outils, affichez les étiquettes de champ et de valeur, et basculez l'alignement des graphiques de l'horizontale à la verticale (champs catégoriels uniquement).

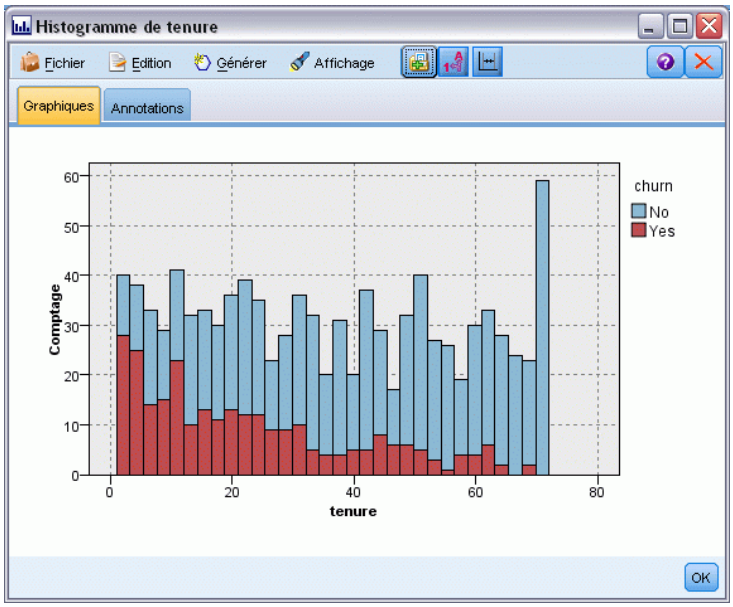
- La barre d'outils ou le menu Editer vous permet en outre de choisir les statistiques à afficher.

Figure 7-7
Afficher les statistiques



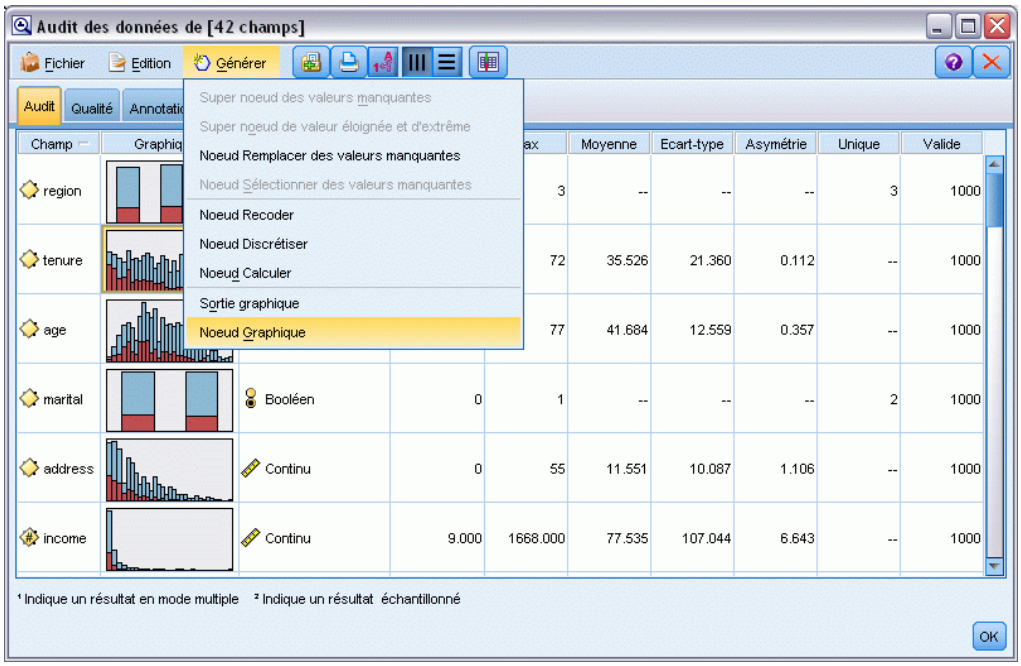
Double-cliquez sur un graphique en miniature dans le rapport d'audit pour afficher ce graphique en taille réelle. Etant donné que *attrition* est le seul champ cible du flux, il est automatiquement utilisé comme champ de superposition. Vous pouvez basculer l'affichage des étiquettes de champ et de valeur à l'aide de la barre d'outils de la fenêtre Graphiques ou cliquer sur le bouton Mode d'édition pour personnaliser le graphique.

Figure 7-8
Histogramme de durée d'affectation



Vous pouvez également sélectionner une ou plusieurs miniatures et générer un noeud Graphiques pour chacune d'elles. Les noeuds générés sont placés dans l'espace de travail de flux. Vous pouvez les ajouter au flux pour recréer le graphique concerné.

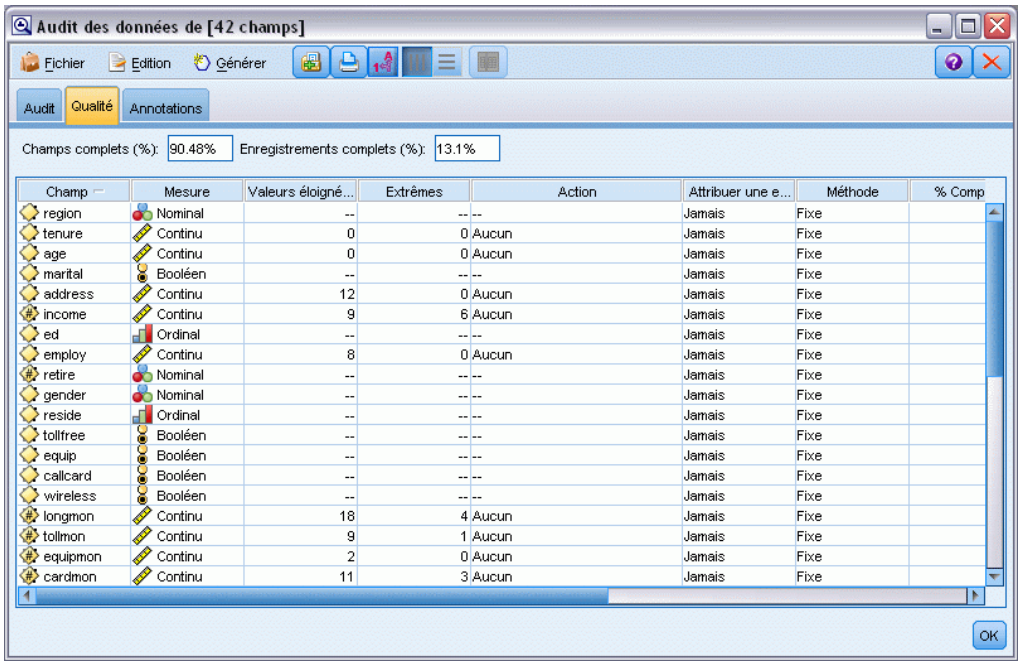
Figure 7-9
Génération d'un noeud Graphiques



Traitement des valeurs éloignées et manquantes

L'onglet Qualité du rapport d'audit contient des informations sur les valeurs éloignées, extrêmes et manquantes.

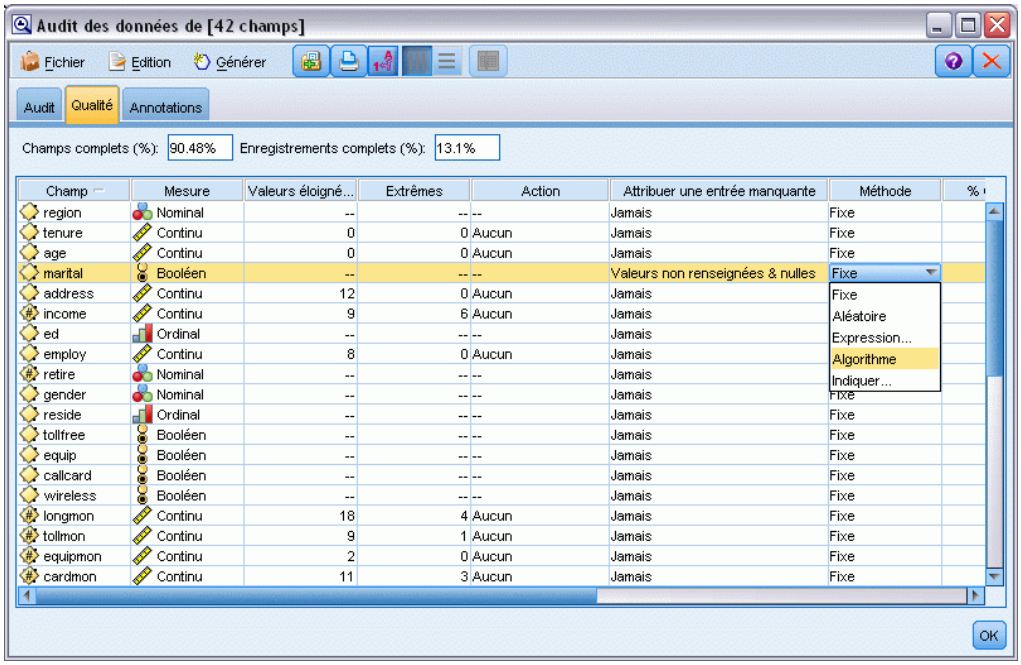
Figure 7-10
Navigateur Audit données - Onglet Qualité



Champ	Mesure	Valeurs éloigné...	Extrêmes	Action	Attribuer une e...	Méthode	% Comp
region	Nominal	--	--	--	Jamais	Fixe	
tenure	Continu	0	0	Aucun	Jamais	Fixe	
age	Continu	0	0	Aucun	Jamais	Fixe	
marital	Booléen	--	--	--	Jamais	Fixe	
address	Continu	12	0	Aucun	Jamais	Fixe	
income	Continu	9	6	Aucun	Jamais	Fixe	
ed	Ordinal	--	--	--	Jamais	Fixe	
employ	Continu	8	0	Aucun	Jamais	Fixe	
retire	Nominal	--	--	--	Jamais	Fixe	
gender	Nominal	--	--	--	Jamais	Fixe	
reside	Ordinal	--	--	--	Jamais	Fixe	
tollfree	Booléen	--	--	--	Jamais	Fixe	
equip	Booléen	--	--	--	Jamais	Fixe	
callcard	Booléen	--	--	--	Jamais	Fixe	
wireless	Booléen	--	--	--	Jamais	Fixe	
longmon	Continu	18	4	Aucun	Jamais	Fixe	
tollmon	Continu	9	1	Aucun	Jamais	Fixe	
equipmon	Continu	2	0	Aucun	Jamais	Fixe	
cardmon	Continu	11	3	Aucun	Jamais	Fixe	

Vous pouvez également définir des méthodes de gestion des valeurs et générer des super noeuds qui appliquent automatiquement les transformations. Par exemple, vous pouvez sélectionner un ou plusieurs champs et choisir d'attribuer ou de remplacer les valeurs manquantes de ces champs à l'aide de diverses méthodes, dont l'algorithme C&RT.

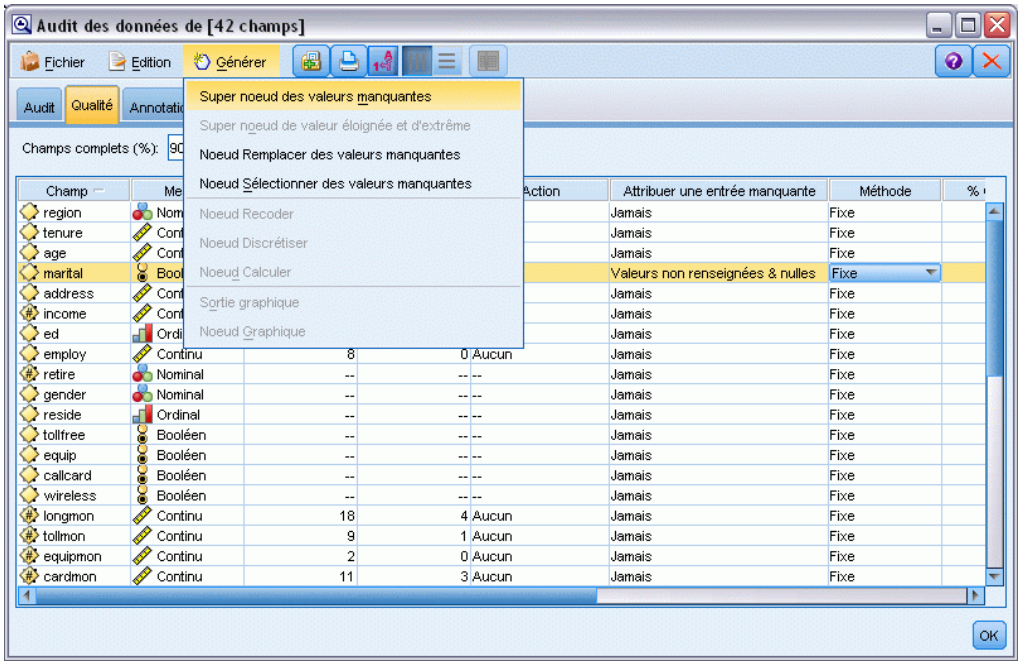
Figure 7-11
Choix d'une méthode d'attribution



Après avoir spécifié une méthode d'attribution pour un ou plusieurs champs, pour générer un super noeud Valeurs Manquantes, dans les menus choisissez :

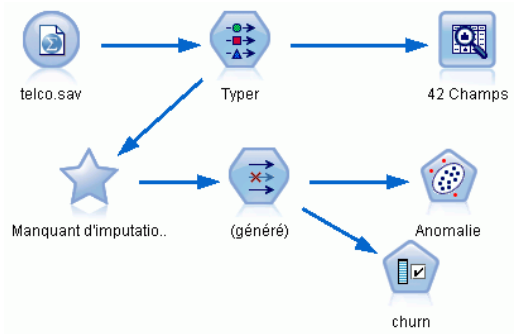
Générer > Super noeud des valeurs manquantes

Figure 7-12
Génération du Super noeud



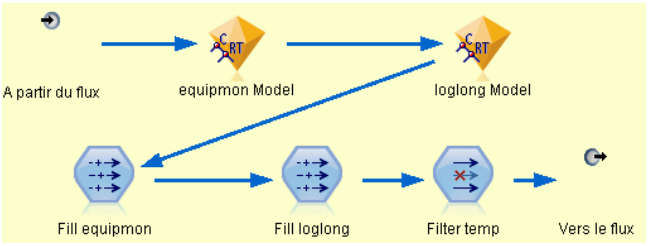
Le super noeud généré est ajouté à l'espace de travail de flux, où vous pouvez le connecter au flux pour appliquer les transformations.

Figure 7-13
Flux avec super noeud Valeurs manquantes



Le super noeud contient en réalité plusieurs noeuds qui exécutent les transformations requises. Pour comprendre son fonctionnement, modifiez le super noeud et cliquez sur Zoom avant.

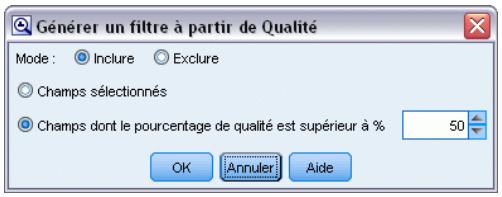
Figure 7-14
Zoom avant sur le super noeud



Chaque champ auquel une valeur est attribuée à l'aide de la méthode algorithmique, par exemple, est associé à un modèle C&RT distinct et à un noeud Remplacer qui remplace les valeurs non renseignées et les valeurs nulles par la valeur prédite par le modèle. Vous pouvez ajouter, modifier ou supprimer des noeuds précis dans le super noeud pour personnaliser encore davantage son comportement.

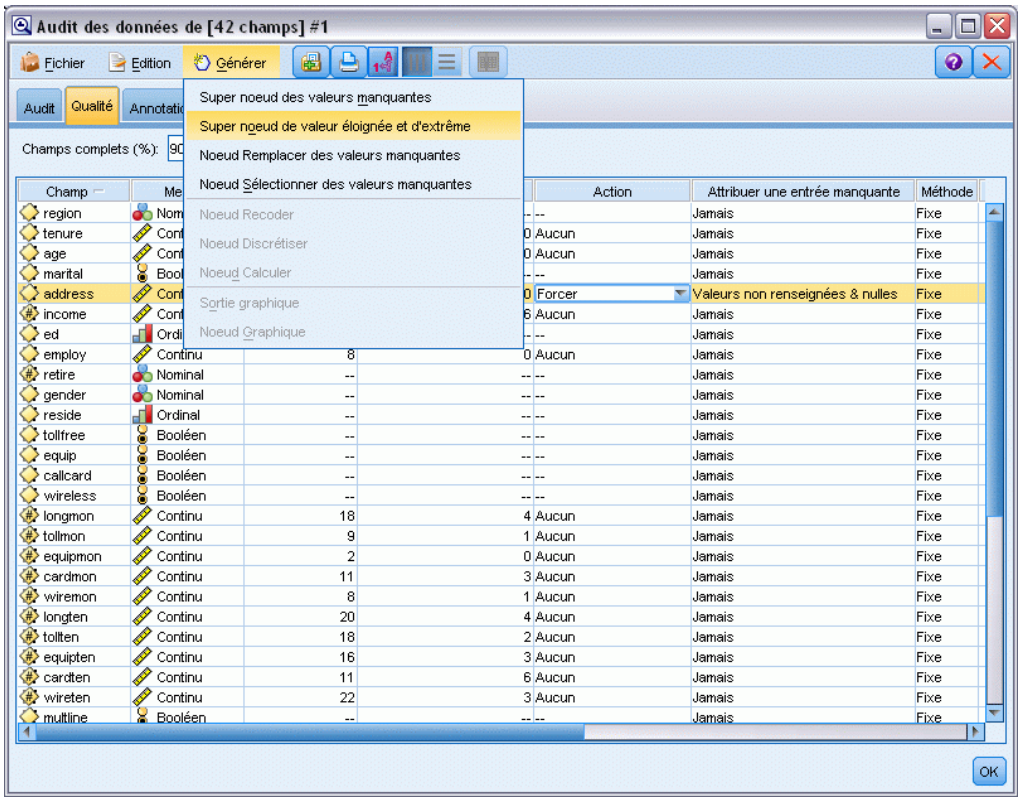
Vous pouvez également générer un noeud Sélectionner ou Filtrer pour supprimer les champs ou les enregistrements où des valeurs manquent. Par exemple, vous pouvez filtrer les champs dont le pourcentage de qualité est inférieur au seuil défini.

Figure 7-15
Génération d'un noeud Filtrer



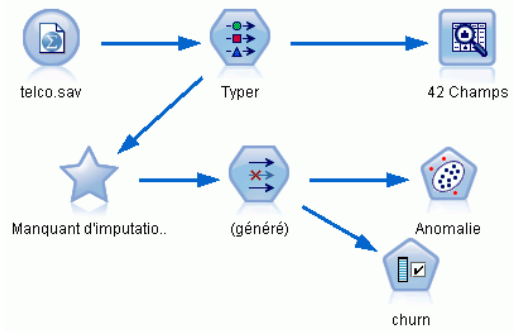
Les valeurs éloignées et extrêmes peuvent être gérées de manière similaire. Indiquez l'action à appliquer à chaque champ (Forcer, Supprimer ou Rendre nul) et générez un super noeud pour appliquer les transformations.

Figure 7-16
Génération d'un noeud Filtrer



Une fois l'audit terminé et les noeuds générés ajoutés au flux, vous pouvez poursuivre l'analyse. Vous pouvez également effectuer une analyse plus poussée des données grâce à la méthode Détection des anomalies ou Sélection de fonction, ou à d'autres méthodes.

Figure 7-17
Flux avec super noeud Valeurs manquantes



Chapitre
8

Traitements par médicaments (Graphiques exploratoires/C5.0)

Pour cette section, imaginez que vous êtes un chercheur et que vous souhaitez compiler des données pour une étude médicale. Vous avez rassemblé des données sur un ensemble de patients, souffrant tous de la même maladie. Lors du traitement, chaque patient a réagi à l'un des cinq médicaments. Votre travail consiste à utiliser le Data mining pour savoir quel médicament pourrait convenir à un futur patient atteint de la même maladie.

Cet exemple utilise le flux intitulé *druglearn.str*, qui référence le fichier de données *DRUG1n*. Ces fichiers sont disponibles dans le répertoire *Demos* de n'importe quelle installation de IBM® SPSS® Modeler. Ce dossier est accessible à partir du groupe de programmes IBM® SPSS® Modeler dans le menu Démarrer de Windows. Le fichier *druglearn.str* se trouve dans le répertoire des *flux*.

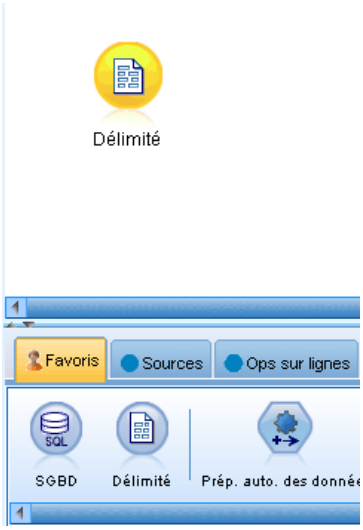
Les champs de données utilisés dans la démo sont :

Champ de données	Description
Age	(chiffre)
Sexe	*M* ou *F*
TA	Tension artérielle : *ELEVEE*, *NORMALE*, ou *BASSE*
Cholestérol	Taux de cholestérol dans le sang : *NORMAL* ou *ELEVE*
Na	Concentration de sodium dans le sang
K	Concentration de potassium dans le sang
Médicament	Médicament prescrit auquel le patient a réagi

Lecture de données texte

Vous pouvez lire des données texte délimitées à l'aide d'un **noeud Délimité**. Vous pouvez ajouter un noeud Délimité depuis les palettes en cliquant sur l'onglet Sources pour rechercher le noeud ou utiliser l'onglet Favoris qui contient ce noeud par défaut. Ensuite, double-cliquez sur le noeud que vous venez de placer pour ouvrir la boîte de dialogue correspondante.

Figure 8-1
Ajout d'un noeud Délimité



Pour sélectionner le répertoire dans lequel IBM® SPSS® Modeler est installé sur votre système, cliquez sur le bouton représentant des points de suspension (...), à droite de la zone Fichier. Ouvrez le répertoire *Demos*, puis sélectionnez le fichier *DRUG1n*.

Vérifiez que vous avez sélectionné Lire noms des champs à partir du fichier et notez les valeurs et champs qui ont été chargés dans la boîte de dialogue.

Figure 8-2
Boîte de dialogue Délimité

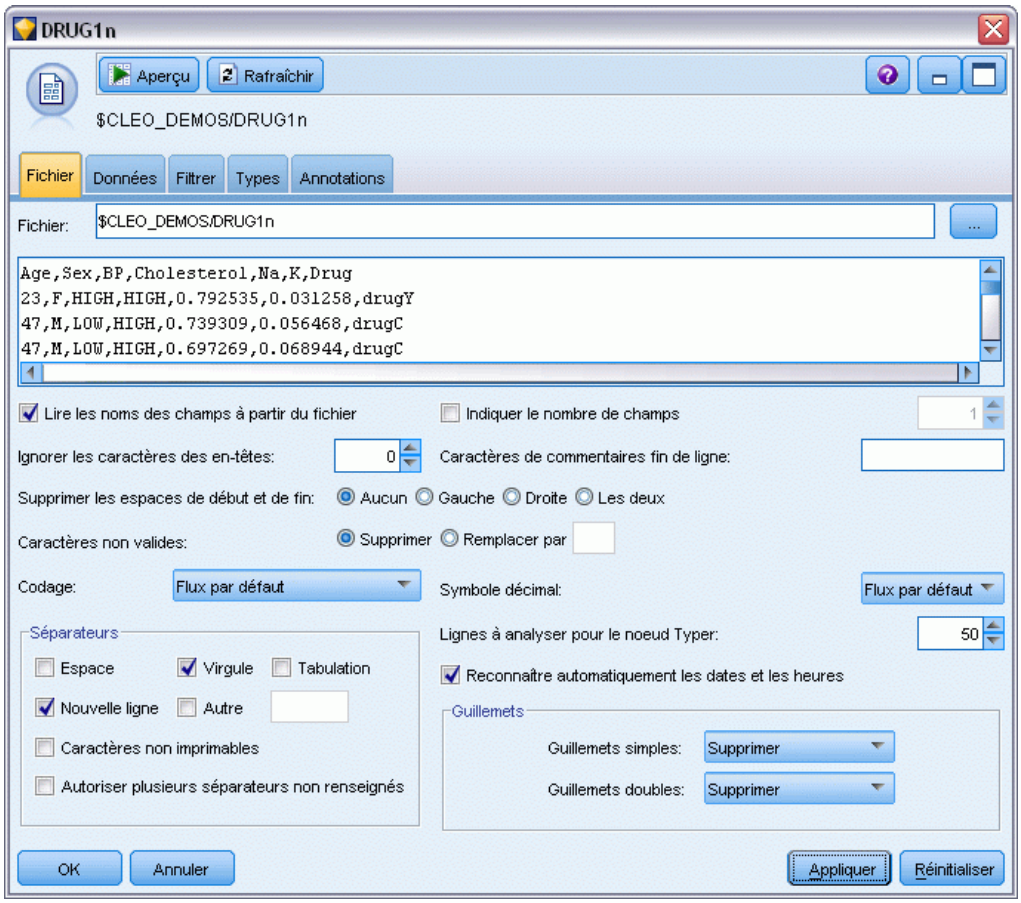


Figure 8-3
Modification du type de stockage pour un champ

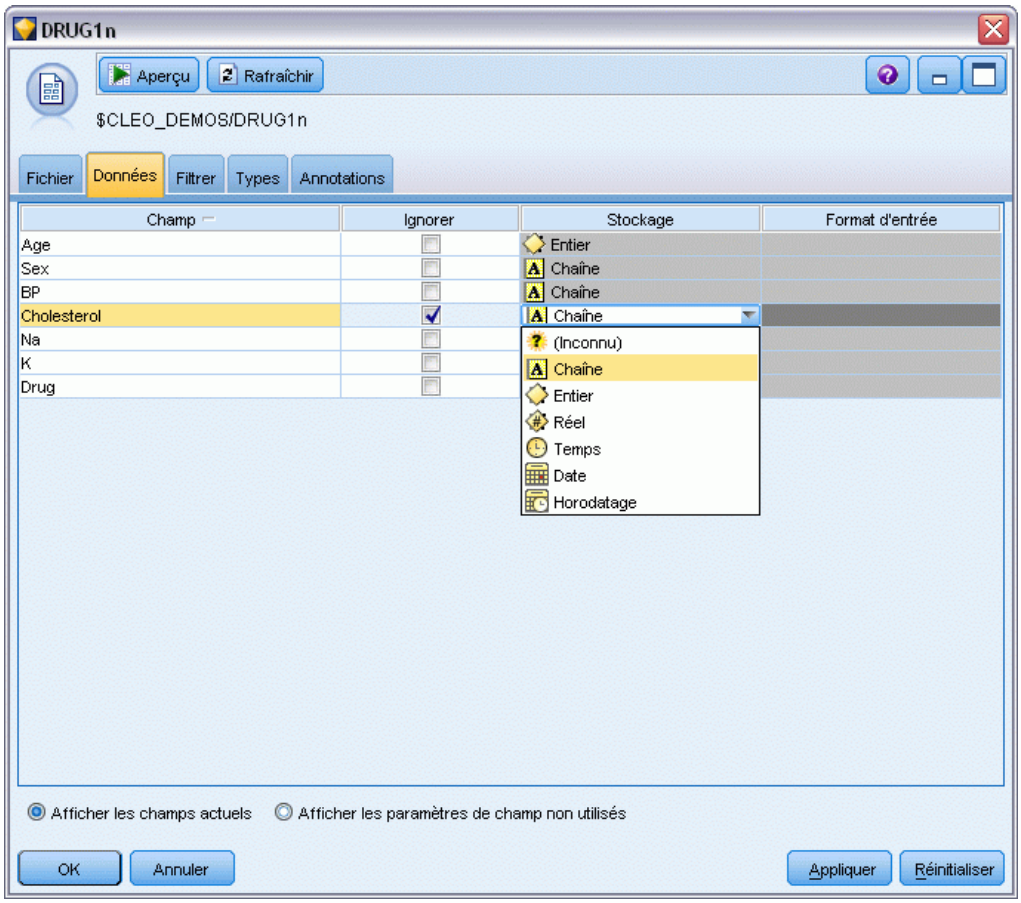
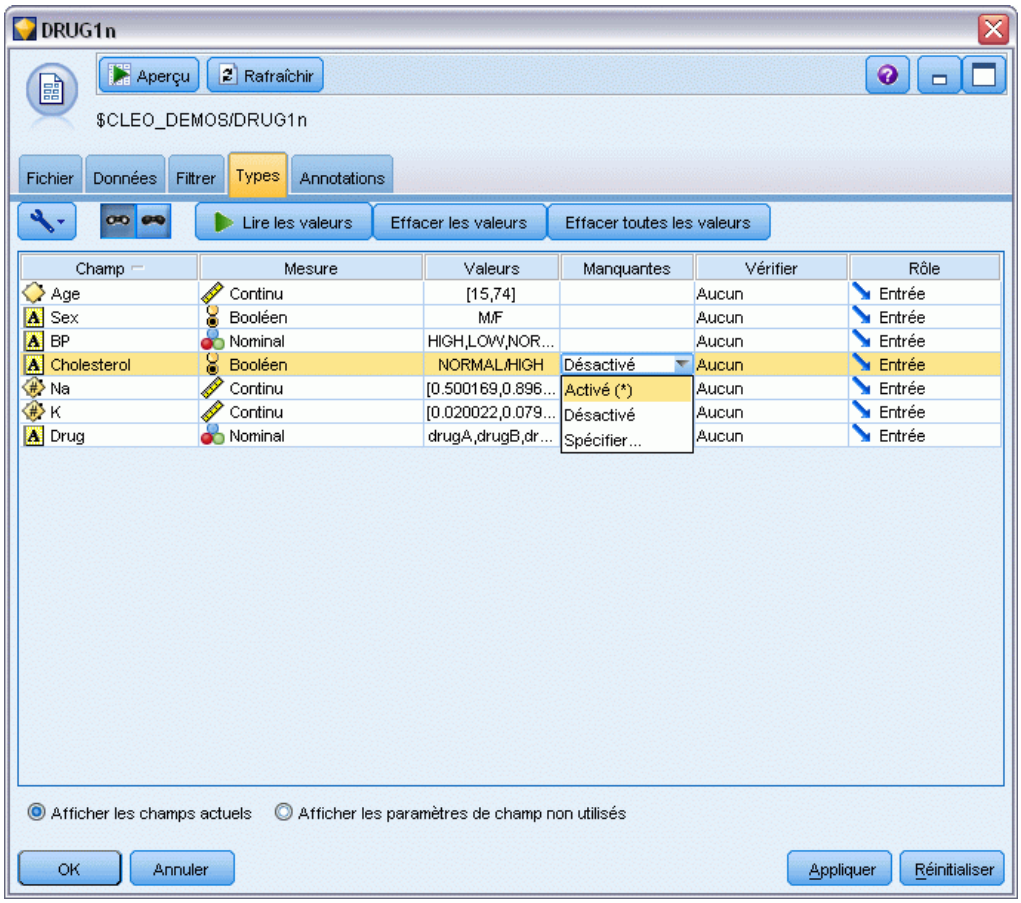


Figure 8-4
Sélection des options de valeur dans l'onglet Types



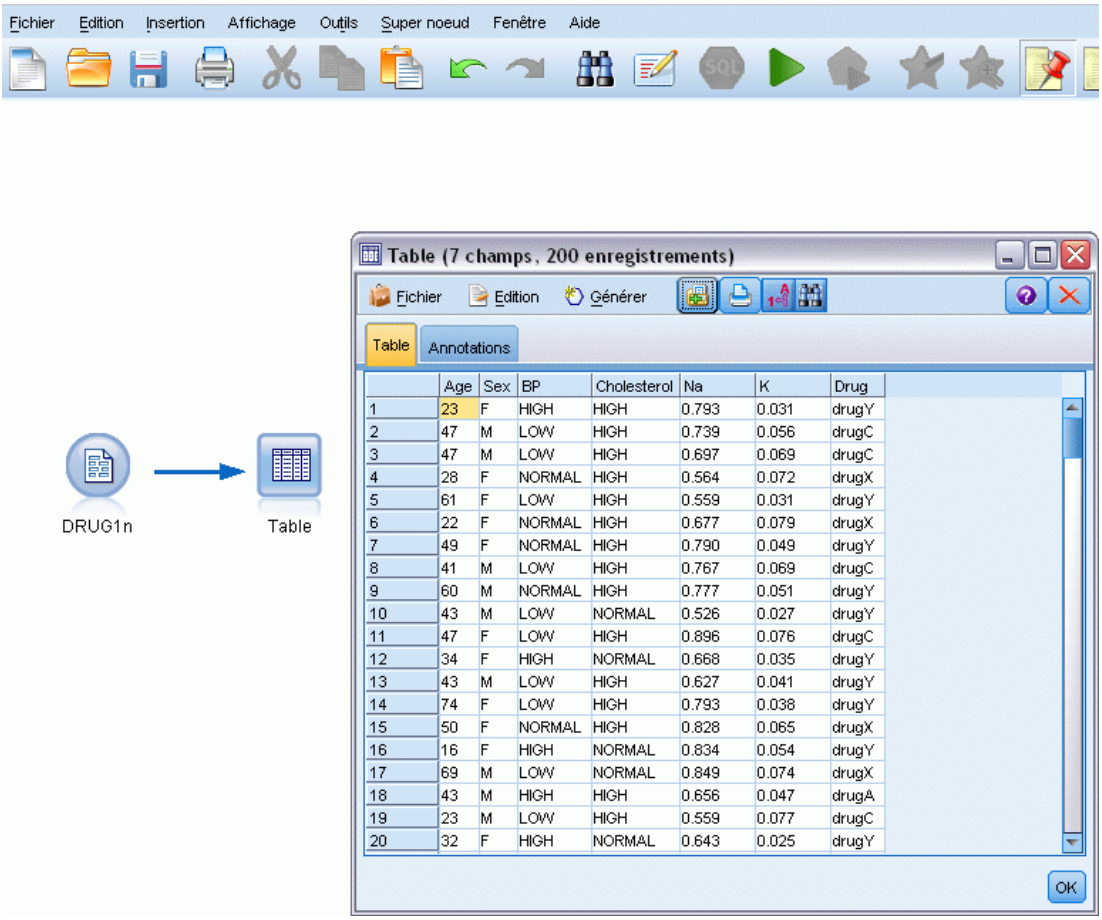
Cliquez sur l'onglet Données pour ignorer et modifier le **Stockage** d'un champ. Veuillez noter que le stockage est différent des **Mesures**, c'est-à-dire, le niveau de mesure (ou le type d'utilisation) du champ des données. L'onglet Types vous permet d'obtenir des informations supplémentaires sur le type de champs de vos données. Vous pouvez également choisir Lire les valeurs pour afficher les valeurs réelles de chaque champ en fonction des sélections effectuées dans la colonne *Valeurs*. Ce processus est appelé **instanciation**.

Ajout d'une table

Maintenant que le fichier de données est chargé, vous pouvez examiner les valeurs de certains enregistrements. Pour ce faire, vous pouvez, par exemple, créer un flux incluant un noeud Table. Pour placer un noeud Table dans le flux, double-cliquez sur l'icône de la palette ou faites-la glisser vers l'espace de travail.

Figure 8-5
Noeud Table relié à la source de données

Figure 8-6
Exécution d'un flux à partir de la barre d'outils



lorsque vous double-cliquez sur un noeud de la palette, il est automatiquement relié au noeud sélectionné dans l'espace de travail. Si les noeuds ne sont pas reliés, vous pouvez également utiliser le bouton central de votre souris pour relier le noeud source au noeud Table. Pour simuler l'action du bouton central de la souris, maintenez la touche Alt enfoncée tout en déplaçant la souris. Pour afficher la table, cliquez dans la barre d'outils sur le bouton représentant une flèche verte afin d'exécuter le flux, ou cliquez avec le bouton droit de la souris sur le noeud Table et sélectionnez Exécuter.

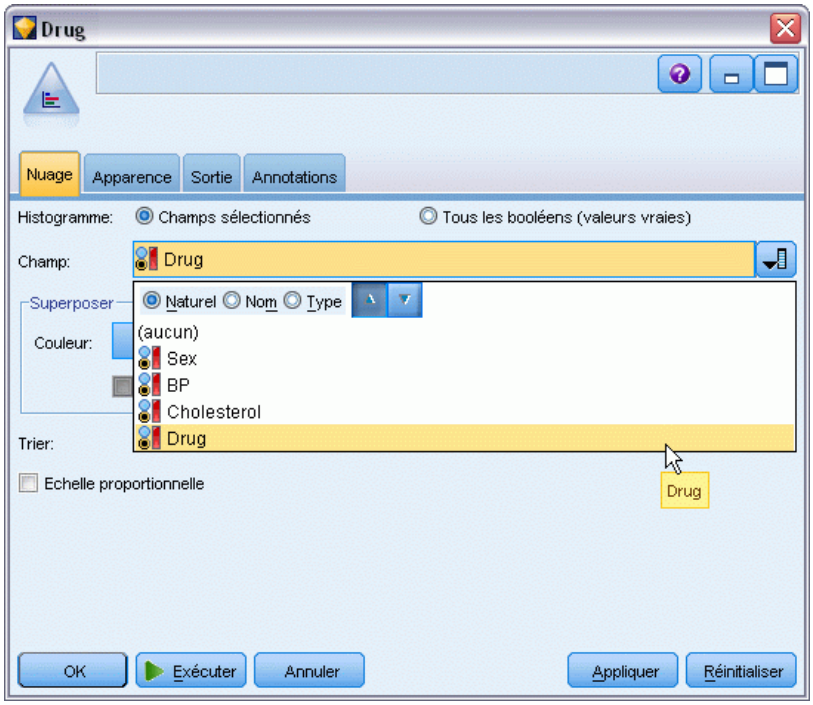
Création d'un graphique Proportion

Dans le cadre du Data mining, il est souvent utile d'explorer les données en créant des récapitulatifs visuels. IBM® SPSS® Modeler propose plusieurs types de graphiques en fonction du genre de données à récapituler. Par exemple, pour connaître la proportion des patients ayant réagi à chaque médicament, utilisez un noeud Proportion.

Ajoutez un noeud Proportion au flux, connectez-le au noeud source, puis double-cliquez dessus pour éditer les options d'affichage.

Sélectionnez *Médicament* comme champ cible dont vous souhaitez afficher la proportion. Ensuite, cliquez sur le bouton Exécuter de la boîte de dialogue.

Figure 8-7
Sélection du médicament en tant que champ cible



Le graphique obtenu vous permet de visualiser « l'aspect » des données. Il démontre que les patients ont réagi le plus souvent au médicament *Y* et moins souvent aux médicaments *B* et *C*.

Figure 8-8
Proportion des réactions à un type de médicament

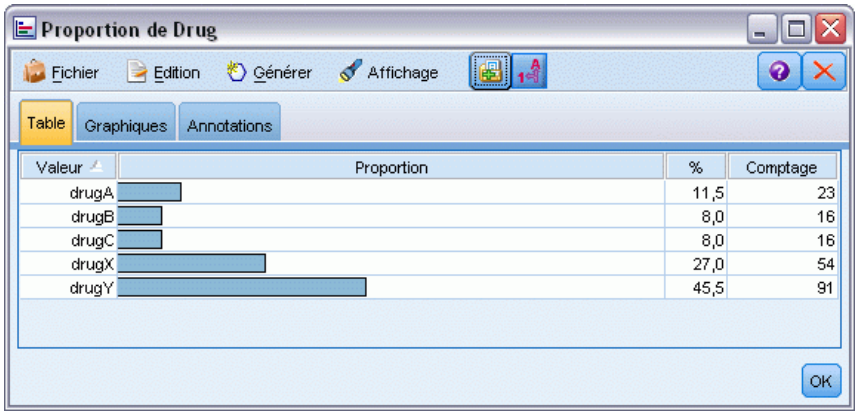
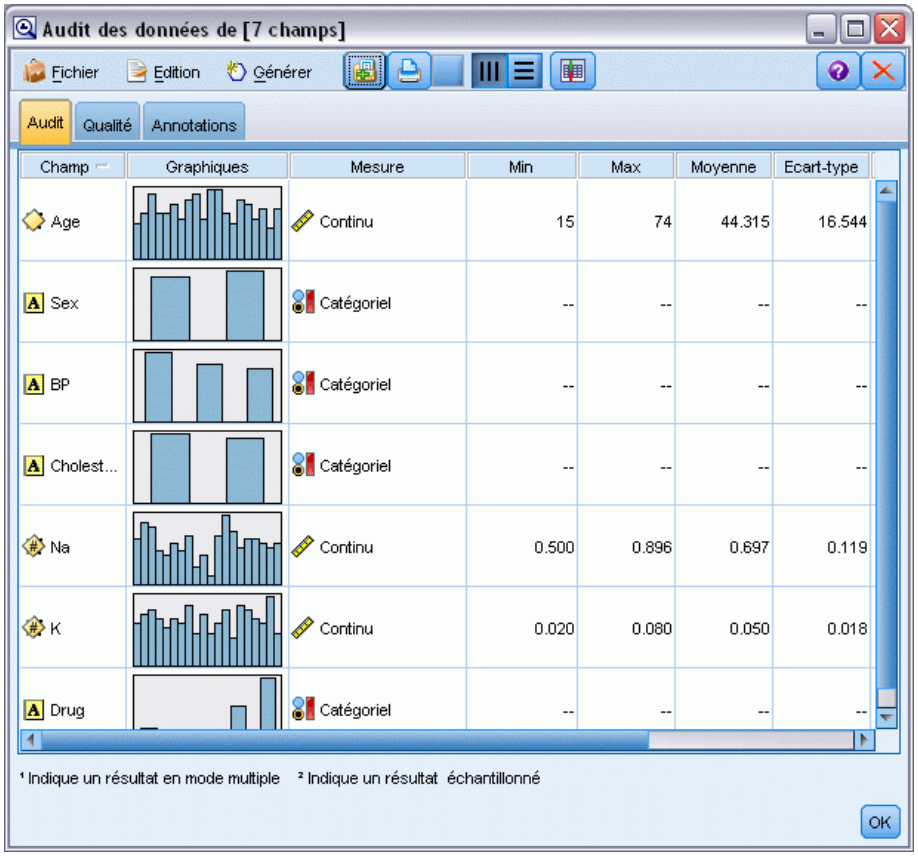


Figure 8-9
Résultats d'un audit de données



Vous pouvez également relier et exécuter un noeud Audit données pour obtenir un aperçu des proportions et des histogrammes de tous les champs simultanément. Le noeud Audit données est disponible dans l'onglet Sortie.

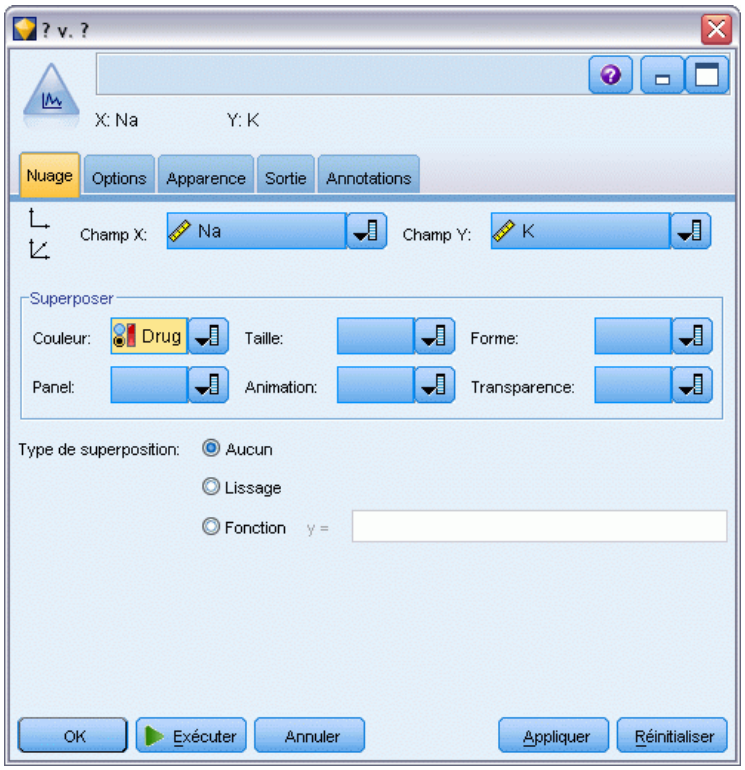
Création d'un diagramme de dispersion

A présent, examinons les facteurs susceptibles d'influencer *Médicament*, la variable cible. En tant que chercheur, vous savez que les concentrations de sodium et de potassium dans le sang sont des facteurs importants. Etant donné qu'il s'agit de valeurs numériques, vous pouvez créer un diagramme de dispersion pour comparer les valeurs du sodium et du potassium, et utiliser les catégories de médicaments en tant que valeurs de superposition.

Placez un noeud Nuage dans l'espace de travail, connectez-le au noeud source, puis double-cliquez dessus pour l'éditer.

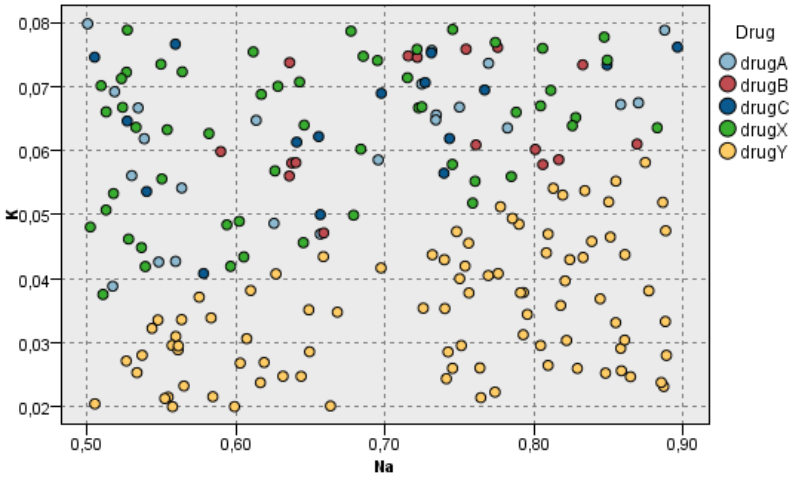
Dans l'onglet Nuage, sélectionnez *Na* comme champ X, *K* comme champ Y et *Médicament* comme champ de superposition. Puis cliquez sur Exécuter.

Figure 8-10
Création d'un diagramme de dispersion



Le nuage montre clairement qu'il existe un seuil au-delà duquel le médicament approprié est toujours le médicament *Y* et en dessous lequel le médicament approprié n'est jamais le médicament *Y*. Ce seuil est un rapport : le rapport entre le sodium (*Na*) et le potassium (*K*).

Figure 8-11
Diagramme de dispersion de la proportion de médicaments

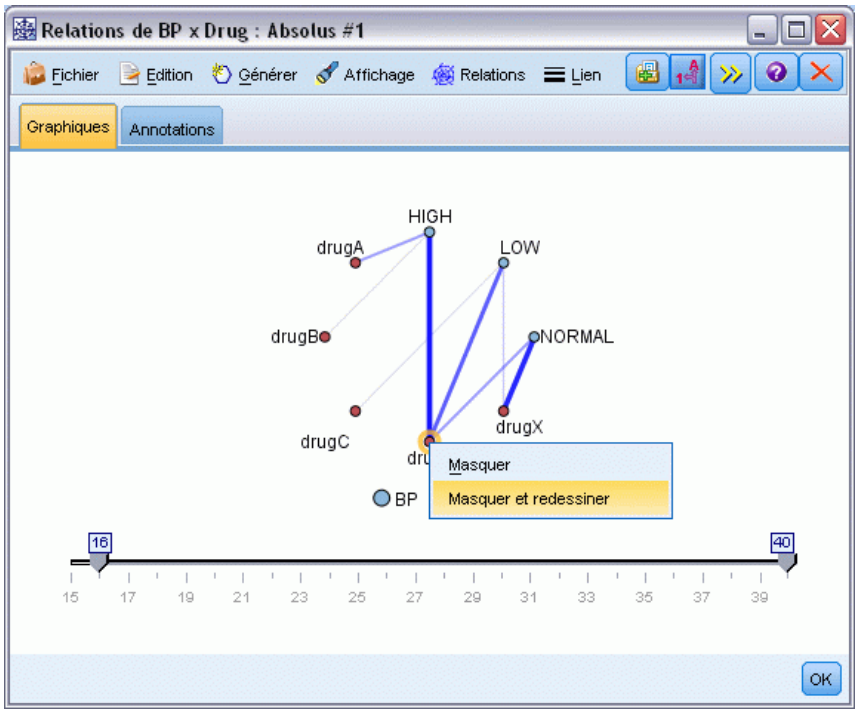


Création d'un graphique Relations

La plupart des champs de données étant de type catégoriel, vous pouvez également essayer de tracer un graphique Relations qui réalise le mappage des associations entre les différentes catégories. Connectez un noeud Relations au noeud source dans l'espace de travail. Dans la boîte de dialogue du noeud Relations, sélectionnez *TA* (pression artérielle) et *Médicament*. Puis cliquez sur Exécuter.

Dans le graphique, il semble que le médicament *Y* soit associé aux trois niveaux de pression artérielle. Ceci n'est pas surprenant : vous aviez déjà déterminé dans quel cas du médicament *Y* est approprié. Pour étudier les autres médicaments, vous pouvez masquer le médicament *Y*. Dans le menu Affichage, choisissez Mode d'édition, puis faites un clic droit sur le médicament *Y* et choisissez Masquer et redessiner.

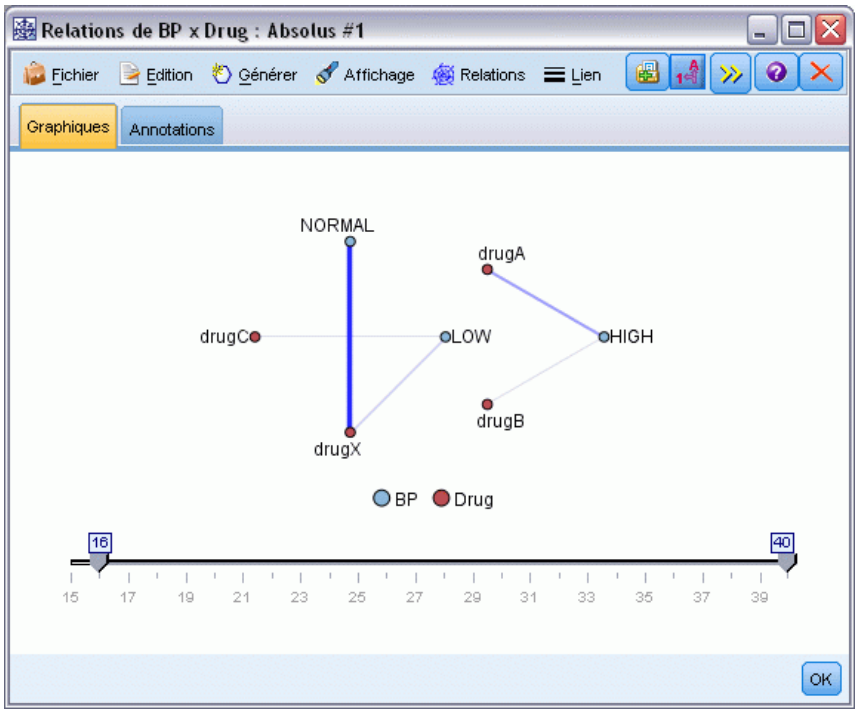
Figure 8-12
Graphique Relations des médicaments en fonction de la tension artérielle



Dans le graphique simplifié, le médicament *Y* et tous ses liens sont masqués Maintenant, vous pouvez voir clairement que seuls les médicaments *A* et *B* sont associés à une pression artérielle élevée. Seuls les médicaments *C* et *X* sont associés à une pression artérielle faible. Seul le médicament *X* est associé à une pression artérielle normale. Cependant, vous ne savez toujours

pas comment choisir entre le médicament *A* et le médicament *B*, ou entre les médicaments *C* et *X* pour un patient donné. C'est dans un cas comme celui-ci que la modélisation peut vous aider.

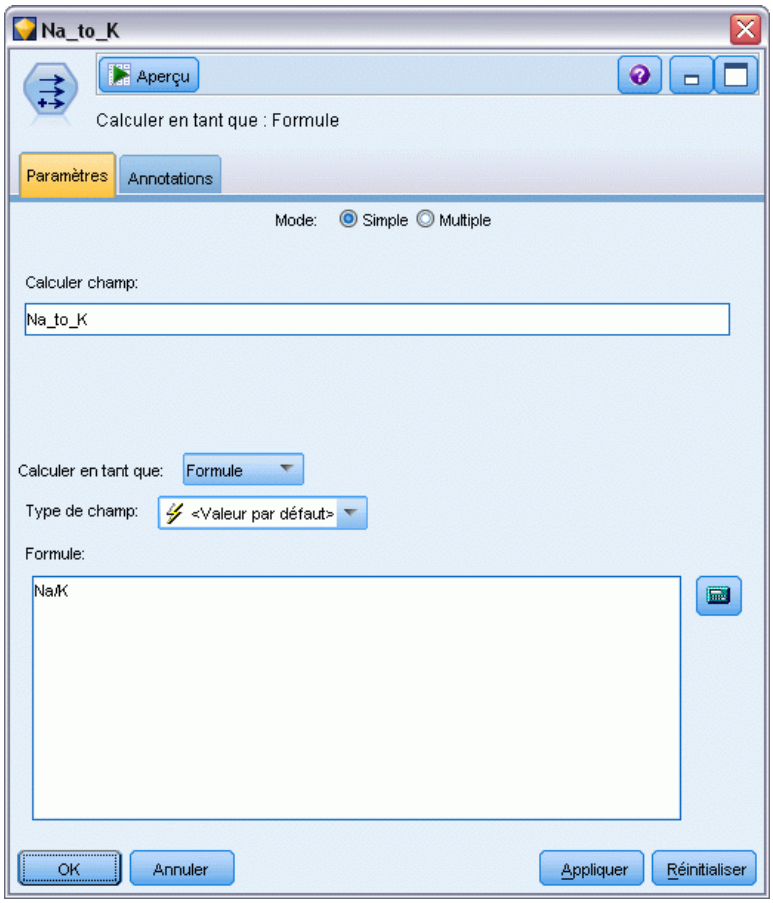
Figure 8-13
Graphique Relations avec médY et ses liens masqués



Calcul d'un nouveau champ

Etant donné que le rapport sodium/potassium semblait indiquer quand utiliser le médicament *Y*, vous pouvez calculer un champ contenant la valeur de ce rapport pour chaque enregistrement. Ce champ peut s'avérer utile par la suite, lors de la création d'un modèle permettant de savoir quand utiliser chacun des cinq médicaments. Pour simplifier la présentation du flux, commencez par effacer tous les noeuds à l'exception du noeud source DRUG1. Reliez un noeud Calculer (Onglet Options de champ) à DRUG1n, puis double-cliquez sur le noeud Calculer pour le modifier.

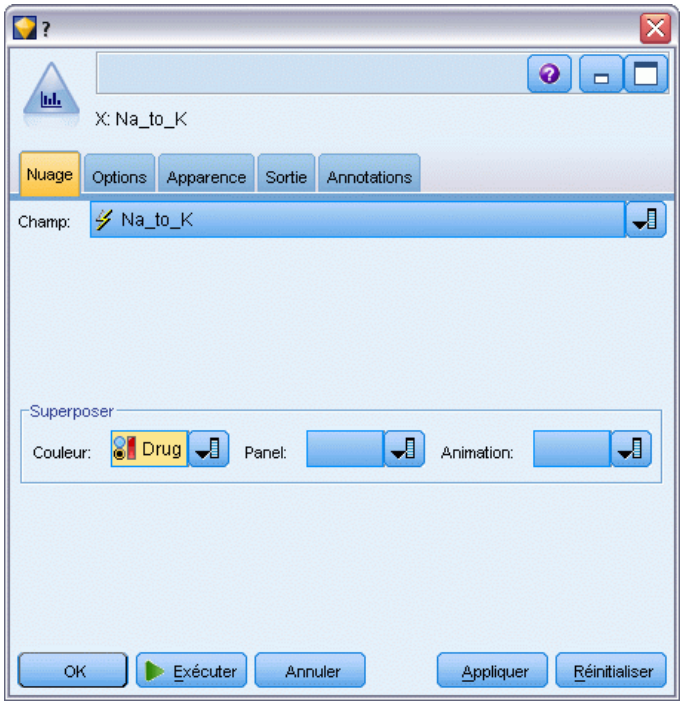
Figure 8-14
Edition du noeud Calculer



Appelez le nouveau champ *Na_sur_K*. Etant donné que vous obtenez le nouveau champ en divisant la valeur du sodium par la valeur du potassium, entrez Na/K dans le champ Formule. Vous pouvez également créer une formule en cliquant sur l'icône située juste à droite du champ. Le Générateur de formules apparaît. Il permet de créer des formules de façon interactive en utilisant des listes de fonctions intégrées, des opérandes, ainsi que des champs et leurs valeurs.

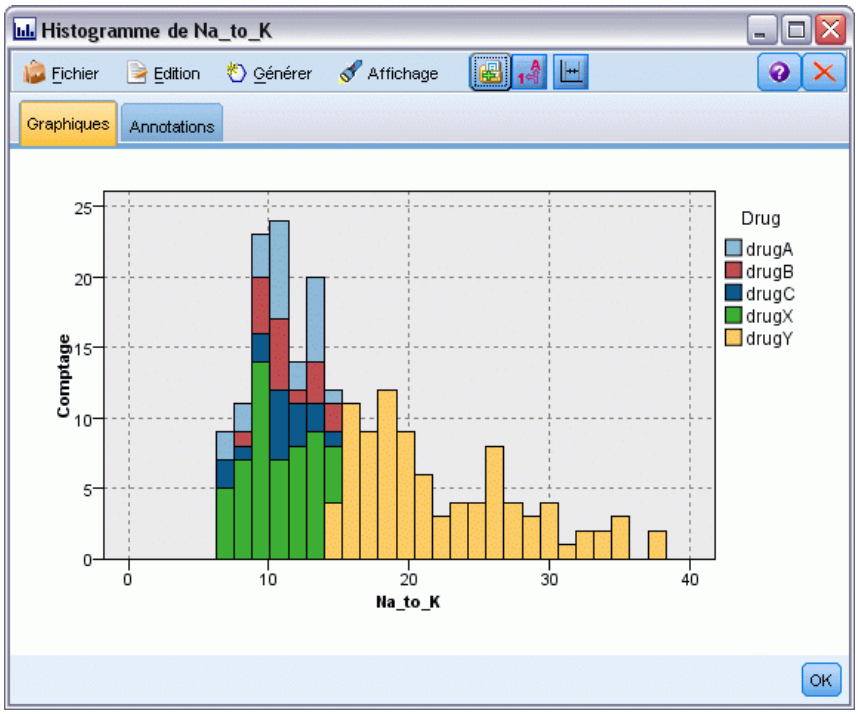
Vous pouvez observer la proportion du nouveau champ en reliant un noeud Histogramme au noeud Calculer. Dans la boîte de dialogue du noeud Histogramme, indiquez que *Na_sur_K* constitue le champ à reporter et *Médicament* le champ de superposition.

Figure 8-15
Edition du noeud Histogramme



Lorsque vous exécutez le flux, vous obtenez le graphique affiché ici. En fonction des éléments affichés, vous pouvez conclure que lorsque la valeur *Na_sur_K* est égale ou supérieure à 15, le médicament recommandé est le médicament *Y*.

Figure 8-16
Histogramme

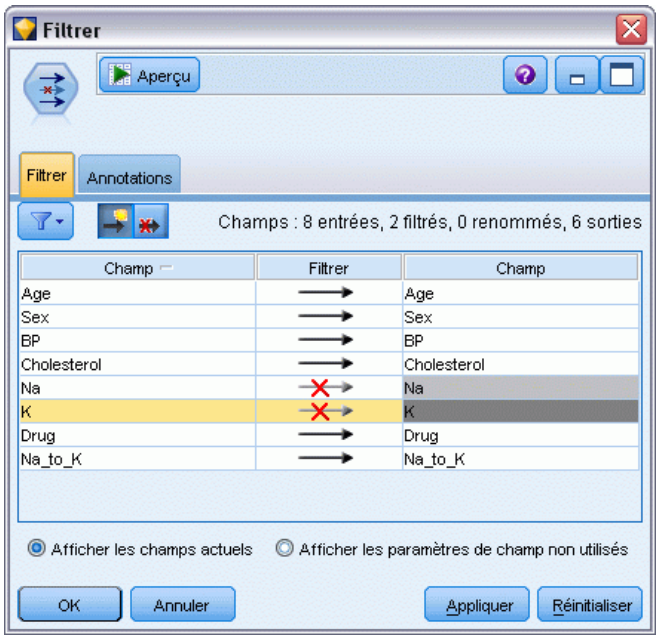


Création d'un modèle

En exploitant et en manipulant les données, vous avez formulé des hypothèses. Le rapport entre le sodium et le potassium dans le sang semble influer sur le choix du médicament, tout comme la pression artérielle. Mais vous ne pouvez pas encore expliquer totalement tous les liens existant entre ces facteurs. La modélisation vous fournira probablement des réponses. Pour cela, vous pouvez essayer d'ajuster les données à l'aide d'un modèle de création de règle, le modèle C5.0.

Etant donné que vous utilisez un champ déjà calculé, *Na_sur_K*, vous pouvez filtrer les champs d'origine, *Na* et *K*, afin d'éviter qu'ils soient utilisés deux fois dans l'algorithme de modélisation. Vous pouvez utiliser pour cela un noeud Filtrer.

Figure 8-17
Edition du noeud Filtrer

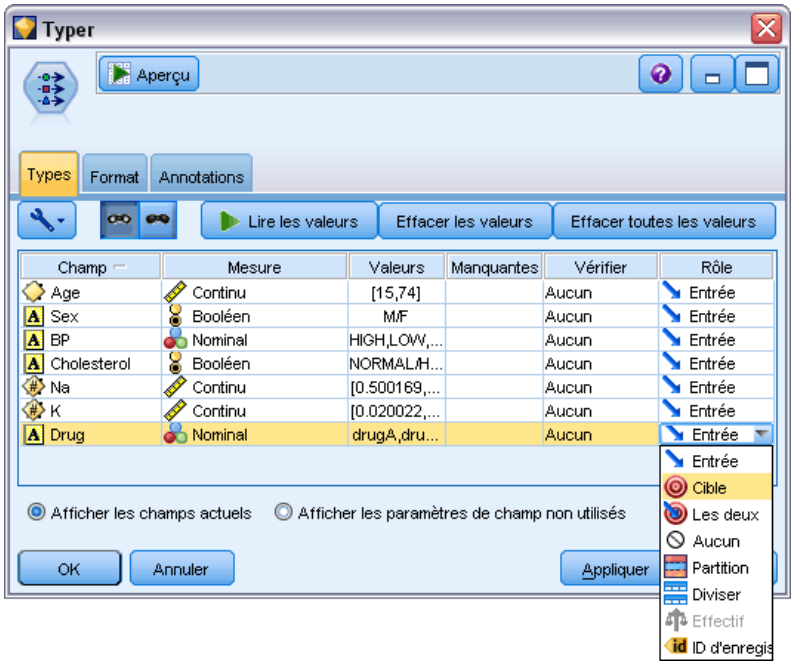


Dans l'onglet Filtrer, cliquez sur les flèches à côté de *Na* et *K*. Des X rouges apparaissent au-dessus des flèches pour indiquer que les champs sont désormais filtrés.

Reliez ensuite un noeud Typer connecté au noeud Filtrer. Le noeud Typer vous permet d'indiquer les types de champ utilisés, ainsi que la façon dont ils seront utilisés pour prédire les résultats.

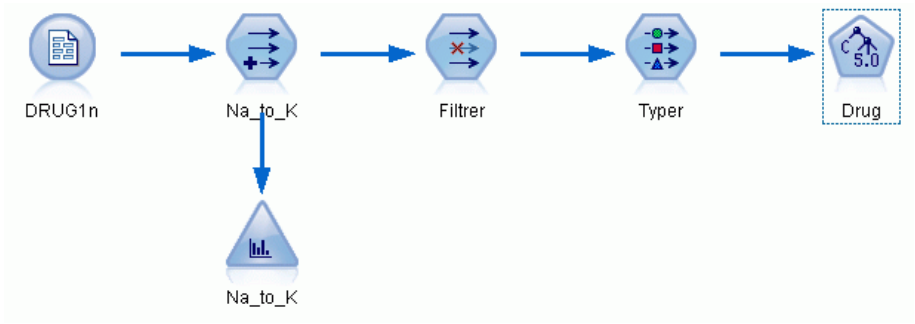
Dans l'onglet Types, attribuez le rôle *Médicament* au champ Cible ; vous indiquez ainsi que le champ *Médicament* est celui sur lequel porte l'analyse. Laissez les autres champs paramétrés sur le rôle Entrée de sorte qu'ils soient utilisés comme variables indépendantes.

Figure 8-18
Edition du noeud Typer



Pour évaluer le modèle, placez un noeud C5.0 dans l'espace de travail et reliez-le à la fin du flux, comme l'indique l'illustration. Puis cliquez sur le bouton vert de la barre d'outil Exécuter pour exécuter le flux.

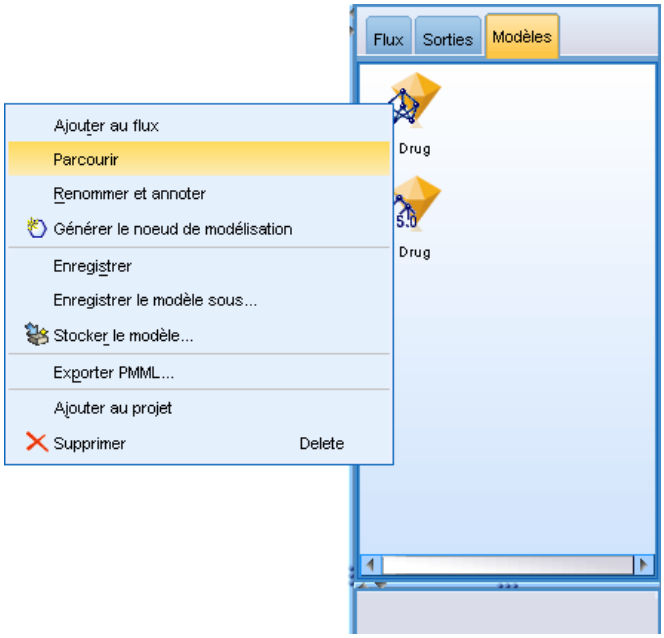
Figure 8-19
Ajout d'un noeud C5.0



Navigation dans le modèle

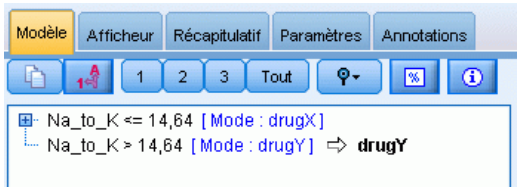
Lorsque le noeud C5.0 est exécuté, le nugget de modèle est ajouté au flux et également à la palette Modèles en haut à droite de la fenêtre. Pour parcourir le modèle, cliquez avec le bouton droit de la souris sur une des icônes, puis sélectionnez Modifier ou Parcourir dans le menu contextuel.

Figure 8-20
Navigation dans le modèle



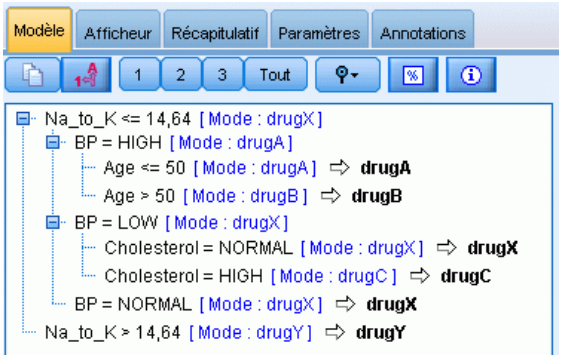
Le navigateur de règles affiche l'ensemble des règles générées par le noeud C5.0 sous forme d'arbre décision. A l'ouverture du navigateur, l'arbre est réduit. Pour le développer et afficher tous les niveaux, cliquez sur le bouton Tout.

Figure 8-21
Navigateur de règles



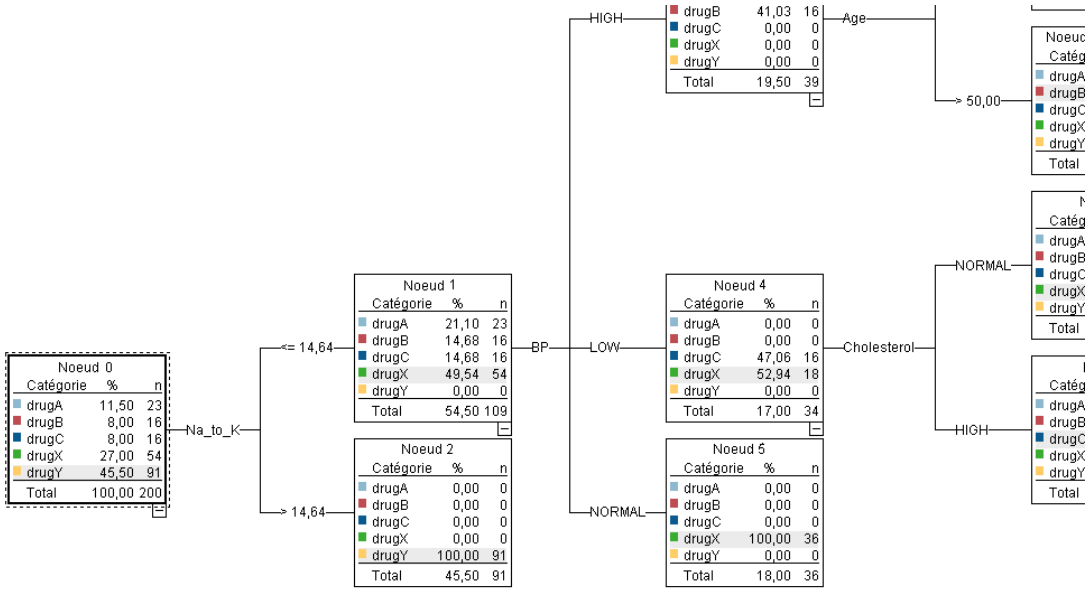
Ainsi, vous pouvez visualiser les éléments manquants. Pour les personnes ayant un rapport *Na*-sur-*K* inférieur à 14,64 et une pression artérielle élevée, l'âge détermine le choix du médicament. Pour les personnes présentant une faible pression artérielle, le taux de cholestérol semble être la variable indépendante optimale.

Figure 8-22
Navigateur de règles développé au maximum



Vous pouvez consulter ce même arbre dans un format de graphique plus élaboré. Pour ce faire, cliquez sur l'onglet Afficheur. Vous pouvez voir plus facilement le nombre d'observations contenues dans chaque catégorie de pression artérielle, ainsi que le pourcentage d'observations.

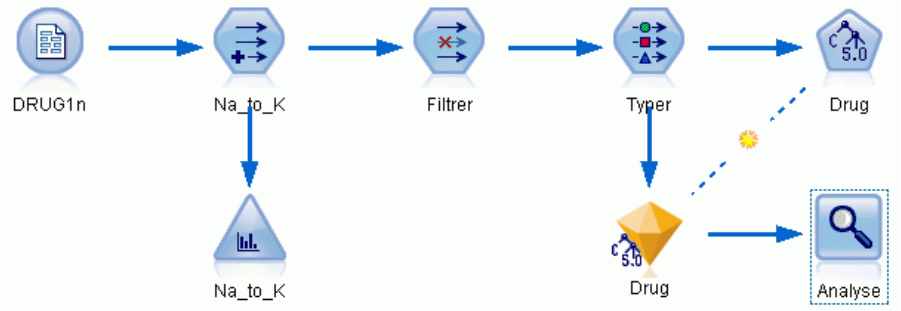
Figure 8-23
Arbre décision en format graphique



Utilisation d'un noeud Analyse

Vous pouvez évaluer l'exactitude du modèle à l'aide d'un noeud Analyse. Reliez un noeud Analyse (de la palette du noeud Sortie) au nugget de modèle, ouvrez le noeud Analyse et cliquez sur Exécuter.

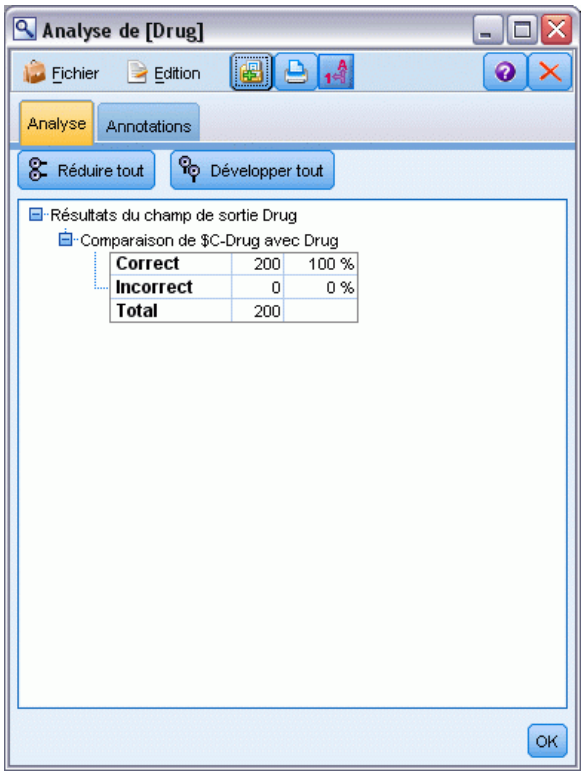
Figure 8-24
Ajout d'un noeud Analyse



Le résultat du noeud Analyse indique que, avec cet ensemble de données artificielles, le modèle a réalisé une prévision correcte du choix de médicament pour chaque enregistrement de l'ensemble de données. Avec un ensemble de données réel, il est peu probable que vous soyez confronté à

une précision de 100 %. Vous pouvez néanmoins utiliser le noeud Analyse pour déterminer si le modèle a une précision acceptable pour votre application.

Figure 8-25
Sortie du noeud Analyse



Chapitre

9

Filtrage des variables indépendantes (sélection de fonction)

Le noeud Sélection de fonction vous permet d'identifier les champs les plus importants pour la prévision de certains résultats. A partir de centaines, voire de milliers de variables indépendantes, le noeud Sélection de fonction filtre, classe et sélectionne celles qui peuvent être les plus importantes. En fin de compte, vous pouvez obtenir un modèle plus rapide et plus efficace (qui utilise moins de variables indépendantes, s'exécute plus rapidement et est plus compréhensible).

Les données utilisées dans cet exemple représentent l'entrepôt de données d'un opérateur de téléphonie fictif et contiennent des informations concernant les réponses données par 5 000 clients de l'opérateur à une promotion spéciale. Ces données incluent un grand nombre de champs comprenant l'âge, la profession et les revenus des clients, ainsi que les statistiques d'utilisation de leur téléphone. Trois champs « cible » indiquent si le client a répondu à chacune des trois offres. L'opérateur souhaite utiliser ces données pour connaître les clients les plus susceptibles de répondre à des offres similaires à l'avenir.

Cet exemple utilise le flux nommé *featureselection.str*, qui fait référence au fichier de données nommé *customer_dbase.sav*. Ces fichiers sont disponibles dans le répertoire *Demos* de n'importe quelle installation de IBM® SPSS® Modeler. Ce dossier est accessible à partir du groupe de programmes IBM® SPSS® Modeler dans le menu Démarrer de Windows. Le fichier *featureselection.str* se trouve dans le répertoire des *flux*.

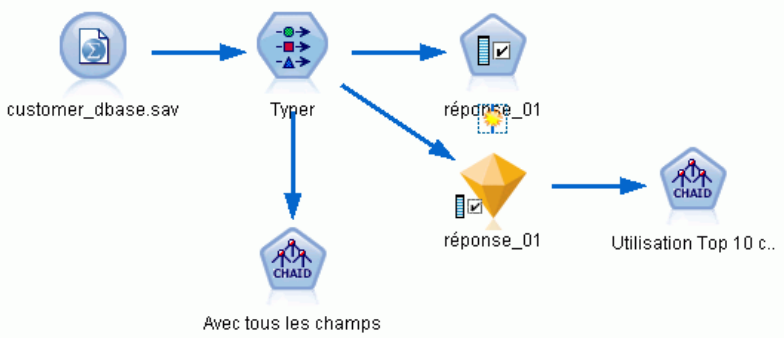
Cet exemple n'emploie comme cible que l'une des offres. Il utilise le noeud de création d'arbre CHAID pour développer un modèle visant à décrire les clients les plus susceptibles de répondre à la promotion. Il met en opposition deux approches :

- Sans la sélection de fonction. Tous les champs variables indépendantes de l'ensemble de données sont employés comme entrées pour l'arbre CHAID.
- Avec la sélection de fonction. Le noeud Sélection de fonction est utilisé pour sélectionner les 10 premières variables indépendantes. Ces variables indépendantes sont ensuite employées comme entrées pour l'arbre CHAID.

Si nous comparons les deux modèles d'arbre obtenus, nous constatons que la sélection de fonction produit des résultats efficaces.

Création du flux

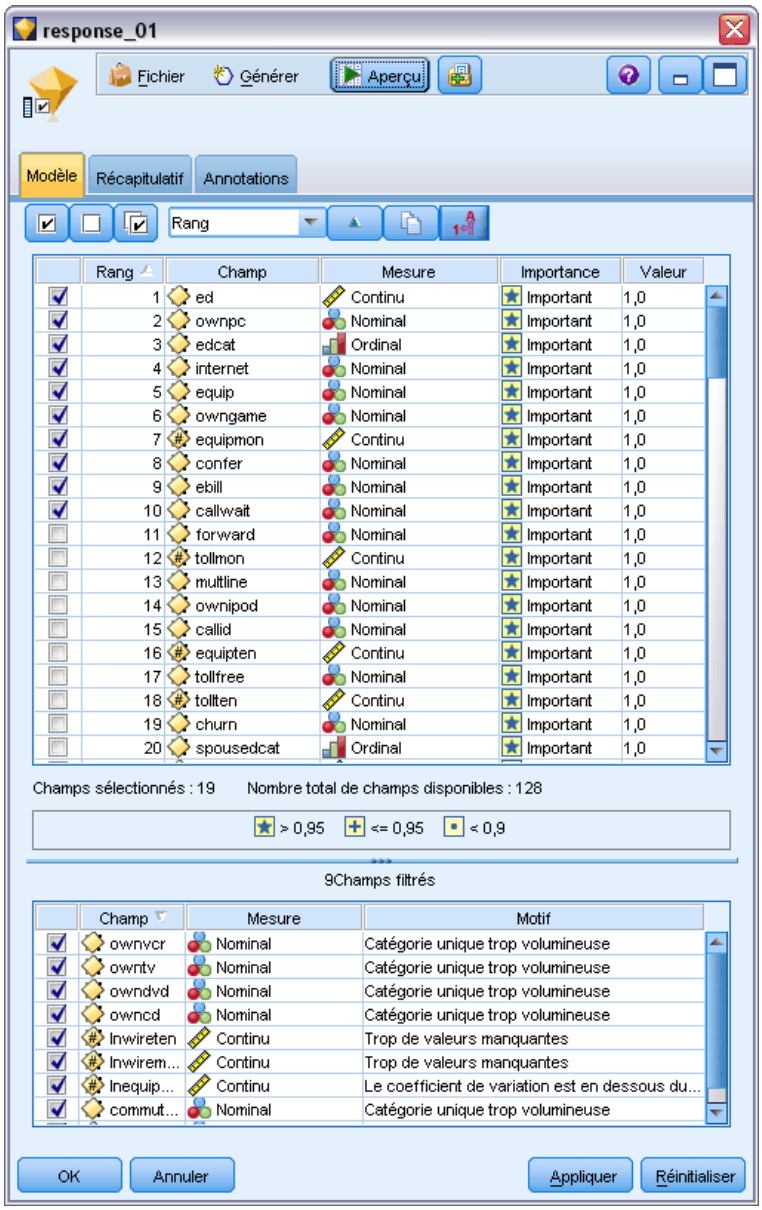
Figure 9-1
Exemple de flux Sélection de fonction



- Placez un noeud source Statistics sur un espace de travail de flux vide. Faites pointer ce noeud vers le fichier de données exemple *customer_dbase.sav*, disponible dans le répertoire *Demos* de votre installation IBM® SPSS® Modeler. (Vous pouvez également ouvrir le fichier de flux exemple *featureselection.str*, dans le répertoire des *flux* .)
- Ajoutez un noeud Typer. Dans l'onglet Types, défilez vers le bas et modifiez le rôle du champ *response_01* en *Cible*. Modifiez le rôle en *Aucun* pour les autres champs de réponse (*response_02*) et *response_03*) ainsi que l'ID client (*custid* en haut de la liste. Laissez le rôle défini sur *Entrée* pour tous les autres champs, et cliquez sur le bouton Lire les valeurs, puis cliquez sur OK.
- Ajoutez au flux un noeud de modélisation Sélection de fonction. Dans ce noeud, vous pouvez définir les règles et les critères de filtrage ou de désactivation des champs.
- Exécutez le flux afin de créer le nugget de modèle Sélection de fonction.

▶ Faites un clic droit sur le nugget de modèle dans le flux ou dans la palette Modèles et choisissez Modifier ou Parcourir pour afficher les résultats.

Figure 9-2
Onglet Modèle dans le nugget de modèle Sélection de fonction



Le panneau supérieur contient les champs considérés comme utiles pour la prévision. Ils sont classés en fonction de leur importance. Le panneau inférieur indique les champs filtrés, et la raison du filtrage. En examinant les champs du panneau supérieur, vous pouvez décider de ceux à utiliser lors des sessions de modélisation suivantes.

- Nous pouvons désormais sélectionnez les champs à employer en aval. Bien que 34 champs aient été identifiés à l'origine comme étant importants, nous souhaitons quand même réduire davantage l'ensemble de variables indépendantes.

- Sélectionnez uniquement les 10 premières variables indépendantes en décochant les cases de la première colonne pour désélectionner les variables indépendantes superflues. (Cliquez sur la coche de la ligne 11, maintenez la touche Maj. appuyée et cliquez sur la coche de la ligne 34). Fermez le nugget de modèle.

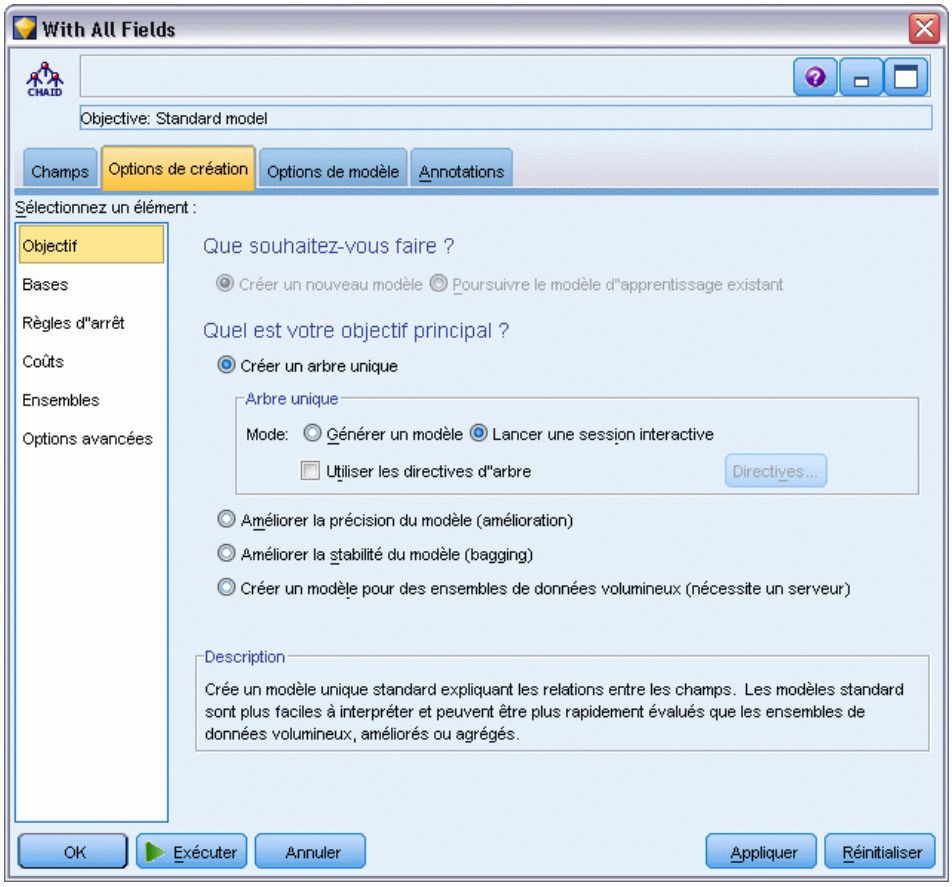
- Pour comparer les résultats sans sélection de fonction, ajoutez deux noeuds de modélisation CHAID au flux : un noeud utilisant la sélection de fonction et un noeud ne s'en servant pas.

- Connectez un noeud CHAID au noeud Typer et l'autre au nugget de modèle Sélection de fonction.

- Ouvrez chacun des noeuds CHAID, sélectionnez l'onglet Options de création et vérifiez que les options Créer un nouveau modèle, Créer un seul arbre et Lancer une session interactive sont sélectionnées dans le panneau Objectifs.

  Dans le panneau Options de base, vérifiez que la Profondeur maximale de l'arbre est définie sur 5.

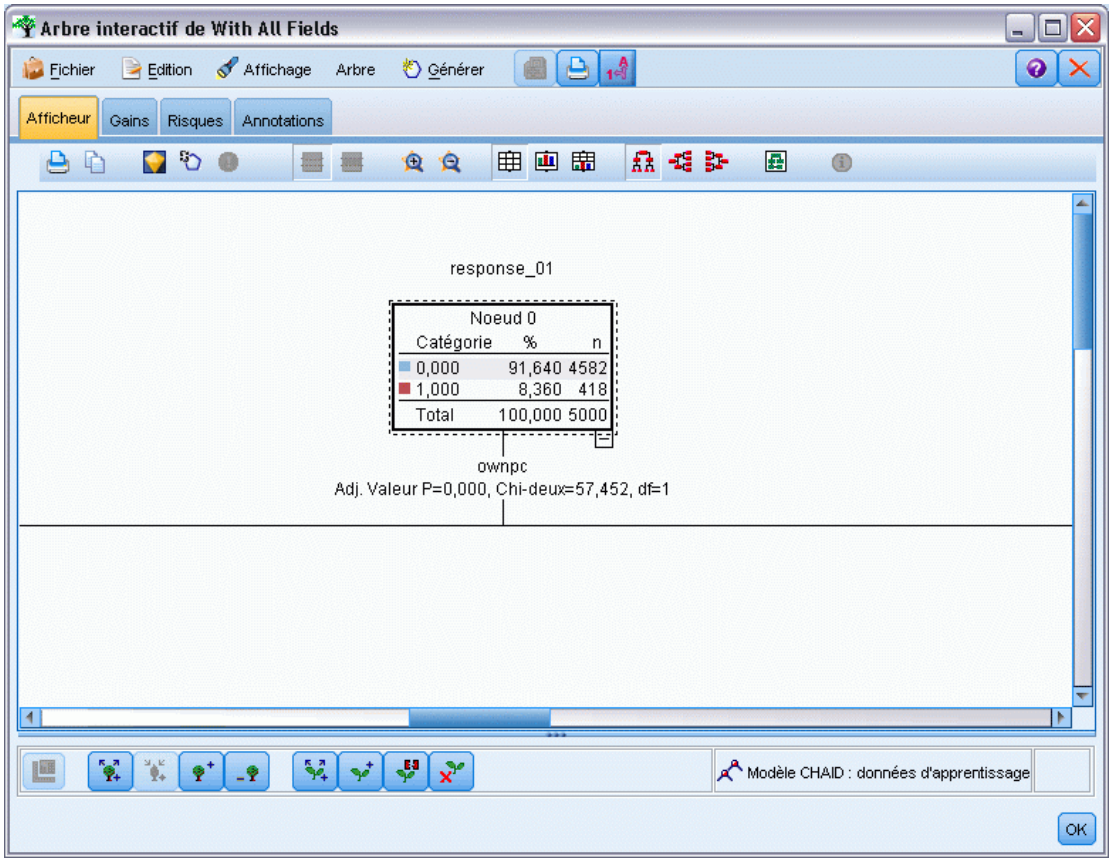
Figure 9-3
Paramètres des objectifs du noeud de modélisation CHAID pour tous les champs variables indépendantes



Création des modèles

- Exécutez le noeud CHAID qui emploie toutes les variables indépendantes de l'ensemble de données (celui connecté au noeud Typer). Notez la durée du traitement. La fenêtre de résultats affiche un tableau.
- Dans les menus, choisissez Arbre > Développer l'arbre pour développer et afficher l'arbre.

Figure 9-4
Développement de l'arbre dans le Générateur d'arbres



- A présent, procédez de même pour l'autre noeud CHAID, qui n'utilise que 10 variables indépendantes. Là encore, développez l'arbre lorsque le Générateur d'arbres s'ouvre.

Le second modèle s'exécute normalement plus vite que le premier. L'ensemble de données considéré étant relativement petit, la différence de temps d'exécution est peut-être de quelques secondes seulement, mais, pour des ensembles de données réels, plus volumineux, cette différence peut s'avérer très importante (plusieurs minutes, voire plusieurs heures). Utiliser la sélection de fonction peut accélérer considérablement vos temps de traitement.

Le second arbre contient également moins de noeuds d'arbre que le premier. Il est plus simple à comprendre. Toutefois, avant de décider de vous en servir, vous devez vérifier qu'il est efficace et le comparer au modèle utilisant toutes les variables indépendantes.

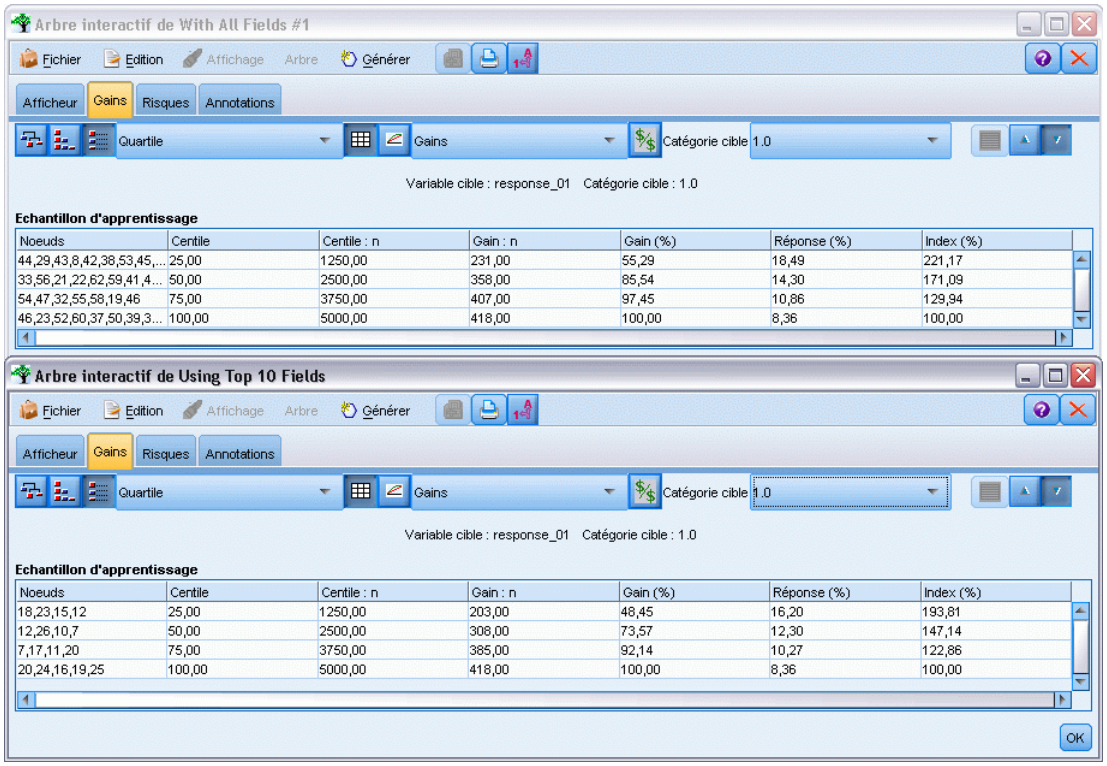
Comparaison des résultats

Pour comparer les deux résultats, nous devons utiliser une mesure d'efficacité. Pour ce faire, utilisons l'onglet Gains du Générateur d'arbres. Portons notre attention sur le **Lift**, qui mesure le degré de probabilité selon lequel les enregistrements d'un noeud peuvent faire partie de la catégorie cible, comparés à tous les enregistrements de l'ensemble de données. Par exemple, une valeur de Lift (augmentation) de 148 % signifie que les enregistrements du noeud ont 1,48 fois plus de chances d'appartenir à la catégorie cible que tous les enregistrements de l'ensemble de données. Le Lift est spécifié dans la colonne *Index* de l'onglet Gains.

- Dans le Générateur d'arbres de l'ensemble complet des variables indépendantes, cliquez sur l'onglet Gains. Définissez la catégorie cible sur 1,0. Passez à un affichage en quartiles. Pour ce faire, cliquez d'abord sur le bouton de la barre d'outils Quantiles. Puis sélectionnez Quartile dans la liste déroulante à droite de ce bouton.

- Répétez cette procédure dans le Générateur d'arbres pour l'ensemble des 10 variables indépendantes, de sorte à avoir deux tableaux de gains similaires à comparer, comme l'illustrent les figures suivantes.

Figure 9-5
Graphiques de gains des deux modèles CHAID



Chaque tableau de gains regroupe les noeuds terminaux de son arbre en quartiles. Pour comparer l'efficacité des deux modèles, examinez l'augmentation (lift) (valeur *Index*) du quartile supérieur dans chaque tableau.

Si toutes les variables indépendantes sont incluses, le modèle affiche une augmentation (Lift) de 221%. Plus précisément, les observations présentant les caractéristiques de ces noeuds ont 2,2 fois plus de chances de répondre à la promotion cible. Pour connaître ces caractéristiques, cliquez sur la ligne supérieure afin de la sélectionner. Passez ensuite à l'onglet Afficheur, où les noeuds correspondants sont désormais mis en évidence en noir. Parcourez l'arbre de haut en bas, jusqu'à chaque noeud terminal mis en évidence, afin de voir comment les variables indépendantes ont été divisées. A lui seul, le quartile supérieur comprend 10 noeuds. Convertis en modèles de scoring réels, 10 profils client différents peuvent être difficiles à gérer.

Avec l'inclusion des 10 premières variables indépendantes (identifiées par la sélection de fonction) seulement, l'augmentation (Lift) est de presque 194 %. Bien que ce modèle ne soit pas aussi performant que celui employant toutes les variables indépendantes, il est indéniablement utile. Dans ce cas, le quartile supérieur n'inclut que quatre noeuds et est donc plus simple. Nous arrivons par conséquent à la conclusion qu'il est préférable d'utiliser le modèle Sélection de fonction au lieu de celui employant toutes les variables indépendantes.

Récapitulatif

Passons à présent en revue les avantages de la sélection de fonction. Utiliser moins de variables indépendantes est plus économique. En effet, vous avez moins de données à collecter, à traiter et à intégrer dans vos modèles. Le temps de calcul s'en trouve amélioré. Dans cet exemple, même avec l'étape supplémentaire de la sélection de fonction, la création du modèle a été nettement plus rapide avec l'ensemble réduit de variables indépendantes. Avec un ensemble de données réel plus volumineux, les gains de temps seraient considérables.

Utiliser moins de variables indépendantes simplifie le scoring. Comme le montre cet exemple, vous ne pouvez identifier que quatre profils de clients susceptibles de répondre à la promotion. Veuillez noter qu'avec des quantités plus importantes de variables indépendantes, vous risqueriez de surajuster votre modèle. Il est possible que le modèle le plus simple se généralise mieux aux autres ensembles de données (mieux vaut néanmoins effectuer un test à titre de vérification).

Pour la sélection de fonction, vous auriez pu utiliser un algorithme de création d'arbre. L'arbre identifie ainsi automatiquement les variables indépendantes les plus importantes. En fait, l'algorithme CHAID est souvent utilisé à cet effet et il est même possible de développer l'arbre niveau par niveau pour en contrôler la profondeur et la complexité. Toutefois, le noeud Sélection de fonction est plus rapide et plus facile à utiliser. Il classe toutes les variables indépendantes en une seule fois, ce qui vous permet d'identifier rapidement les champs les plus importants. En outre, il vous offre la possibilité de changer le nombre de variables indépendantes à inclure. Vous pouvez facilement réappliquer cet exemple, en utilisant cette fois les 15 ou 20 premières variables indépendantes au lieu des 10 premières, afin de comparer les résultats et de déterminer le modèle optimal.

Chapitre
10

Réduction de la longueur des chaînes de données d'entrée (Noeud Recoder)

Réduction de la longueur des chaînes de données d'entrée (Reclassifier)

Pour les modèles de régression logistique et de classificateur automatique qui incluent un modèle de régression logistique binomiale, les champs de type chaîne sont limités à 8 caractères maximum. Lorsque les chaînes contiennent plus de 8 caractères, elles peuvent être recodées à l'aide du noeud Recoder.

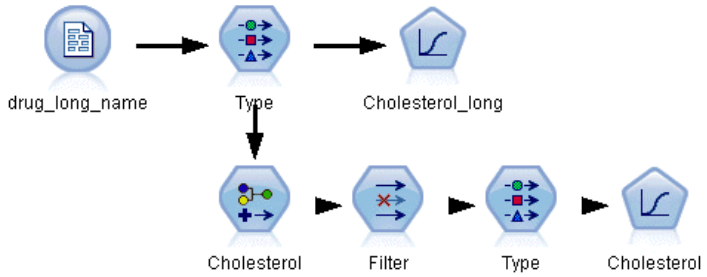
Cet exemple utilise le flux intitulé *reclassify_strings.str*, qui référence le fichier de données *drug_long_name*. Ces fichiers sont disponibles dans le répertoire *Demos* de n'importe quelle installation de IBM® SPSS® Modeler. Ce dossier est accessible à partir du groupe de programmes IBM® SPSS® Modeler dans le menu Démarrer de Windows. Le fichier *reclassify_strings.str* se trouve dans le répertoire des *flux*.

Cet exemple se concentre sur une petite partie d'un flux et présente le type d'erreurs pouvant être générées avec des chaînes trop longues et explique comment utiliser le noeud Recoder pour modifier les détails des chaînes et leur donner une longueur acceptable. Bien que cet exemple utilise un noeud de régression logistique binomiale, il convient également lors de l'utilisation du noeud Classificateur automatique pour générer un modèle de régression logistique binomiale.

Reclassification des données

- A l'aide d'un noeud source Délimité, connectez-vous à l'ensemble de données *drug_long_name* dans le dossier *Demos*.

Figure 10-1
Exemple de flux présentant une reclassification de chaînes pour une régression logistique binomiale



- Ajoutez un noeud Typer au noeud source et sélectionnez Cholesterol_long comme cible.
- Ajoutez un noeud Régression logistique au noeud Typer.

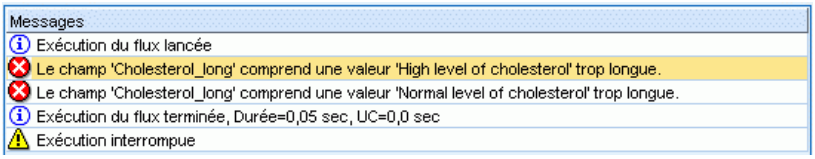
▶ Dans le noeud Régression logistique, cliquez sur l'onglet Modèle et sélectionnez la procédure Binomial.

Figure 10-2
Détails de chaînes de grande longueur dans le champ "Cholesterol_long"

▶ Lorsque vous exécutez le noeud Régression logistique dans *reclassify_strings.str*, un message d'erreur apparaît pour vous prévenir que les valeurs de chaîne Cholesterol_long sont trop longues.

Si vous rencontrez ce genre de messages d'erreur, suivez la procédure expliquée dans le reste de cet exemple pour modifier vos données.

Figure 10-3
Message d'erreur affiché lors de l'exécution du noeud de régression logistique binomiale



▶ Ajoutez un noeud Recoder au noeud Typer.

▶ Dans le champ Reclassifier, sélectionnez Cholesterol_long.

▶ Saisissez Cholesterol comme nouveau nom de champ.

▶ Cliquez sur le bouton Obtenir pour ajouter les valeursCholesterol_long à la colonne de valeurs d'origine.

- Dans la nouvelle colonne de valeurs, saisissez Elevé à côté de la valeur d'origine du Niveau de cholestérol élevé et Normal à côté de la valeur d'origine de Niveau de cholestérol normal.

Figure 10-4
Reclassification des chaînes longues

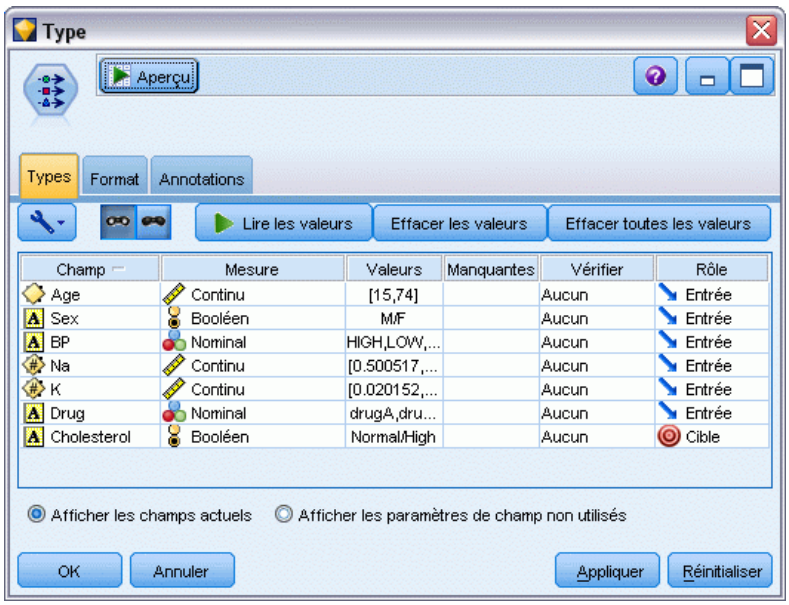
- Ajoutez un noeud Filtrer au noeud Recoder.

▶ Dans la colonne Filtrer, cliquez pour supprimer Cholesterol_long.

Figure 10-5
Filtrage du champ "Cholesterol_long" à partir des données

▶ Ajoutez un noeud Typer au noeud Filtrer et sélectionnez Cholesterol comme cible.

Figure 10-6
Détails de chaînes courtes dans le champ "Cholesterol"

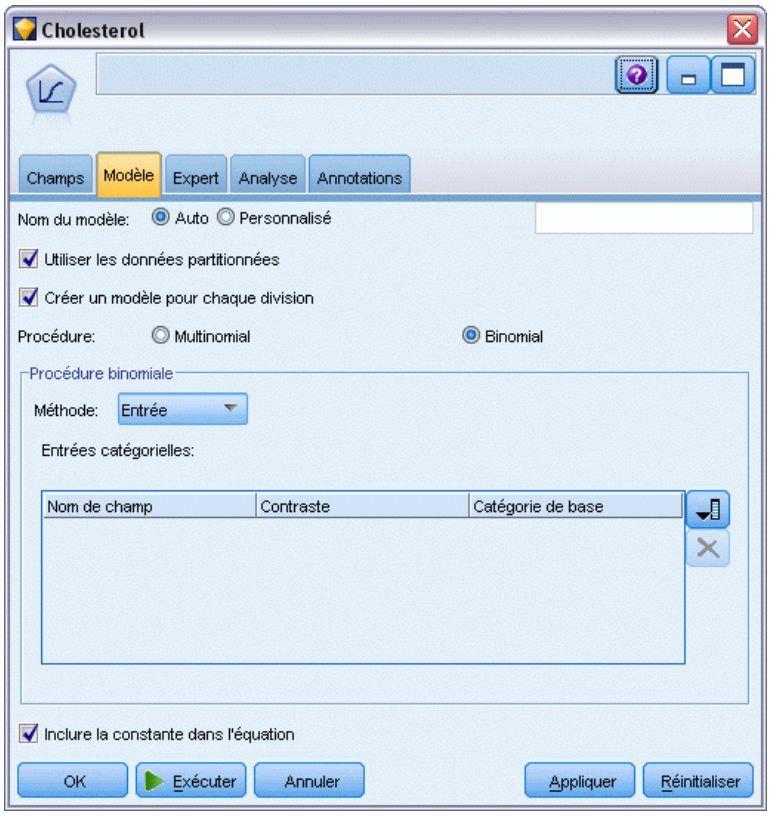


▶ Ajoutez un noeud Logistique au noeud Typer.

▶ Dans le noeud Logistique, cliquez sur l'onglet Modèle et sélectionnez la procédure Binomial.

▶ Vous pouvez maintenant exécuter le noeud Logistique binomiale et générer un modèle sans qu'un message d'erreur ne s'affiche.

Figure 10-7
Choix de la procédure Binomial



Cet exemple ne présente qu'une partie d'un flux. Si vous avez besoin d'informations supplémentaires sur les types de flux dans lesquels vous pouvez avoir besoin de reclassifier de longues chaînes, les exemples suivants sont disponibles :

- Noeud Classificateur automatique. Pour plus d'informations, reportez-vous à la section Modélisation de la réponse client (Classificateur automatique) dans le chapitre 4 sur p. 45.
- noeud Régression logistique binomiale. Pour plus d'informations, reportez-vous à la section Attrition dans le domaine des télécommunications (régression logistique binomiale) dans le chapitre 13 sur p. 163.

Des informations supplémentaires sur l'utilisation de IBM® SPSS® Modeler, telles que le guide de l'utilisateur, le guide de référence des noeuds et le guide des algorithmes, sont disponibles dans le répertoire *Documentation* du disque d'installation.

Partie III:
Exemples de modélisation

Chapitre
11

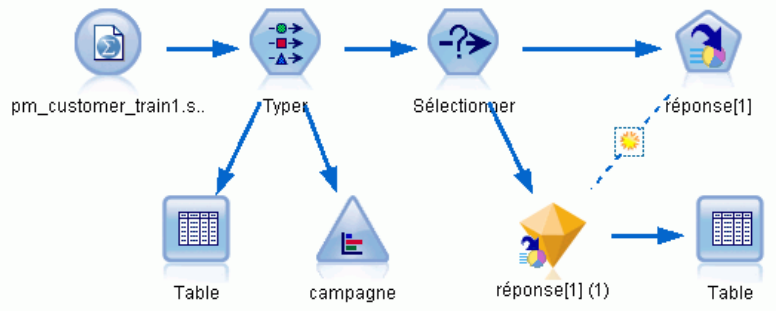
Modélisation de la réponse client (Liste de décision)

L'algorithme Liste de décision génère des règles qui indiquent une probabilité plus ou moins élevée d'obtenir un résultat binaire (oui ou non) donné. Les modèles Liste de décision sont largement utilisés dans la gestion de la relation client, par exemple dans les centres d'appel ou les applications marketing.

Cet exemple repose sur une société fictive qui souhaite obtenir des résultats plus rentables au cours des prochaines campagnes de marketing en présentant à chaque client une offre adaptée. En particulier, l'exemple utilise un modèle Liste de décision pour identifier les caractéristiques des clients les plus à même de répondre favorablement, sur la base des promotions précédentes, et de générer un fichier d'adresses en fonction des résultats.

Les modèles Liste de décisions sont particulièrement adaptés à la modélisation interactive et vous permettent de régler les paramètres du modèle et d'obtenir des résultats immédiats. Si vous souhaitez utiliser une autre approche qui vous permet de créer automatiquement plusieurs modèles différents et de classer les résultats obtenus, utilisez le noeud Classificateur automatique.

Figure 11-1
Exemple de flux Liste de décision

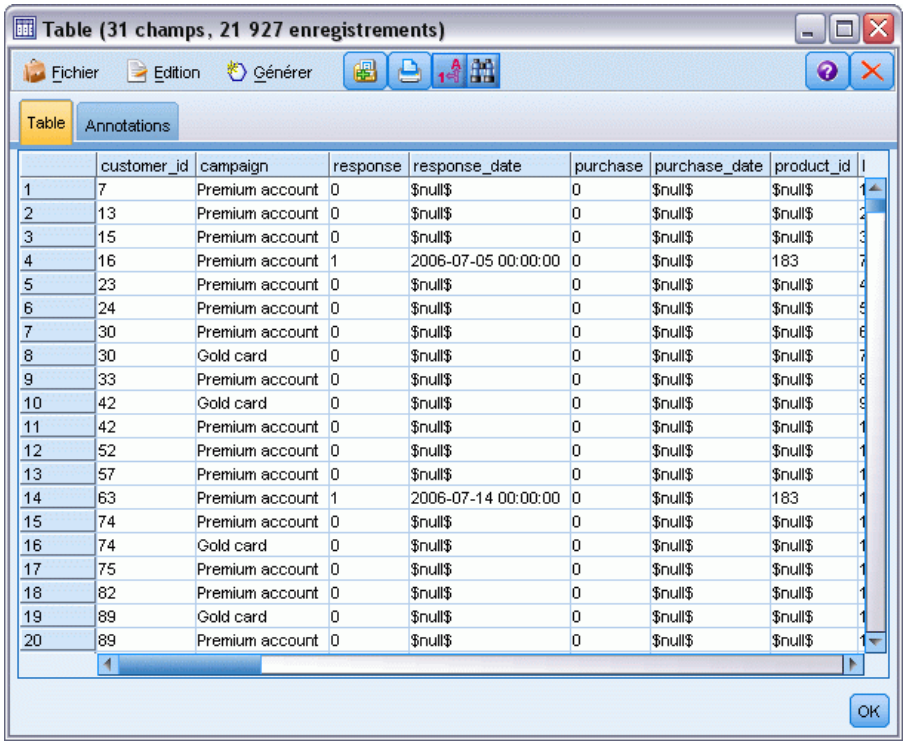


Cet exemple utilise le flux *pm_decisionlist.str*, qui fait référence au fichier de données *pm_customer_train1.sav*. Ces fichiers sont disponibles dans le répertoire *Demos* de n'importe quelle installation de IBM® SPSS® Modeler. Ce dossier est accessible à partir du groupe de programmes IBM® SPSS® Modeler dans le menu Démarrer de Windows. Le fichier *pm_decisionlist.str* se trouve dans le répertoire des *flux*.

Données historiques

Le fichier *pm_customer_train1.sav* comporte des données historiques suivant les offres faites à des clients spécifiques au cours de campagnes passées, comme l'indique la valeur du champ *campaign*. Le plus grand nombre d'enregistrements se trouve dans la campagne *Premium account*.

Figure 11-2
Données sur les anciennes promotions



	customer_id	campaign	response	response_date	purchase	purchase_date	product_id	l
1	7	Premium account	0	$null$	0	$null$	$null$	1
2	13	Premium account	0	$null$	0	$null$	$null$	2
3	15	Premium account	0	$null$	0	$null$	$null$	3
4	16	Premium account	1	2006-07-05 00:00:00	0	$null$	183	7
5	23	Premium account	0	$null$	0	$null$	$null$	4
6	24	Premium account	0	$null$	0	$null$	$null$	5
7	30	Premium account	0	$null$	0	$null$	$null$	6
8	30	Gold card	0	$null$	0	$null$	$null$	7
9	33	Premium account	0	$null$	0	$null$	$null$	8
10	42	Gold card	0	$null$	0	$null$	$null$	9
11	42	Premium account	0	$null$	0	$null$	$null$	1
12	52	Premium account	0	$null$	0	$null$	$null$	1
13	57	Premium account	0	$null$	0	$null$	$null$	1
14	63	Premium account	1	2006-07-14 00:00:00	0	$null$	183	1
15	74	Premium account	0	$null$	0	$null$	$null$	1
16	74	Gold card	0	$null$	0	$null$	$null$	1
17	75	Premium account	0	$null$	0	$null$	$null$	1
18	82	Premium account	0	$null$	0	$null$	$null$	1
19	89	Gold card	0	$null$	0	$null$	$null$	1
20	89	Premium account	0	$null$	0	$null$	$null$	1

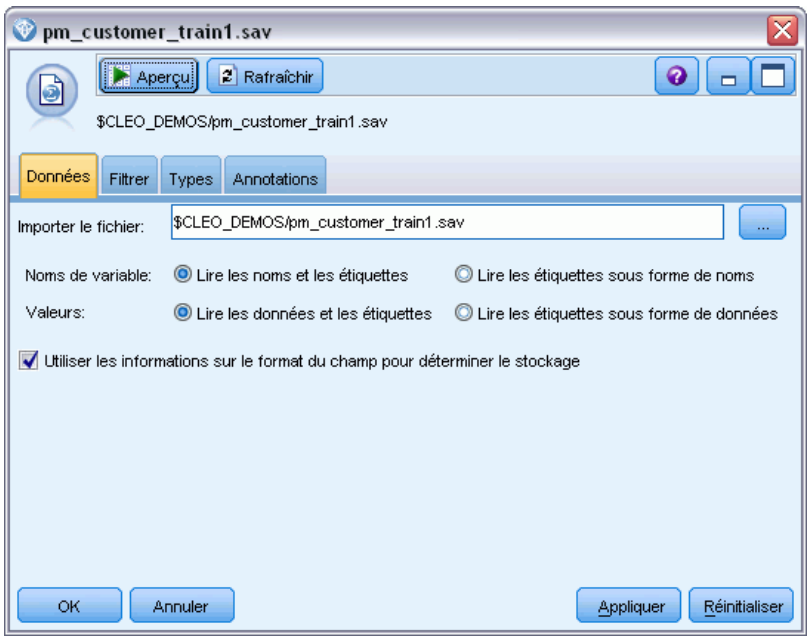
Les valeurs du champ *campaign* sont en fait codées comme des entiers dans les données, avec des étiquettes définies dans le noeud Typer (par exemple *2 = Premium account*). Vous pouvez masquer ou afficher les étiquettes de valeur dans le tableau à l'aide de la barre d'outils.

Le fichier inclut aussi un certain nombre de champs contenant des informations démographiques et financières sur chaque client qui peut servir à créer ou à « former » un modèle qui prévoit les taux de réponse pour différents groupes en fonction de caractéristiques spécifiques.

Création du flux

- Ajoutez un noeud source Statistics qui pointe sur *pm_customer_train1.sav*, dans le dossier *Demos* du répertoire d'installation de IBM® SPSS® Modeler. (Vous pouvez spécifier $CLEO_DEMOS/ dans le chemin du fichier comme raccourci de référence de ce dossier.)

Figure 11-3
Lecture de données



- Ajoutez un noeud Typer, puis sélectionnez *Réponse* en tant que champ cible (Rôle = Cible). Paramétrez le niveau de mesure de ce champ sur Booléen.

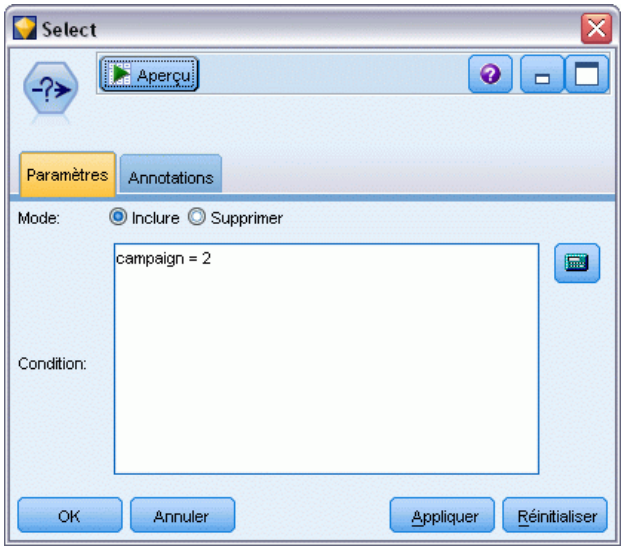
Figure 11-4
Configuration du niveau de mesure et du rôle

- Paramétrez le rôle sur Aucun pour les champs suivants : *customer_id*, *campaign*, *response_date*, *purchase*, *purchase_date*, *product_id*, *Rowid* et *X_random*. Ces champs ont tous des utilisations dans les données mais ne seront pas utilisés pour la création du modèle réel.

- Cliquez sur le bouton Lire les valeurs dans le noeud Typer pour vérifier que les valeurs sont instanciées.

Les données incluent des informations sur quatre campagnes différentes, mais vous vous concentrerez sur l'analyse d'une seule campagne à la fois. Comme le plus grand nombre d'enregistrements se trouve dans la campagne Premium (codée *campaign = 2* dans les données), vous pouvez utiliser un noeud Sélectionner pour n'inclure que ces enregistrements dans le flux.

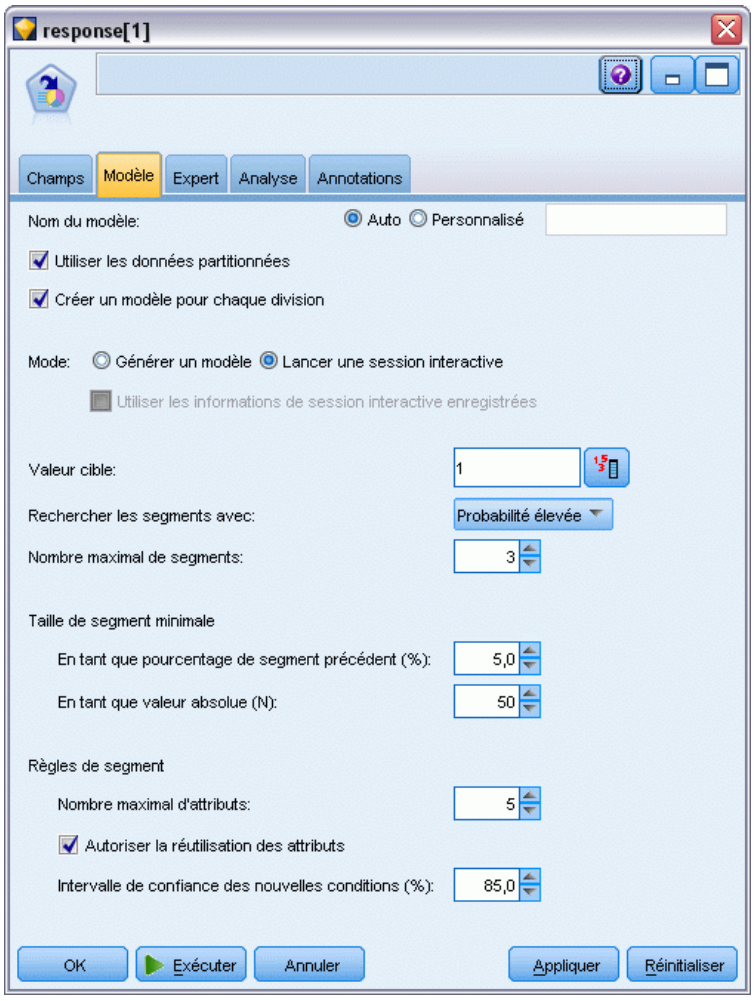
Figure 11-5
Sélection d'enregistrements pour une seule campagne



Création du modèle

- Reliez un noeud Liste de décision au flux. Dans l'onglet Modèle, définissez la Valeur cible sur 1 pour indiquer le résultat que vous souhaitez rechercher. Dans notre exemple, vous recherchez des clients qui ont répondu *Oui* à une offre précédente.

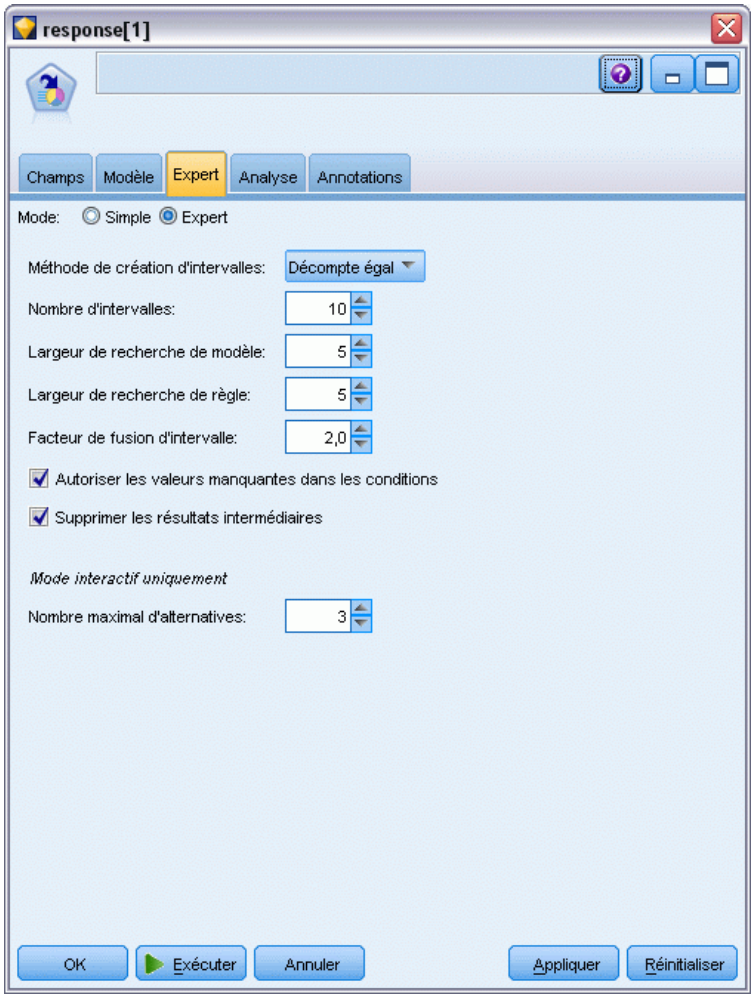
Figure 11-6
Noeud Liste de décision, onglet Modèle



- Sélectionnez Lancer une session interactive.
- Pour conserver la simplicité du modèle pour cet exemple, paramétrez le nombre maximum de segments sur 3.
- Changez l'intervalle de confiance pour les nouvelles conditions à 85 %.

▶ Dans l'onglet Expert, définissez le Mode sur Expert.

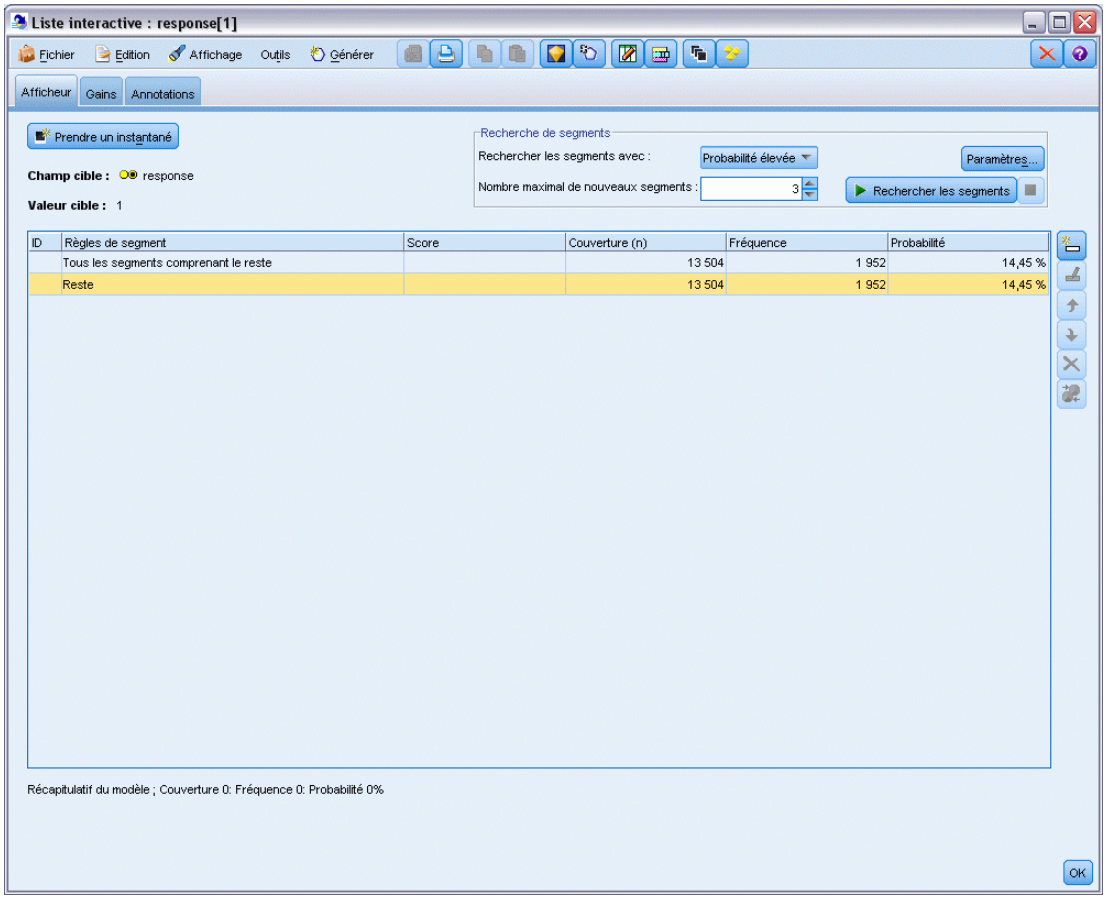
Figure 11-7
Noeud Liste de décision, onglet Expert



▶ Augmentez le Nombre maximal d'alternatives à 3. Cette option fonctionne en association avec le paramètre Lancer une session interactive que vous avez sélectionné dans l'onglet Modèle.

▶ Cliquez sur Exécuter pour afficher l'afficheur Liste interactive.

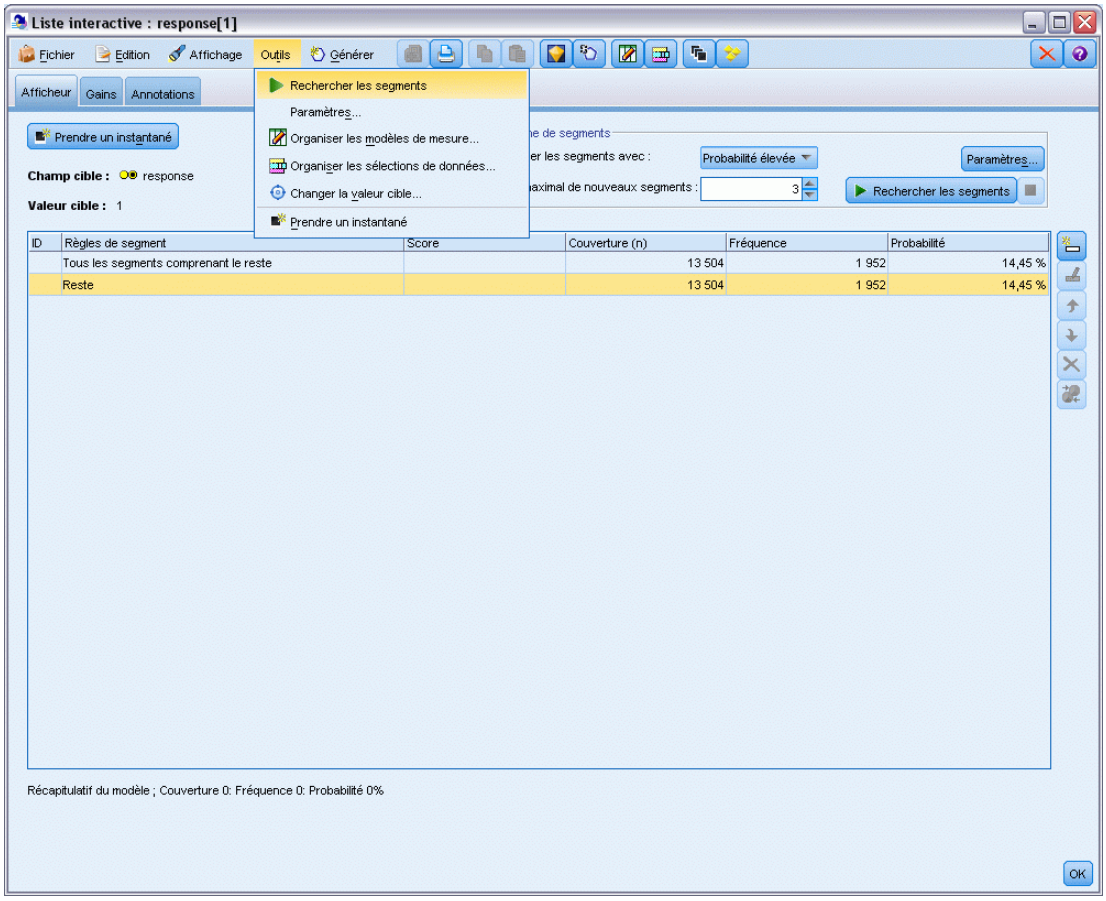
Figure 11-8
Afficheur Liste interactive



Etant donné qu'aucun segment n'a encore été défini, tous les enregistrements sont inclus dans le reste. Sur les 13 504 enregistrements que compte l'échantillon, 1 952 ont dit *Oui*, soit un taux de correspondance global de 14,45%. Vous souhaitez améliorer ce taux en identifiant les segments de clients les plus (ou les moins) susceptibles de donner une réponse favorable.

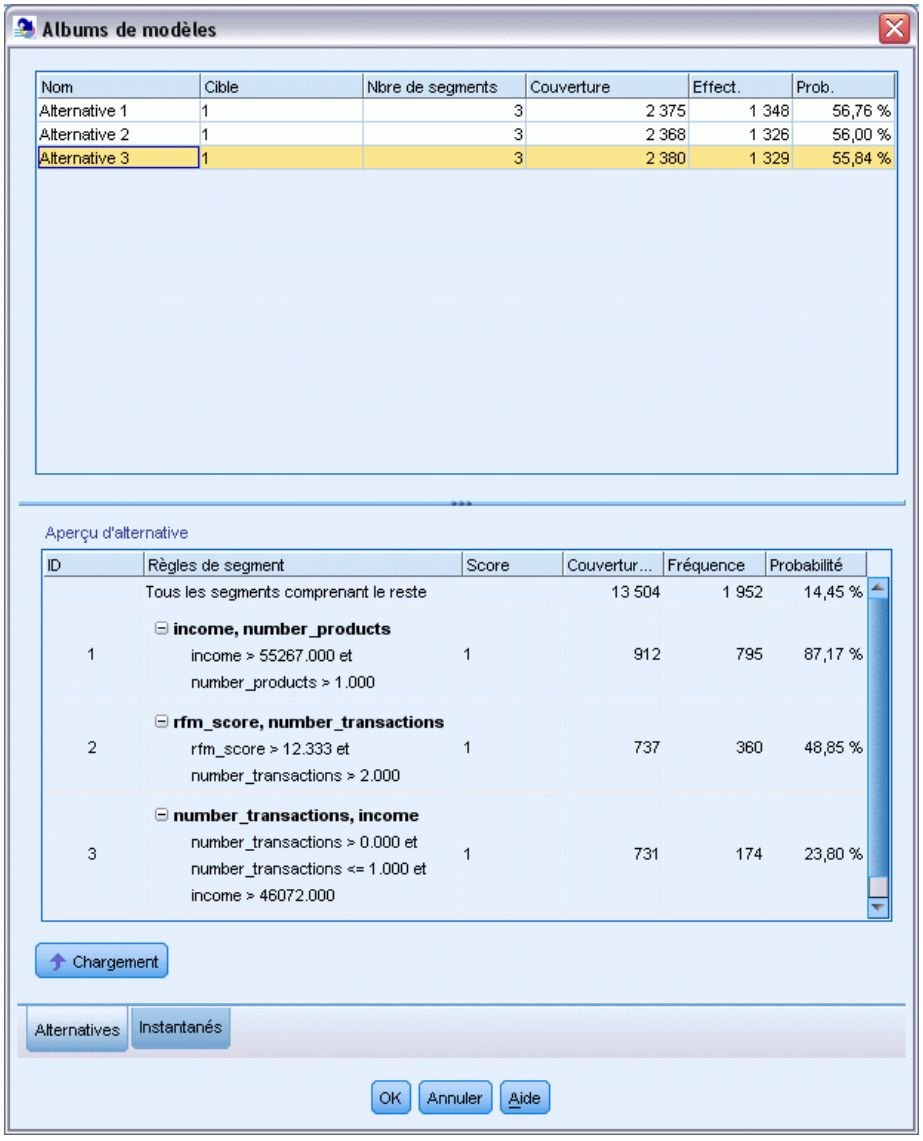
▶ Dans les menus de l'afficheur Liste interactive, choisissez :
Outils > Rechercher les segments

Figure 11-9
Afficheur Liste interactive



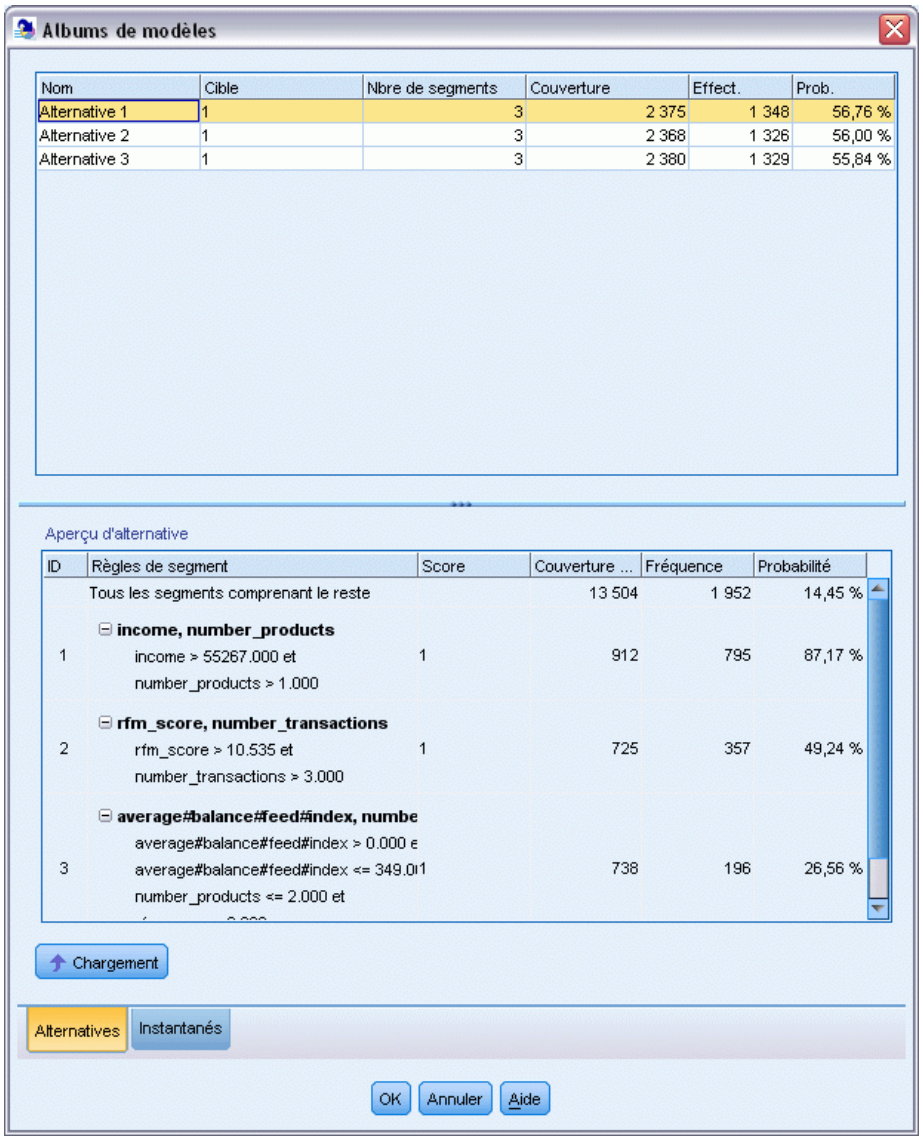
Cette option exécute la tâche exploratoire par défaut sur la base des paramètres que vous avez définis dans le noeud Liste de décision. La tâche terminée renvoie trois modèles alternatifs, qui sont répertoriés dans l'onglet Alternatives de la boîte de dialogue Albums de modèles.

Figure 11-10
Modèles alternatifs disponibles



▶ Sélectionnez la première alternative dans la liste ; ses détails sont affichés dans le panneau Aperçu de l'alternative.

Figure 11-11
Modèle alternatif sélectionné



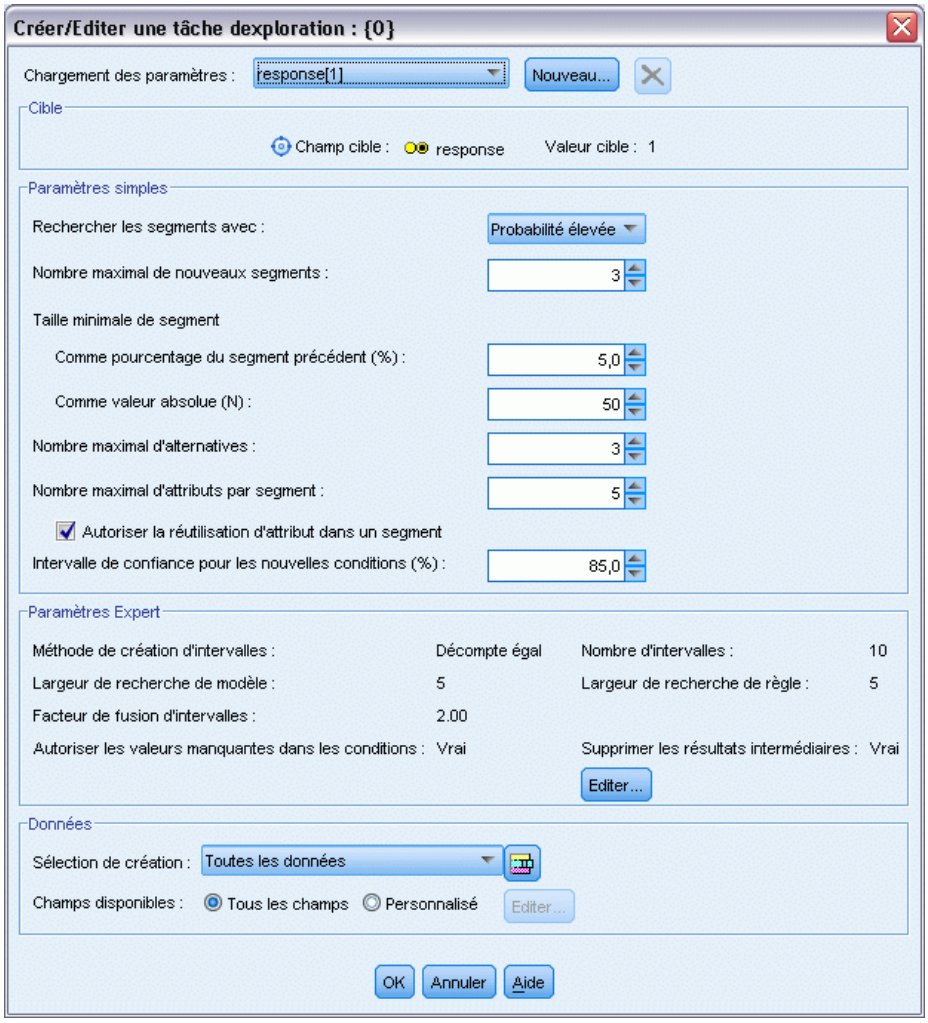
Le panneau Aperçu de l'alternative vous permet de parcourir rapidement plusieurs alternatives sans changer le modèle de travail, ce qui facilite l'expérimentation de différentes approches.

Remarque : Pour mieux voir le modèle, vous pouvez agrandir le panneau Aperçu de l'alternative dans la boîte de dialogue comme l'indique l'illustration. Pour ce faire, faites glisser la bordure du panneau.

En utilisant des règles basées sur les variables indépendantes comme le revenu, le nombre de transactions par mois et le score RFM, le modèle identifie des segments avec des taux de réponse qui sont plus élevés que ceux de l'ensemble de l'échantillon. Lorsque les éléments sont combinés, ce modèle suggère que vous pouvez améliorer votre taux de correspondance jusqu'à 56,76%. Cependant, le modèle ne couvre qu'une petite portion de l'échantillon global et plus de 11 000 enregistrements (dont plusieurs centaines de correspondances) sont inclus dans le reste. Vous recherchez un modèle qui capture davantage de ces correspondances tout en excluant toujours les segments peu performants.

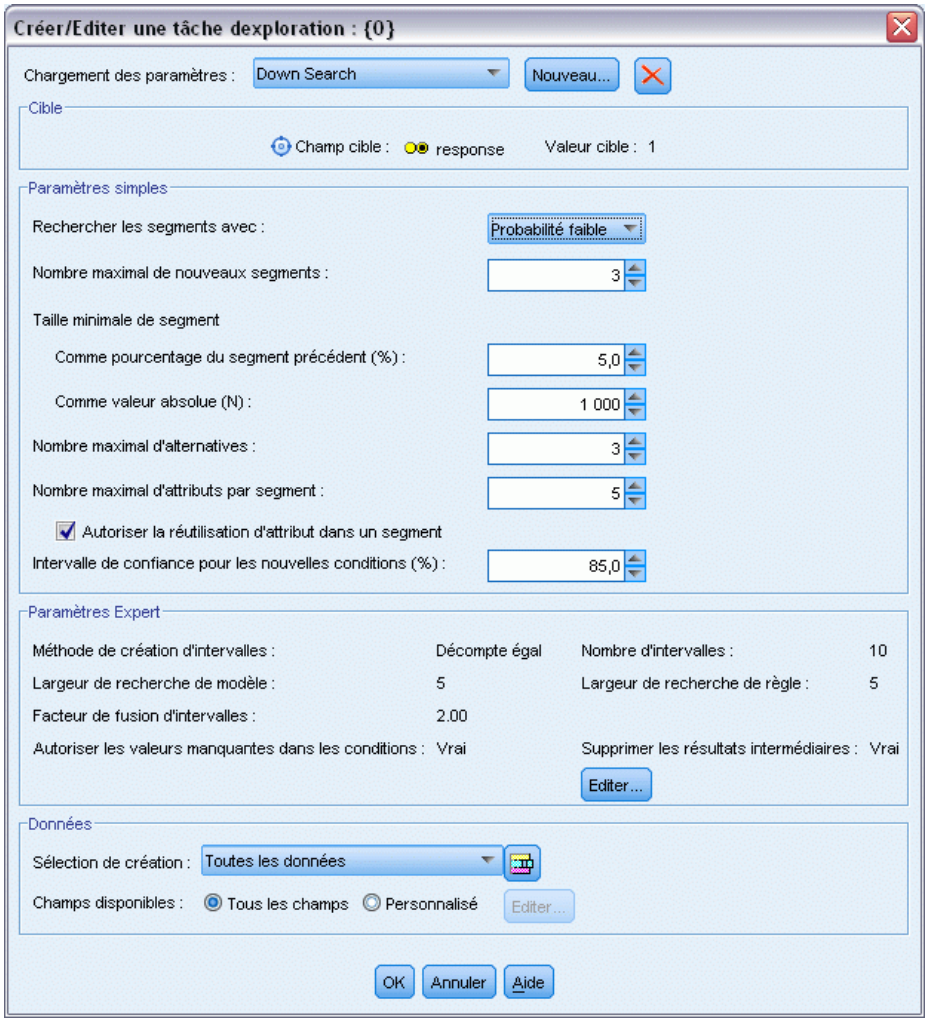
▶ Pour essayer une autre approche de modélisation, sélectionnez les options suivantes dans les menus :
Outils > Paramètres

Figure 11-12
Boîte de dialogue Créer/Editer la tâche d'exploration



- Cliquez sur le bouton Nouveau (dans le coin supérieur droit) pour créer une seconde tâche d'exploration et spécifiez *Down Search* comme nom de tâche dans la boîte de dialogue Nouveaux paramètres.

Figure 11-13
Boîte de dialogue Créer/Editer la tâche d'exploration

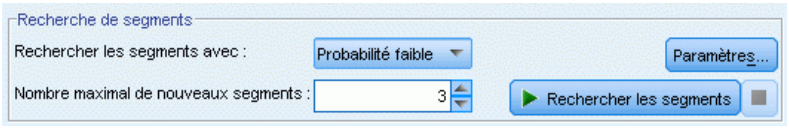


- Faites passer la direction de recherche pour la tâche à Faible probabilité. L'algorithme recherchera les segments avec les taux de réponse *les plus faibles* au lieu des plus élevés.

- Augmentez la taille minimale de segment à 1 000. Cliquez sur OK pour revenir à l'afficheur Liste interactive.

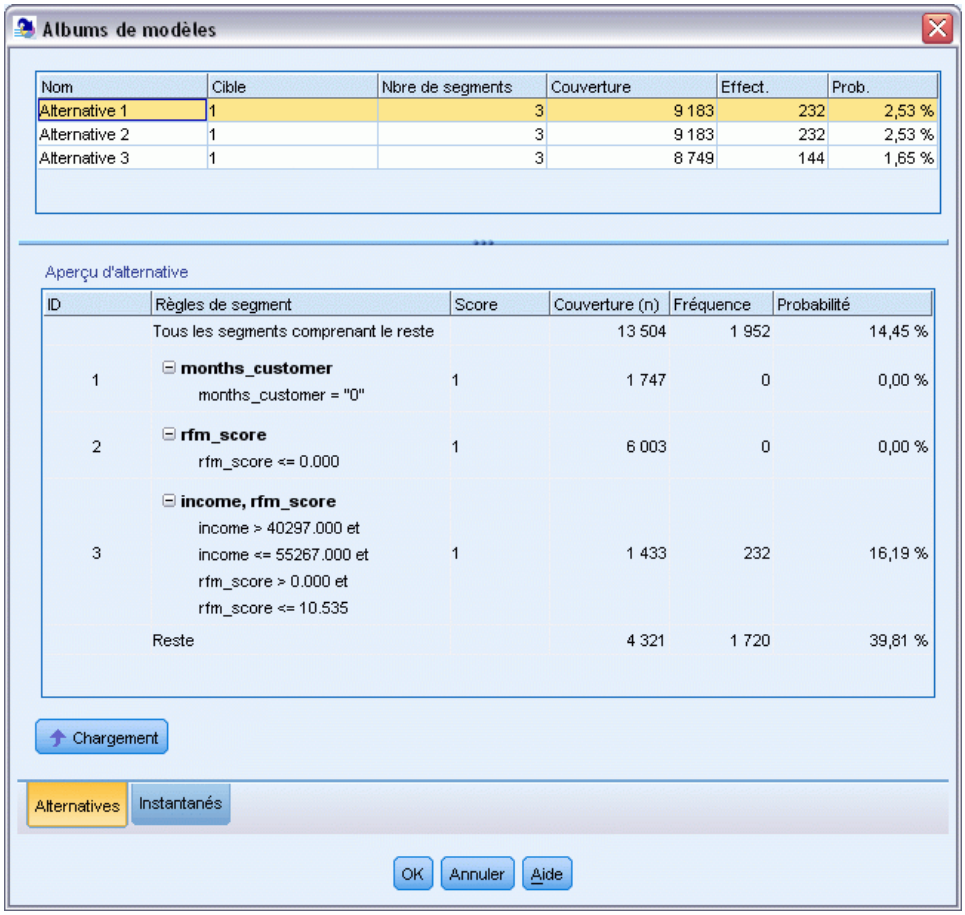
▶ Dans l'afficheur Liste interactive, vérifiez que le panneau *Localisateur de segment* affiche les détails de la nouvelle tâche et cliquez sur Rechercher les segments.

Figure 11-14
Rechercher les segments dans une nouvelle tâche d'exploration



La tâche renvoie un nouvel ensemble d'alternatives, qui est affiché dans l'onglet Alternatives de la boîte de dialogue Albums de modèles et que vous pouvez prévisualiser de la même manière que les résultats précédents.

Figure 11-15
Résultats du modèle obtenus par l'intermédiaire de la tâche Down Search



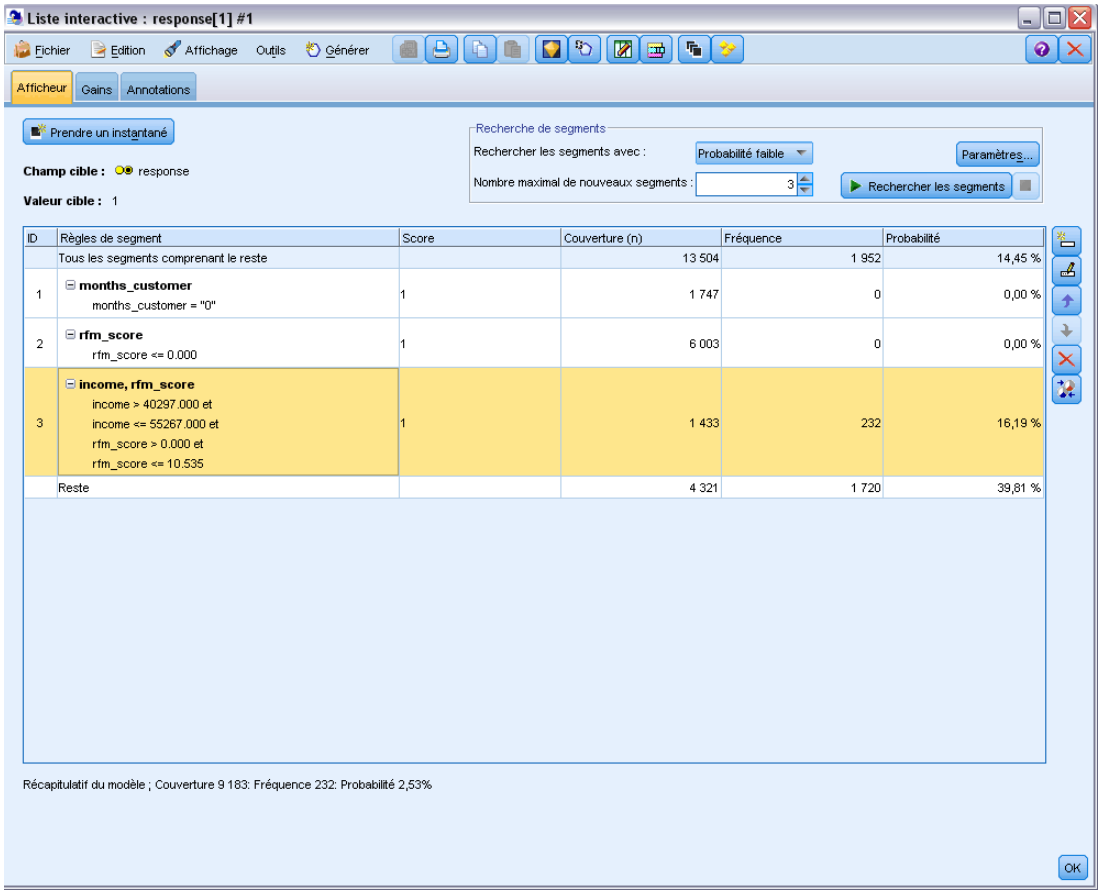
Cette fois, chaque modèle identifie les segments dotés de faibles probabilités de réponse au lieu de fortes probabilités. En examinant la première alternative, vous constatez que le simple fait d'exclure ces segments augmente le taux de correspondance du reste à 39,81%. Ce résultat est

inférieur à celui obtenu avec le modèle précédemment étudié, mais il présente une couverture supérieure (et donc un nombre total de correspondances plus élevé).

En combinant les deux approches (utilisation d'une recherche à faible probabilité pour éliminer les enregistrements inintéressants, suivie d'une recherche à forte probabilité), vous pouvez améliorer ce résultat.

- Cliquez sur Charger pour que la première alternative Down Search devienne le modèle de travail et cliquez sur OK pour fermer la boîte de dialogue Albums de modèles.

Figure 11-16
Exclusion d'un segment



- Cliquez à droite sur chacun des deux premiers segments et sélectionnez Exclure le segment. Ensemble, ces segments capturent presque 8 000 enregistrements avec zéro correspondance entre elles, il est donc souhaitable de les exclure des futures offres. (Les segments exclus auront un score nul pour le signaler.)

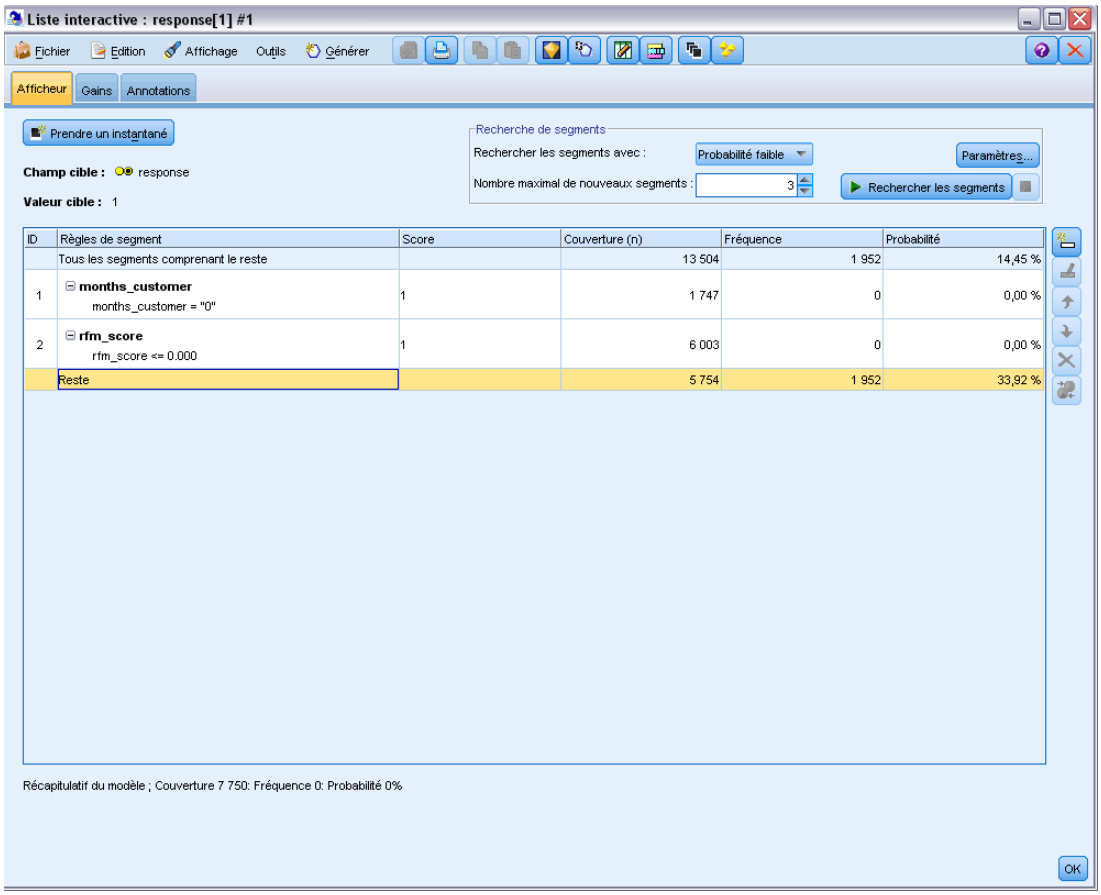
- Cliquez avec le bouton droit de la souris sur le troisième segment et sélectionnez Supprimer le segment. Le taux de correspondance de 16,19 % de ce segment n'est pas très différent du taux de référence de 14,45 %, et il n'ajoute donc pas assez d'informations pour justifier sa conservation.

Remarque : la suppression d'un segment et son exclusion sont deux opérations différentes. L'exclusion d'un segment modifie uniquement son score, alors que sa suppression le retire complètement du modèle.

Une fois que vous avez exclu les segments ayant les plus basses performances, vous pouvez rechercher les segments avec les plus hautes performances dans le reste.

- Dans la table, cliquez sur la ligne du reste pour la sélectionner de telle manière que la prochaine tâche exploratoire s'applique uniquement au reste.

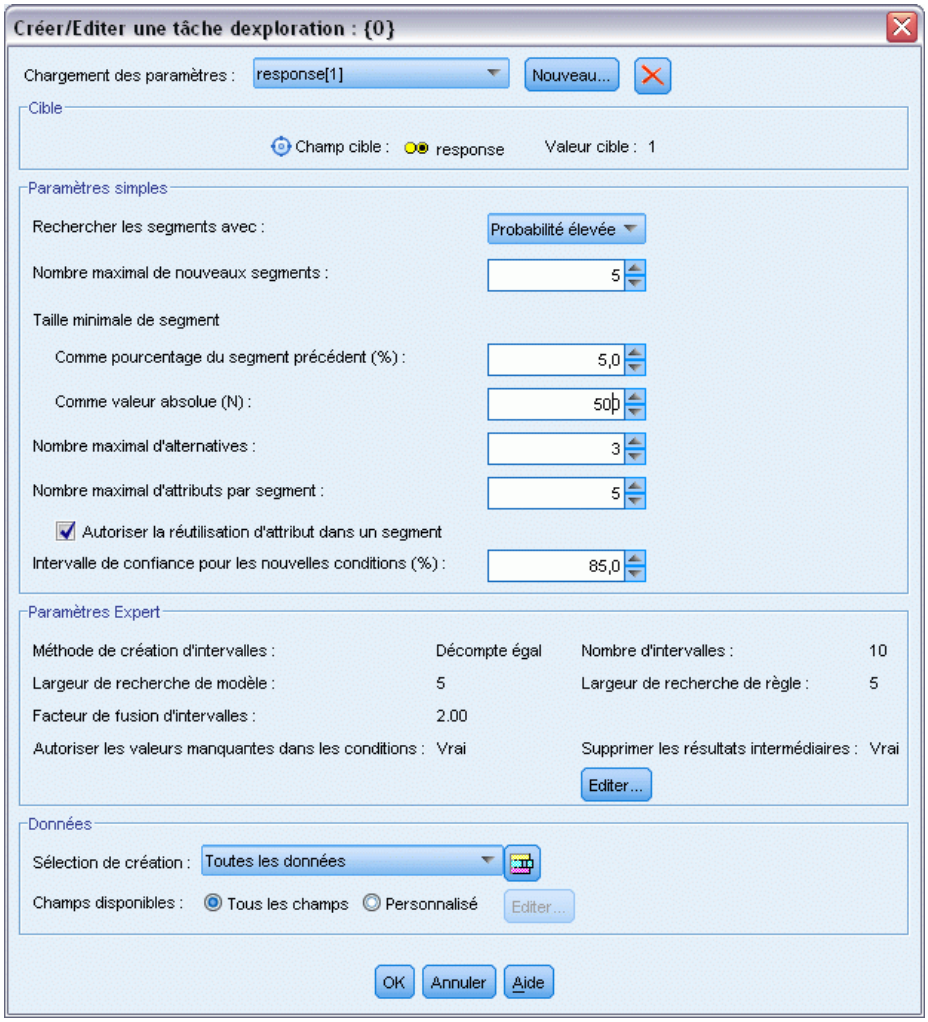
Figure 11-17
Sélection d'un segment



- Le reste étant sélectionné, cliquez sur Paramètres pour ouvrir à nouveau la boîte de dialogue Créer/Editer une tâche d'exploration.
- En haut, dans Charger les paramètres, sélectionnez la tâche d'exploration par défaut : réponse[1].
- Modifiez les Paramètres simples pour augmenter le nombre de nouveaux segments jusqu'à 5 et la taille minimale de segments à 500.

▶ Cliquez sur OK pour revenir à l'afficheur Liste interactive.

Figure 11-18
Sélection de la tâche d'exploration par défaut.

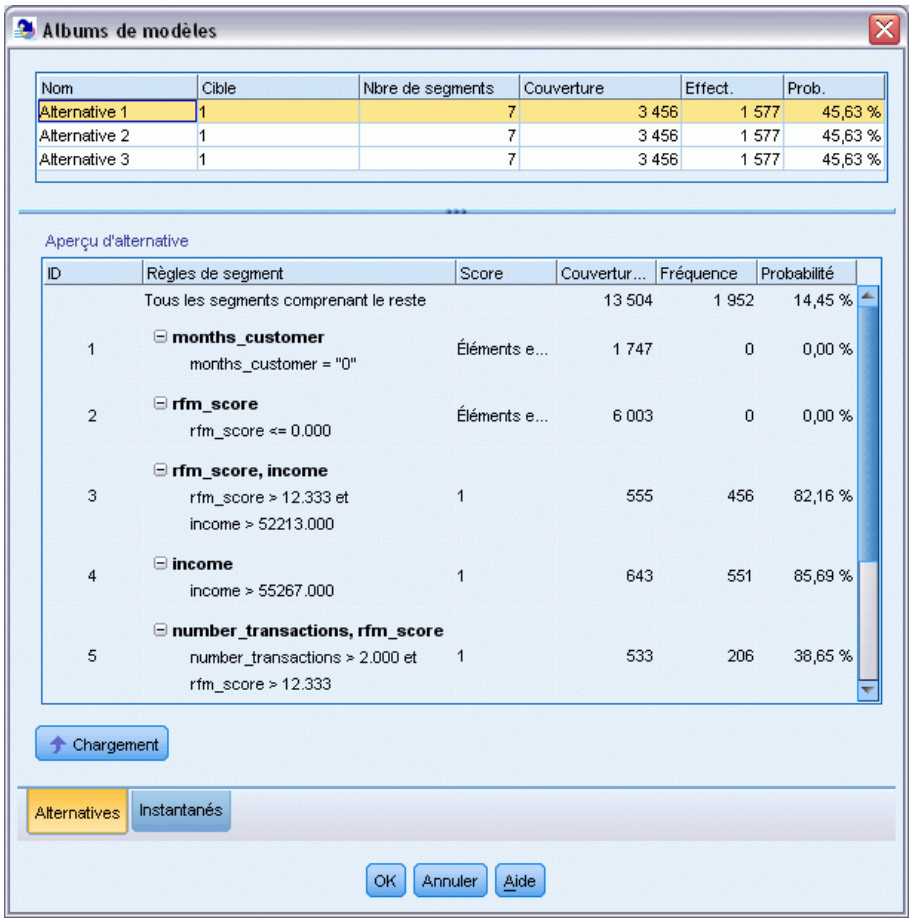


▶ Cliquez sur Rechercher les segments.

Cette action affiche encore un nouvel ensemble de modèles alternatifs. En insérant les résultats d'une tâche exploratoire dans une autre tâche, ces derniers modèles contiennent un mélange de segments très performants et de segments peu performants. Les segments dotés de taux de réponse faibles sont exclus, ce qui signifie que leur score est nul, alors que les segments inclus ont le score 1. Les statistiques globales reflètent ces exclusions, avec un taux de correspondance de

45,63 % pour le premier modèle alternatif et une couverture supérieure (1 577 correspondances sur 3 456 enregistrements) à celle de tous les modèles précédents.

Figure 11-19
Alternatives pour un modèle associé

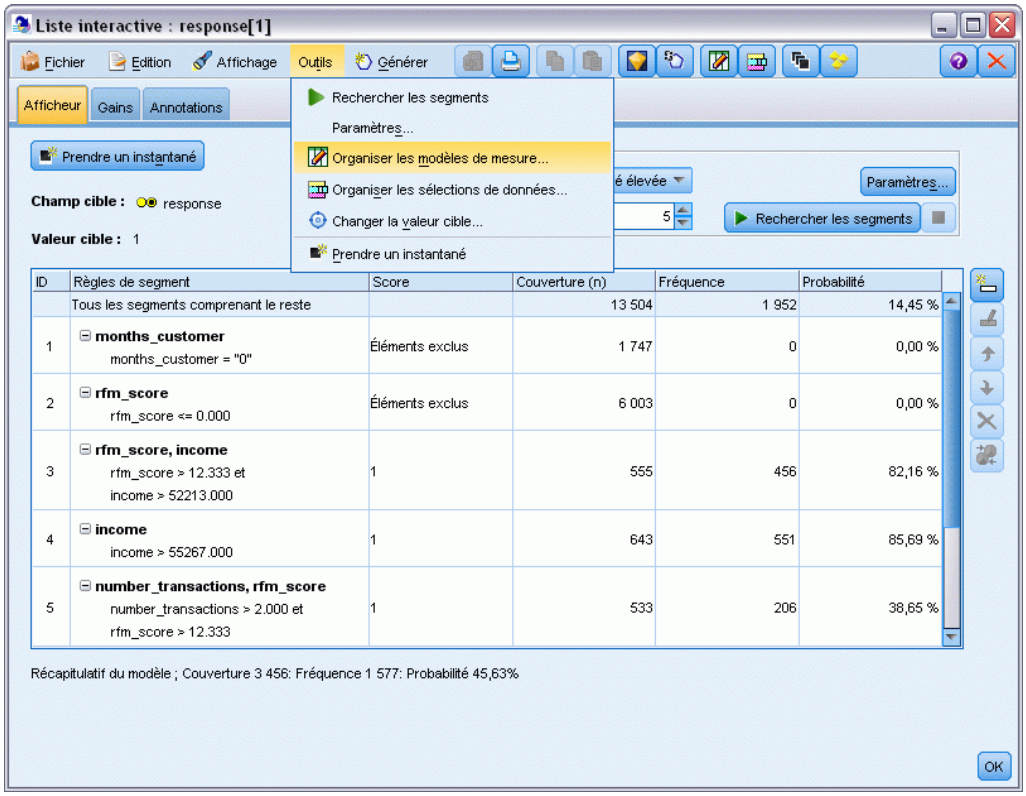


► Prévisualisez la première alternative et cliquez sur Charger pour en faire le modèle de travail.

Calcul des mesures personnalisées avec Excel

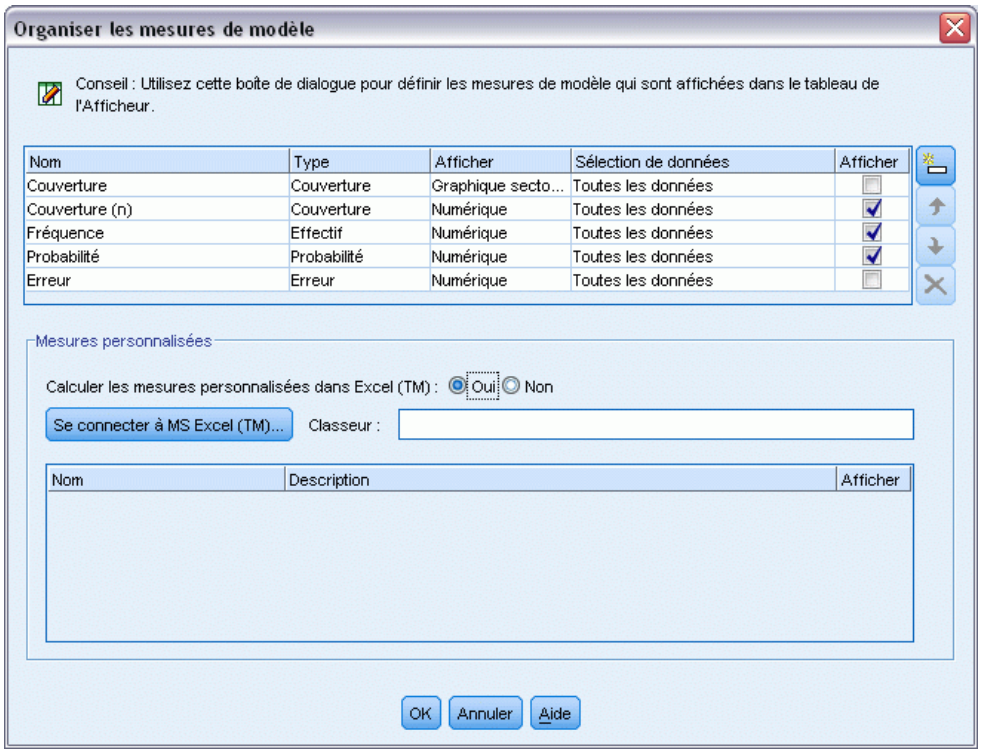
- Pour avoir une meilleure visibilité sur la façon dont le modèle fonctionne en termes pratiques, choisissez Organiser les mesures du modèle dans la barre d'outils.

Figure 11-20
Organisation des mesures de modèle



La boîte de dialogue Organiser les mesures du modèle vous permet de choisir les mesures (ou colonnes) à afficher dans l'Afficheur de liste interactif. Vous pouvez aussi indiquer si les mesures sont calculées sur tous les enregistrements ou sur un sous-ensemble sélectionné, et vous pouvez choisir d'afficher un graphique sectoriel plutôt qu'un nombre le cas échéant.

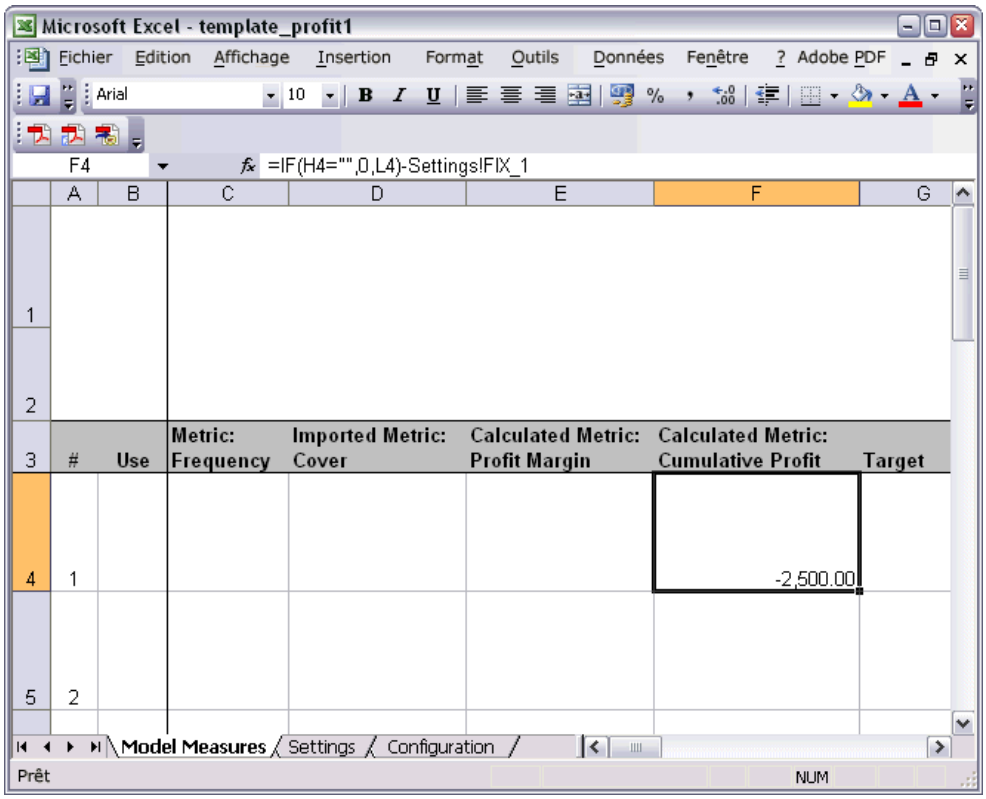
Figure 11-21
Boîte de dialogue Organiser les mesures du modèle



En outre, si Microsoft Excel est installé, vous pouvez lier un modèle Excel qui calculera les mesures personnalisées et les ajoutera à l'affichage interactif.

- Dans la boîte de dialogue Organiser les mesures du modèle, configurez Calculer les mesures personnalisées dans Excel (TM) sur Oui.
- Cliquez sur Se connecter à MS Excel (TM)
- Sélectionnez le classeur *template_profit.xlt*, situé dans le répertoire des *flux* dans le dossier *Demos* de votre installation IBM® SPSS® Modeler et cliquez sur Ouvrir pour lancer la feuille de calcul.

Figure 11-22
Feuille de calcul Excel de modèles de mesures



Le modèle Excel contient trois feuilles de calcul :

- Mesures du modèle affiche les mesures du modèle importées du modèle et calcule les mesures personnalisées pour les réexporter vers le modèle.
- Paramètres contient les paramètres à utiliser dans le calcul des mesures personnalisées.
- Configuration définit les mesures à importer du modèle et à exporter vers ce modèle.

Les mesures réexportées vers le modèle sont :

- **Marge de profit.** Revenus nets du segment
- **Profit cumulé.** Total des profits de la campagne

Définis par les formules suivantes :

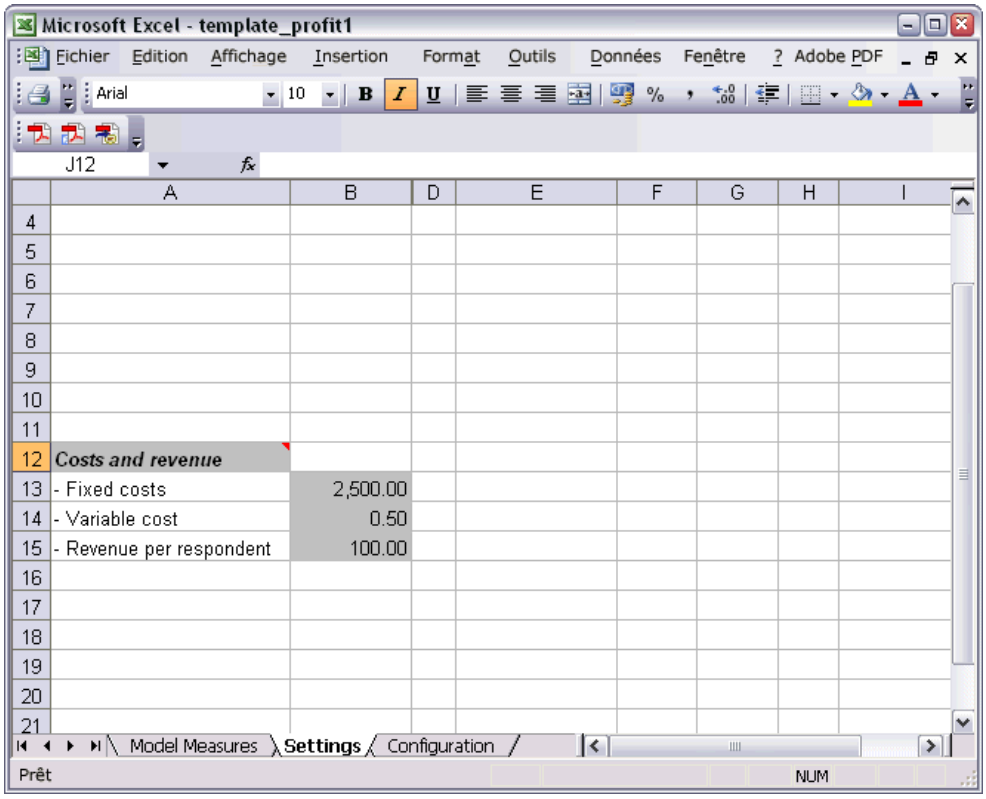
Profit Margin = Frequency * Revenue per respondent - Cover * Variable cost

Cumulative Profit = Total Profit Margin - Fixed cost

Notez que l'effectif et la couverture sont importés du modèle.

Les paramètres de coût et de revenus sont indiqués par l'utilisateur dans la feuille de calcul Paramètres.

Figure 11-23
Feuille de calcul Excel Paramètres



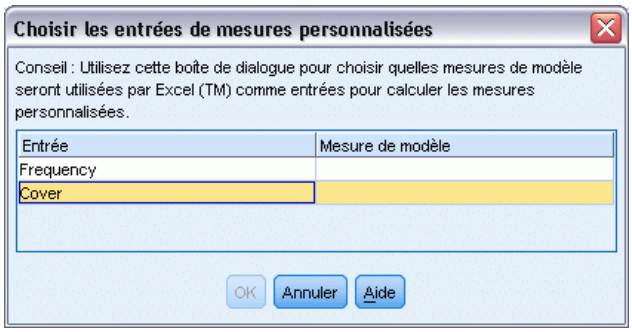
Coût fixe est le coût configuré pour la campagne ; par exemple, conception et planification.

Coût variable est le coût d'extension de l'offre à chaque client, par exemple les enveloppes et les timbres.

Recettes par personne sondée est le revenu net d'un client qui répond à l'offre.

▶ Pour terminer la liaison de retour vers le modèle, utilisez la barre des tâches Windows (ou appuyez sur Alt+Tab) pour revenir à l'afficheur Liste interactive.

Figure 11-24
Choix des entrées de mesures personnalisées



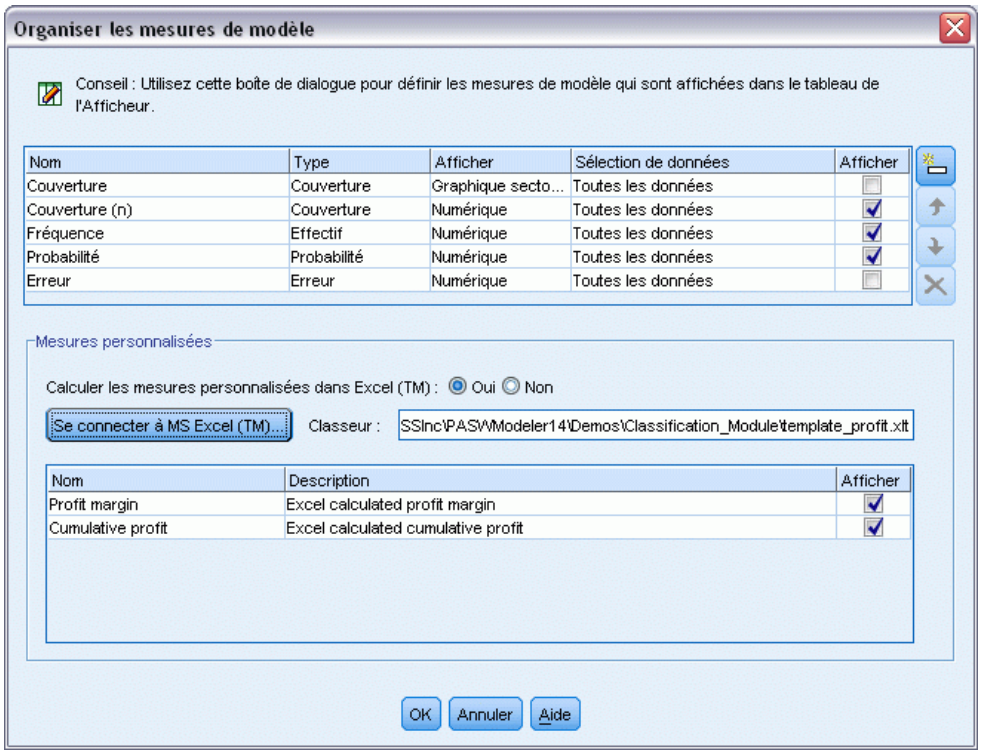
La boîte de dialogue Choisir les entrées de mesures personnalisées s'affiche, vous permettant de faire correspondre les entrées du modèle aux paramètres spécifiques définis dans le modèle. La colonne de gauche répertorie les mesures disponibles et la colonne de droite les fait correspondre aux paramètres de la feuille de calcul définis dans la feuille de calcul Configuration.

▶ Dans la colonne Mesures de modèle, sélectionnez Effectif et Couverture (n) pour les entrées respectives puis cliquez sur OK.

Dans ce cas, les noms du paramètre du modèle (Effectif et Couverture (n)) correspondent aux entrées, mais des noms différents peuvent aussi être utilisés.

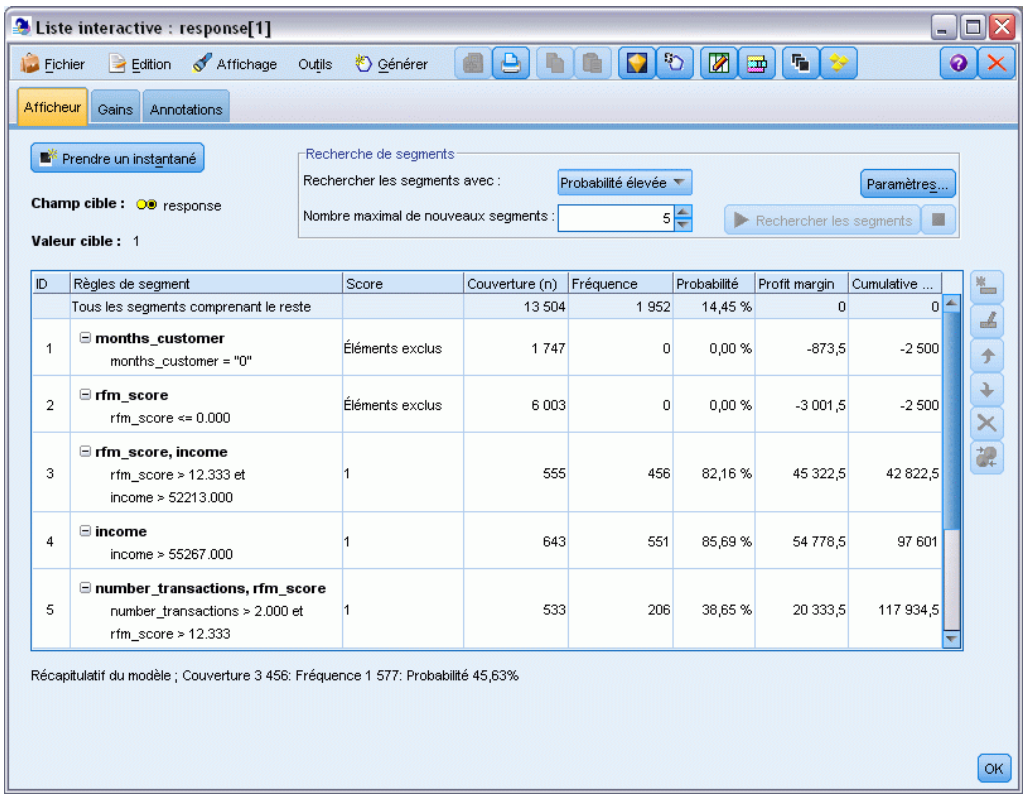
- Cliquez sur OK dans la boîte de dialogue Organiser les mesures du modèle pour mettre à jour l'afficheur de la liste interactive.

Figure 11-25
Boîte de dialogue Organiser les mesures du modèle avec les mesures personnalisées d'Excel



Les nouvelles mesures sont maintenant ajoutées en tant que nouvelles colonnes dans la fenêtre et seront recalculées chaque fois que le modèle sera mis à jour.

Figure 11-26
Mesures personnalisées d'Excel affichées dans l'afficheur Liste interactive



En éditant le modèle Excel, vous pouvez créer autant de mesures personnalisées que vous le souhaitez.

Modification du modèle Excel

Bien que IBM® SPSS® Modeler propose un modèle Excel par défaut à utiliser avec l'afficheur Liste de décisions, il est possible de modifier les paramètres ou d'ajouter les votres. Par exemple, les coûts dans le modèle peuvent ne pas correspondre à ceux de votre entreprise et doivent être modifiés.

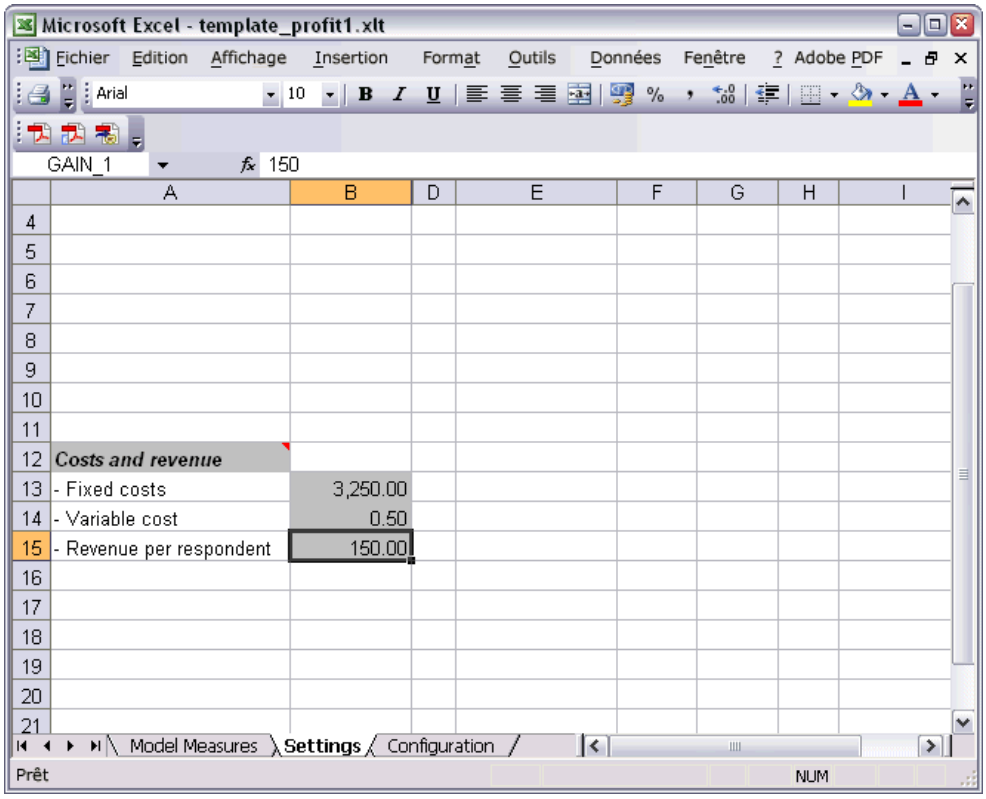
Remarque : Si vous modifiez un modèle existant, ou que vous créez le vôtre, n'oubliez pas de sauvegarder le fichier avec un suffixe *.xlt* d'Excel 2003.

Pour modifier le modèle par défaut et y ajouter de nouveaux coûts et informations sur les revenus et mettre à jour l'afficheur Liste interactive en y ajoutant de nouveaux chiffres :

- Dans l'afficheur Liste interactive, sélectionnez Organiser les mesures de modèle dans le menu Outils.
- Dans la boîte de dialogue Organiser les mesures du modèle, cliquez sur Connecter à Excel™.

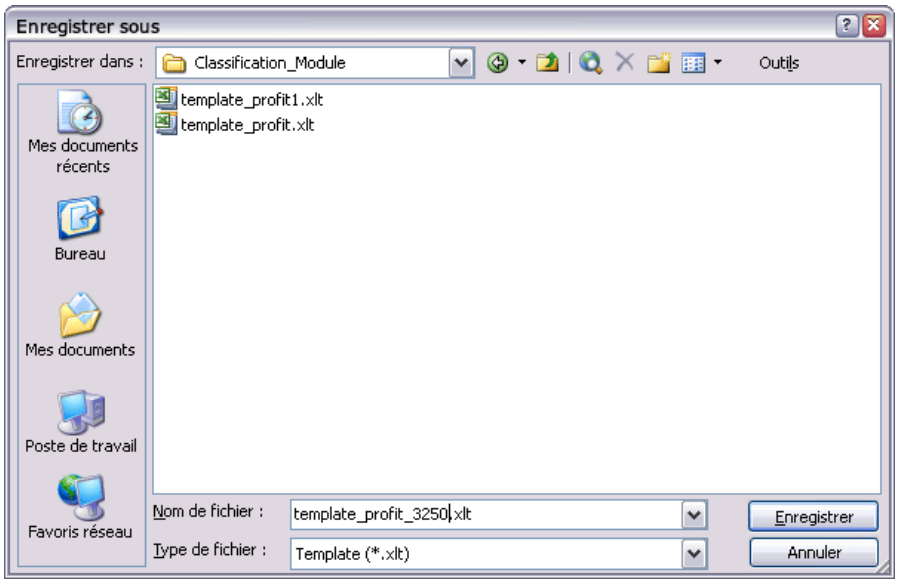
- Sélectionnez le classeur *template_profit.xlt* et cliquez sur Ouvrir pour lancer la feuille de calcul.
- Sélectionnez la feuille de calcul Paramètres.
- Modifiez les coûts fixes sur 3250,00 et le Revenu par personne interrogée sur 150,00.

Figure 11-27
Valeurs modifiées sur la feuille de calcul Excel Paramètres



- Sauvegardez le modèle modifié en utilisant un nom de fichier unique et approprié. Vérifiez qu'il possède une extension *.xlt* d'Excel 2003.

Figure 11-28
Enregistrement d'un modèle Excel modifié



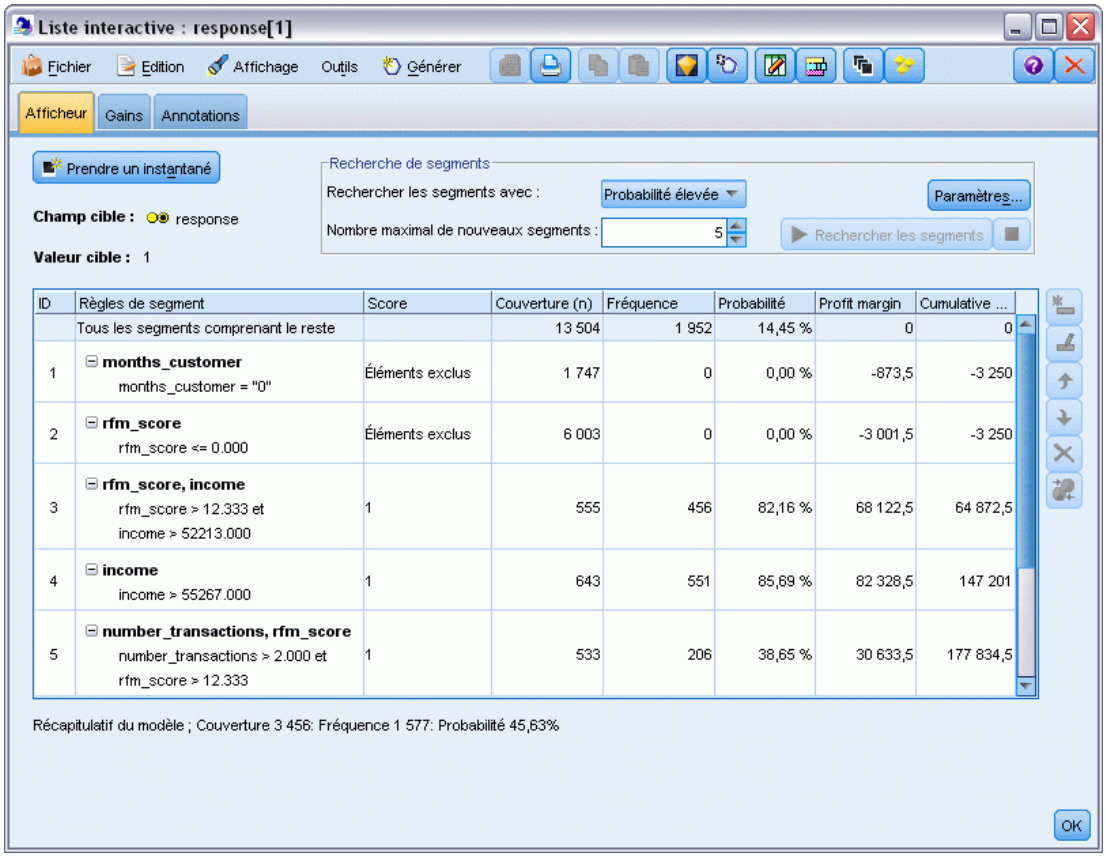
- Utilisez la barre de tâches de Windows (ou appuyez sur Alt+Tab) pour retourner à l'afficheur Liste interactive.

  Dans la boîte de dialogue Choisir les entrées de mesures personnalisées, sélectionnez les mesures à afficher et cliquez sur OK.

- Cliquez sur OK dans la boîte de dialogue Organiser les mesures du modèle pour mettre à jour l'afficheur de la liste interactive.

Bien sûr, cet exemple ne présente qu'une seule façon de modifier le modèle Excel. Vous pouvez effectuer d'autres modifications qui extraient des données de l'afficheur Liste Interactive ou qui lui transmettent des données, ou travailler depuis Excel pour produire d'autres entrées, tels que des graphiques.

Figure 11-29
Mesures personnalisées d'Excel modifiées affichées dans l'afficheur Liste interactive



Enregistrement des résultats

Pour enregistrer un modèle afin de pouvoir l'utiliser ultérieurement au cours de la session interactive, vous pouvez prendre un instantané du modèle, qui apparaîtra dans l'onglet Instantanés. Vous pouvez accéder aux instantanés enregistrés à tout moment au cours de la session interactive.

Ainsi, vous pouvez tester d'autres tâches d'exploration pour rechercher des segments supplémentaires. Vous pouvez également éditer des segments existants, insérer des segments personnalisés sur la base de vos propres règles commerciales, créer des sélections de données pour optimiser le modèle pour des groupes précis et personnaliser le modèle de différentes manières. Enfin, vous pouvez inclure ou exclure explicitement chaque segment, selon vos besoins, pour préciser comment chacun d'eux sera évalué.

Lorsque vous êtes satisfait des résultats, utilisez le menu Générer pour générer un modèle qui peut être ajouté aux flux ou déployé à des fins de scoring.

Une autre solution pour enregistrer l'état actuel de la session interactive et y revenir un autre jour consiste à choisir Mettre à jour le noeud de modélisation dans le menu Fichier. Ainsi, le noeud de modélisation Liste de décision sera mis à jour avec les paramètres en cours, y compris les tâches d'exploration, les instantanés de modèle, les sélections de données et les mesures personnalisées. Lorsque vous réexécutez le flux, vérifiez que l'option Utiliser les informations de session interactive enregistrées est sélectionnée dans le noeud de modélisation Liste de décision pour restaurer l'état actuel de la session. Pour plus d'informations, reportez-vous à la section Liste de décision dans le chapitre 9 dans *Noeuds de modélisation de IBM SPSS Modeler 15*.

Chapitre
12

Classification des clients de télécommunications (régression logistique multinomiale)

La régression logistique est une technique statistique de classification des enregistrements sur la base des valeurs des champs d'entrée. Excepté le fait qu'elle utilise un champ cible catégoriel et non pas numérique, cette régression est semblable à la régression linéaire.

Par exemple, supposons qu'un fournisseur de télécommunications ait segmenté sa base de clientèle par modèles d'utilisation de service, classant ses clients en quatre groupes. Si les données démographiques peuvent être utilisées pour prévoir les groupes d'affectation, vous pouvez personnaliser les offres pour chaque client éventuel.

Cet exemple utilise le flux *telco_custcat.str*, qui fait référence au fichier de données *telco.sav*. Ces fichiers sont disponibles dans le répertoire *Demos* de n'importe quelle installation de IBM® SPSS® Modeler. Ce dossier est accessible à partir du groupe de programmes IBM® SPSS® Modeler dans le menu Démarrer de Windows. Le fichier *telco_custcat.str* se trouve dans le répertoire des *flux* .

Cet exemple est axé sur l'utilisation des données démographiques dans le but de prévoir des modèles d'utilisation. Le champ cible *custcat* possède quatre valeurs possibles qui correspondent aux quatre groupes de clients suivants :

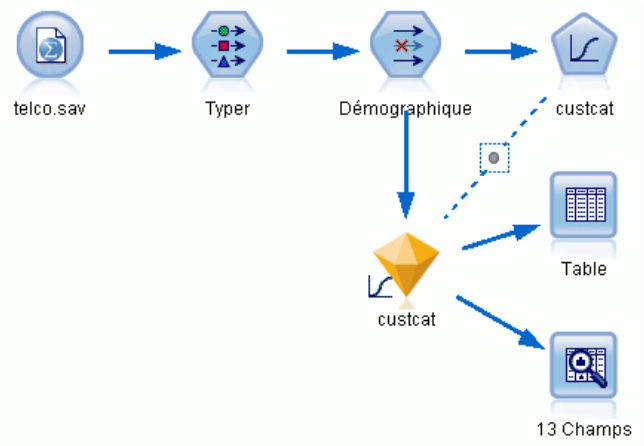
Valeur	Etiquette
1	Basic Service
2	E-Service
3	Plus Service
4	Total Service

Comme le champ cible contient plusieurs catégories, un modèle multinomial est utilisé. Dans le cas d'un champ cible comprenant deux catégories distinctes, telles que oui/non, vrai/faux ou attrition/absence d'attrition, un modèle binomial peut être créé. Pour plus d'informations, reportez-vous à la section Attrition dans le domaine des télécommunications (régression logistique binomiale) dans le chapitre 13 sur p. 163.

Création du flux

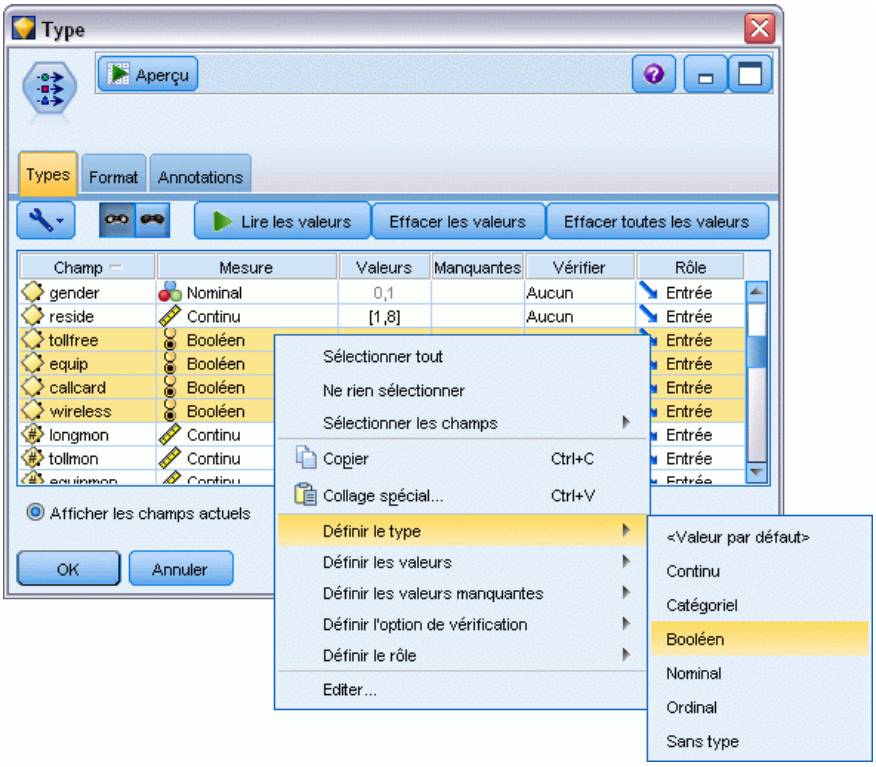
- Ajoutez un noeud source Fichier de statistiques pointant vers *telco.sav* dans le dossier *Demos*.

Figure 12-1
Exemple de flux permettant de classifier les clients par régression logistique multinomiale



- Ajoutez un noeud Typer et cliquez sur Lire les valeurs, en vous assurant que tous les niveaux de mesure sont correctement paramétrés. Par exemple, la majorité des champs avec des valeurs 0 et 1 peuvent être considérés comme des champs booléens.

Figure 12-2
Configuration du niveau de mesure pour plusieurs champs



Astuce : Pour modifier les propriétés de plusieurs champs contenant des valeurs similaires (telles que 0/1), cliquez sur l'en-tête de colonne *Valeurs* afin de trier les champs en fonction de cette valeur. Maintenez la touche Maj enfoncée tout en utilisant la souris ou les touches fléchées pour sélectionner tous les champs à modifier. Cliquez ensuite sur la sélection avec le bouton droit de la souris pour modifier le niveau de mesure ou les autres attributs des champs sélectionnés.

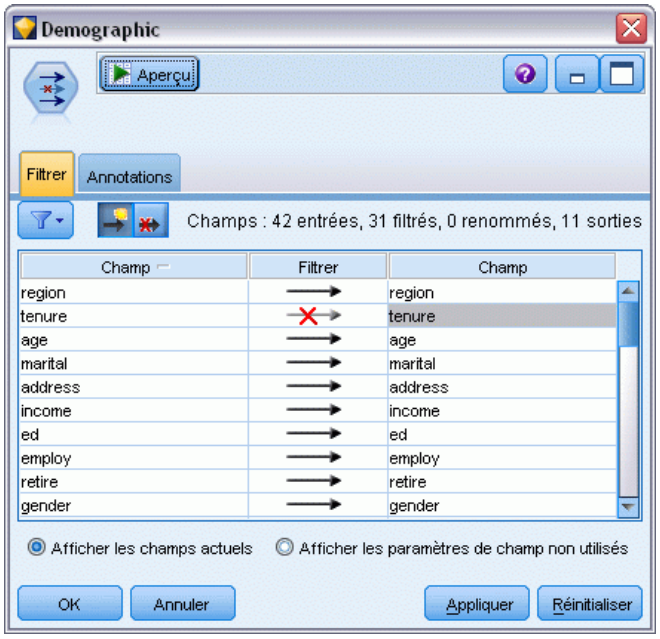
Veuillez noter que puisqu'il est plus correct de considérer le *sexe* comme un champ avec un ensemble de deux valeurs plutôt que comme un champ booléen, laissez sa valeur de mesure sur Nominal.

- Définissez le rôle du champ *custcat* sur Cible. Le rôle de tous les autres champs doit être défini sur Entrée.

Figure 12-3
Définition du rôle de champ

Cet exemple étant axé sur les données démographiques, utilisez un noeud Filtrer pour n'inclure que les champs pertinents (*region*, *age*, *marital*, *address*, *income*, *ed*, *employ*, *retire*, *gender*, *reside* et *custcat*). Les autres champs peuvent être exclus pour cette analyse.

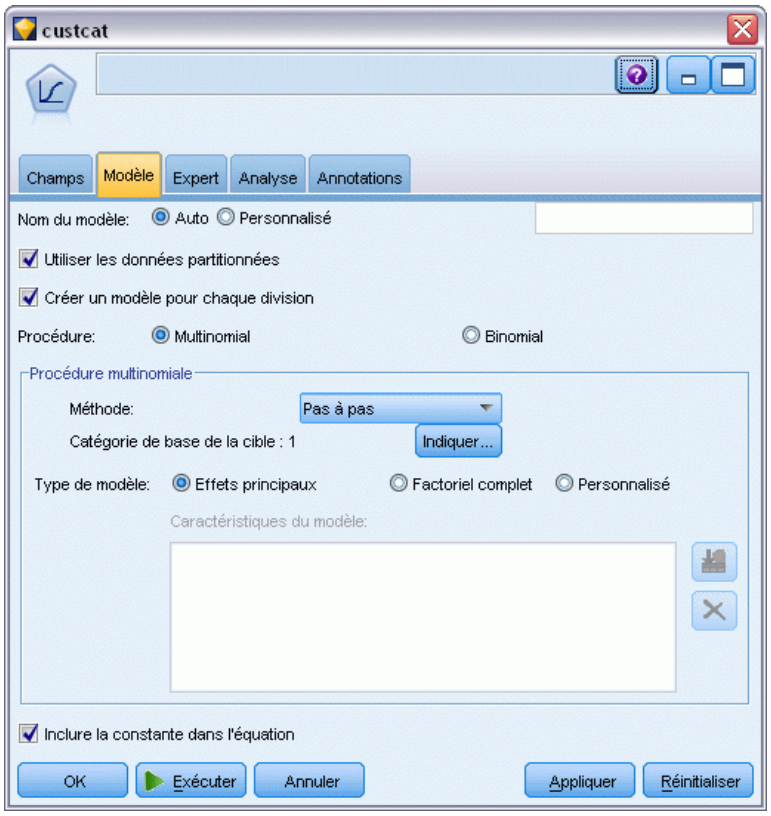
Figure 12-4
Filtrage des champs démographiques



(Vous pouvez également paramétrer le rôle sur Aucun pour ces champs plutôt que de les exclure, ou sélectionner les champs que vous souhaitez utiliser dans le noeud de modélisation.)

▶ Dans le noeud Logistique, cliquez sur l'onglet Modèle et sélectionnez la méthode Pas à pas. Sélectionnez Multinomial, Effets principaux et Inclure la constante dans l'équation.

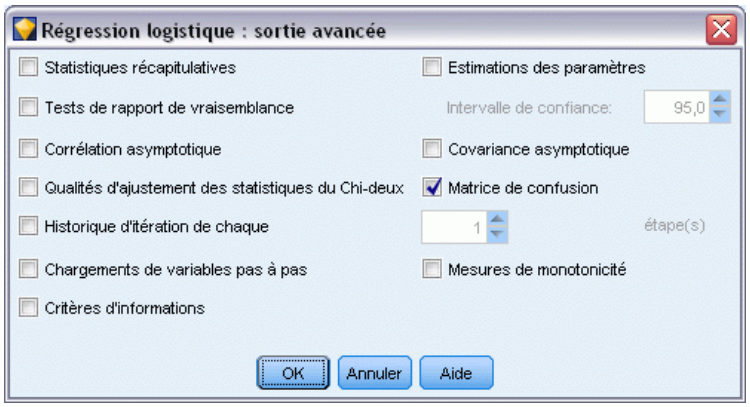
Figure 12-5
Choix des options de modèle



Laissez la catégorie de base de la cible définie sur 1. Le modèle comparera les autres clients à ceux qui sont abonnés au Basic Service.

▶ Dans l'onglet Expert, sélectionnez le mode Expert, puis Sortie et, dans la boîte de dialogue Sorties avancées, sélectionnez Matrice de confusion.

Figure 12-6
Choix des options de sortie

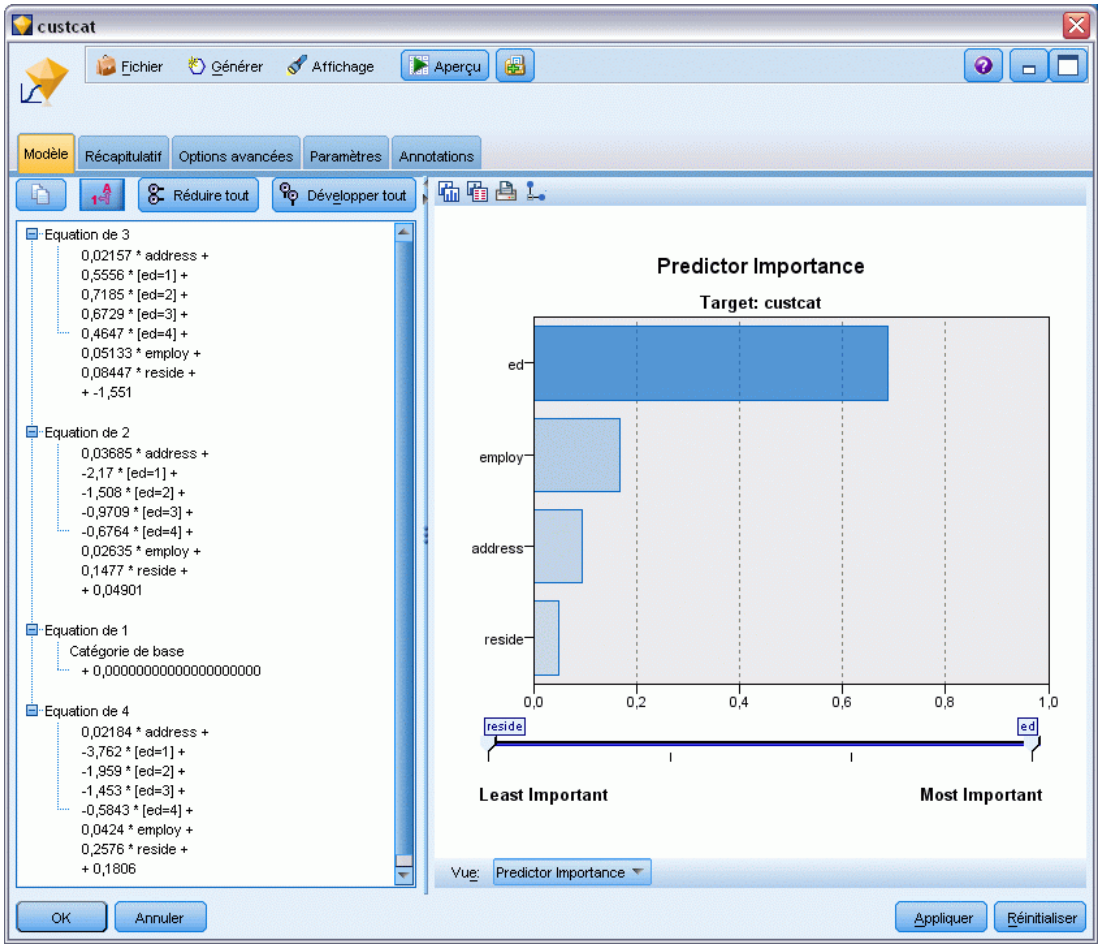


Navigation dans le modèle

▶ Exécutez le noeud pour générer le modèle, qui est ajouté à la palette Modèles dans l'angle supérieur droit. Pour afficher ses détails, cliquez avec le bouton droit de la souris sur le noeud du modèle généré et sélectionnez Parcourir.

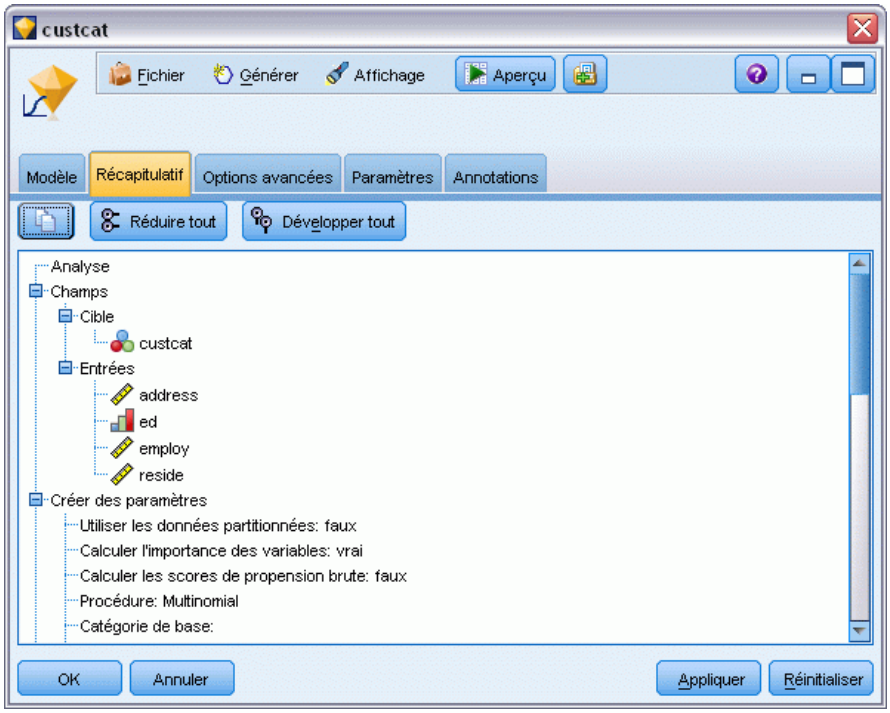
L'onglet Modèle affiche les équations utilisées pour affecter les enregistrements à chaque catégorie du champ cible. Il existe quatre catégories possibles, l'une d'elles est la catégorie de base pour laquelle aucun détail d'équation ne s'affiche. Les détails sont affichés pour les trois équations restantes, où la catégorie 3 représente le Plus Service et ainsi de suite.

Figure 12-7
Navigation dans les résultats du modèle



L'onglet Récapitulatif affiche (entre autres) la cible et les entrées (champs variables indépendantes) utilisées par le modèle. Ces champs sont ceux qui ont été réellement choisis sur la base de la méthode Pas à pas, et non la liste complète soumise.

Figure 12-8
Récapitulatif du modèle avec champs cible et champs d'entrée

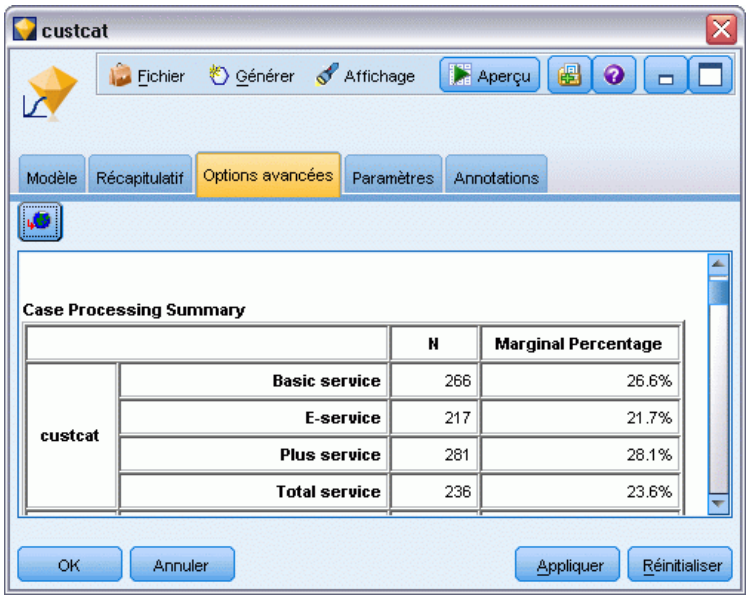


Les éléments affichés dans l'onglet Options avancées dépendent des options sélectionnées dans la boîte de dialogue Sorties avancées, dans le noeud de modélisation.

L'élément Récapitulatif du traitement des observations est systématiquement affiché. Il indique le pourcentage d'enregistrements inclus dans chaque catégorie du champ cible. Vous pouvez ainsi utiliser un modèle nul servant de base à la comparaison.

Sans créer de modèle qui utilise des variables indépendantes, votre meilleure prévision consiste à affecter tous les clients au groupe le plus commun, le groupe du Plus Service.

Figure 12-9
Récapitulatif du traitement des observations

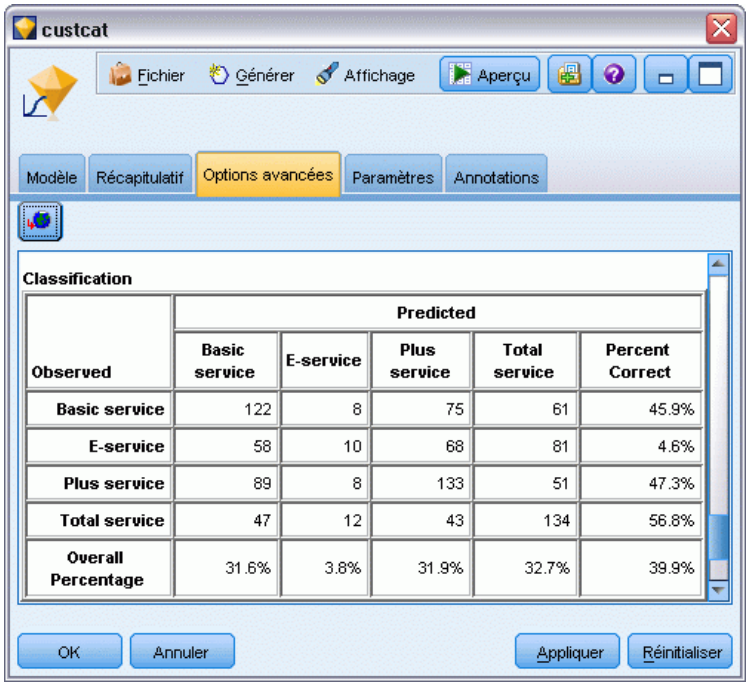


En fonction des données d'apprentissage, si vous avez affecté tous les clients au modèle nul, votre prévision est correcte 281/1000 = 28,1 % du temps. L'onglet Options avancées contient des informations supplémentaires qui vous permettent d'examiner les prévisions du modèle. Vous pouvez ensuite comparer les prévisions aux résultats du modèle nul pour voir le fonctionnement de votre modèle avec vos données.

En bas de l'onglet Options avancées, la table de classification supervisée affiche les résultats de votre modèle, qui est correct 39,9 % du temps.

Votre modèle est particulièrement performant à l'heure d'identifier les clients Total Service (catégorie 4), mais fonctionne très mal pour l'identification des clients E-service (catégorie 2). Si vous souhaitez une meilleure précision pour les clients de la catégorie 2, vous devez trouver une autre variable indépendante pour les identifier.

Figure 12-10
Tableau de classement



En fonction de ce que vous souhaitez prévoir, le modèle peut s'avérer parfaitement adapté à vos besoins. Par exemple, si l'identification des clients de la catégorie 2 ne vous intéresse pas, le modèle peut être assez précis pour vous. Cela peut être le cas lorsque le E-service est un produit d'appel qui ne génère que peu de bénéfices.

Si, par exemple, votre plus grand retour sur investissement provient des clients des catégories 3 ou 4, il est possible que le modèle vous fournisse les informations nécessaires.

Pour évaluer le niveau d'adéquation du modèle aux données, divers diagnostics sont disponibles dans la boîte de dialogue Sorties avancées lorsque vous créez le modèle. Pour plus d'informations, reportez-vous à la section Nugget de modèle Logistique - Sorties avancées dans le chapitre 10 dans *Noeuds de modélisation de IBM SPSS Modeler 15*. Des explications sur les fondements mathématiques des méthodes de modélisation utilisées dans IBM® SPSS® Modeler sont présentées dans le *Guide des algorithmes SPSS Modeler*, disponible dans le répertoire *\Documentation* du disque d'installation.

Sachez également que ces résultats sont établis uniquement d'après les données d'apprentissage. Pour évaluer comment le modèle peut se généraliser à d'autres données dans le monde réel, vous pouvez utiliser un noeud Partitionner destiné à contenir un sous-ensemble d'enregistrements à des fins de test et de validation. Pour plus d'informations, reportez-vous à la section Noeud Partitionner dans le chapitre 4 dans *Noeuds source, exécution et de sortie de IBM SPSS Modeler 15*.

Chapitre
13

Attrition dans le domaine des télécommunications (régression logistique binomiale)

La régression logistique est une technique statistique de classification des enregistrements sur la base des valeurs des champs d'entrée. Excepté le fait qu'elle utilise un champ cible catégoriel et non pas numérique, cette régression est semblable à la régression linéaire.

Cet exemple utilise le flux *telco_churn.str*, qui fait référence au fichier de données *telco.sav*. Ces fichiers sont disponibles dans le répertoire *Demos* de n'importe quelle installation de IBM® SPSS® Modeler. Ce dossier est accessible à partir du groupe de programmes IBM® SPSS® Modeler dans le menu Démarrer de Windows. Le fichier *telco_churn.str* se trouve dans le répertoire des *flux* .

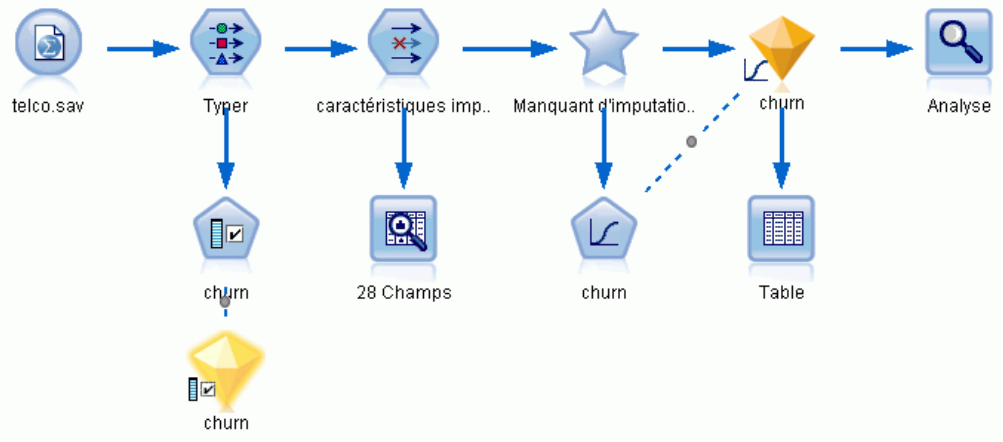
Par exemple, supposons qu'un fournisseur de télécommunications souhaite connaître le nombre de clients qui partent à la concurrence. Si les données d'utilisation du service permettent de prédire les clients susceptibles de passer à un autre fournisseur, les offres peuvent être personnalisées afin de retenir autant de clients que possible.

Cet exemple explique comment se servir des données d'utilisation pour prédire la perte de clients (attrition). Etant donné que la cible présente deux catégories distinctes, un modèle binomial est utilisé. Si la cible présente plus de deux catégories, un modèle multinomial peut être créé à la place. Pour plus d'informations, reportez-vous à la section Classification des clients de télécommunications (régression logistique multinomiale) dans le chapitre 12 sur p. 153.

Création du flux

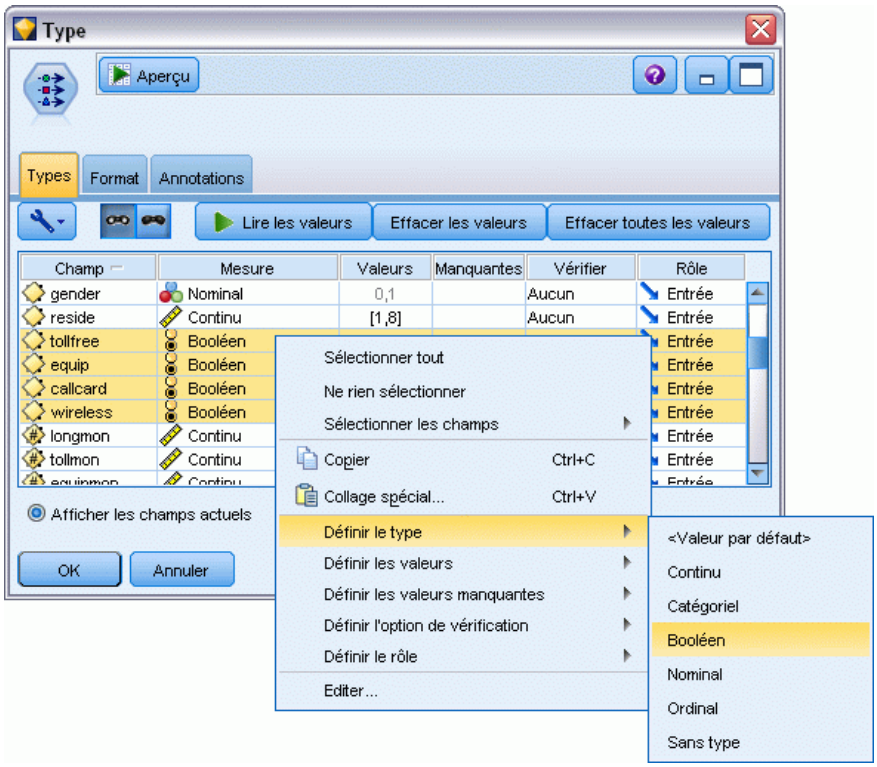
- Ajoutez un noeud source Fichier de statistiques pointant vers *telco.sav* dans le dossier *Demos*.

Figure 13-1
Exemple de flux permettant de classifier les clients par régression logistique binomiale



- Ajoutez un noeud Typer pour définir des champs, en vous assurant que tous les niveaux de mesure sont correctement paramétrés. Par exemple, la plupart des champs dont les valeurs sont 0 et 1 peuvent être considérés comme des champs booléens. Cependant, certains champs, tels que celui indiquant le genre, doivent être considérés comme des champs nominaux à deux valeurs.

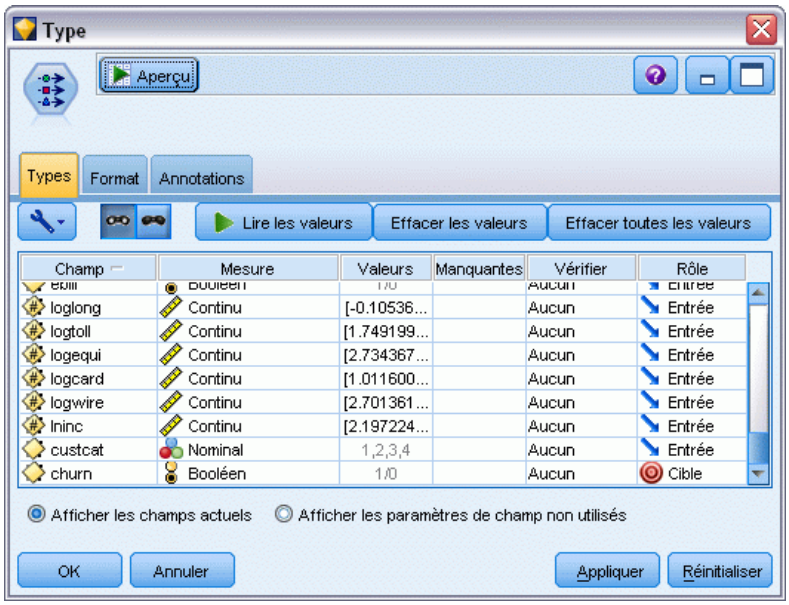
Figure 13-2
Configuration du niveau de mesure pour plusieurs champs



Astuce : Pour modifier les propriétés de plusieurs champs contenant des valeurs similaires (telles que 0/1), cliquez sur l'en-tête de colonne *Valeurs* afin de trier les champs par valeur. Maintenez la touche Maj enfoncée tout en utilisant la souris ou les touches fléchées pour sélectionner tous les champs à modifier. Cliquez ensuite sur la sélection avec le bouton droit de la souris pour modifier le niveau de mesure ou les autres attributs des champs sélectionnés.

▶ Définissez le niveau de mesure pour le champ *attrition* sur Booléen, puis définissez le rôle sur Cible. Le rôle de tous les autres champs doit être défini sur Entrée.

Figure 13-3
Configuration du niveau de mesure et du rôle pour le champ attrition



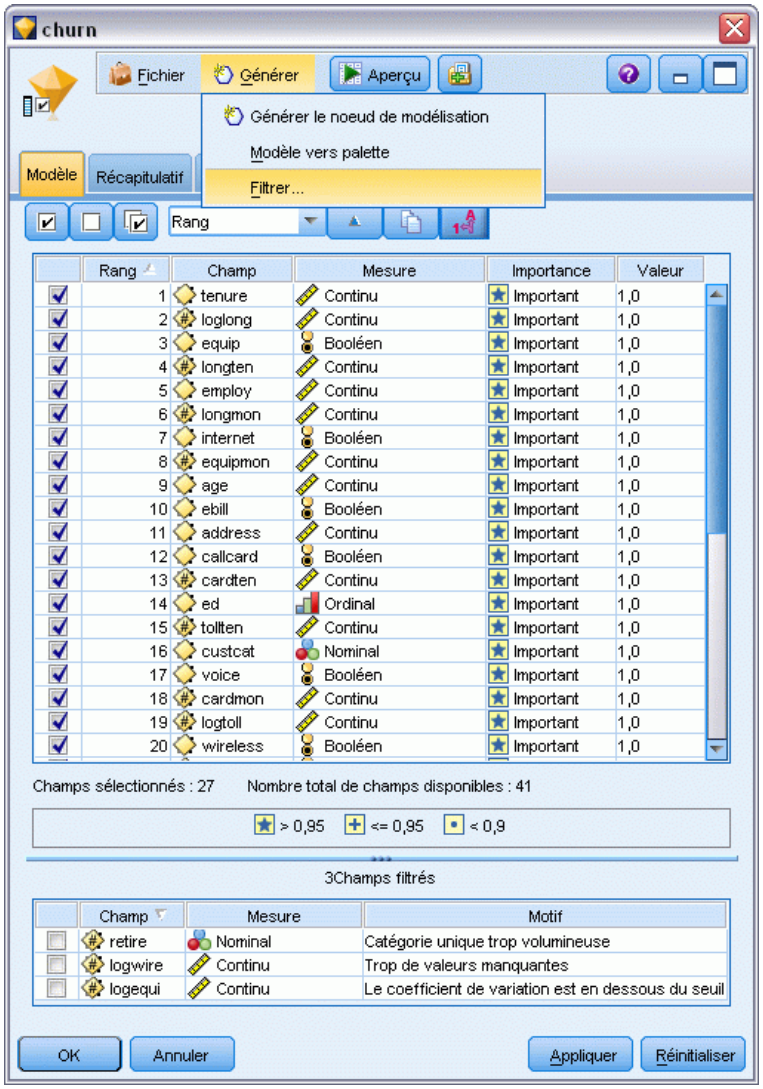
▶ Ajoutez au noeud Typer un noeud de modélisation Sélection de fonction.

L'utilisation d'un nœud Sélection de fonction vous permet de supprimer les variables indépendantes ou les données qui n'apportent aucune information utile en matière de relation variable indépendante/cible.

▶ Exécutez le flux.

▶ Ouvrez le nugget de modèle obtenu, et à partir du menu Générer, sélectionnez Filtrer pour créer un noeud Filtrer.

Figure 13-4
Génération d'un noeud Filtrer à partir d'un noeud Sélection de fonction

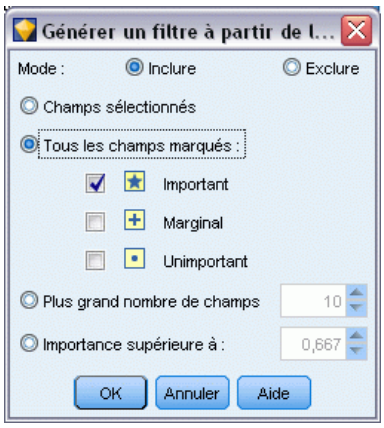


Toutes les données du fichier *telco.sav* ne sont pas utiles à la prévision de l'attrition. Vous pouvez appliquer le filtre pour ne sélectionner que les données considérées comme importantes en tant que variable indépendante.

▶ Dans la boîte de dialogue Générer un filtre, sélectionnez Tous les champs marqués : Important et cliquez sur OK.

▶ Reliez le noeud Filtrer généré au noeud Typer.

Figure 13-5
Sélection des champs importants



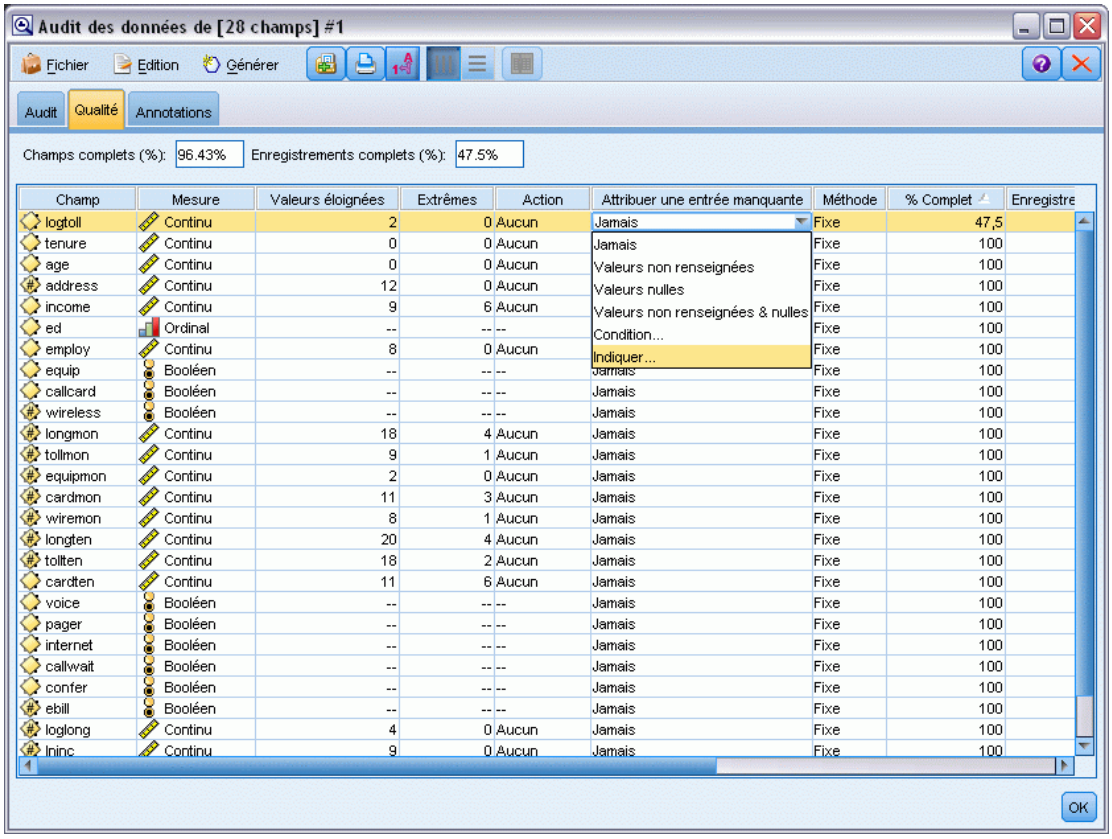
▶ Liez un noeud Audit données au noeud Filtrer généré.

Ouvrez le noeud Audit données, puis cliquez sur Exécuter.

▶ Dans l'onglet Qualité du navigateur Audit données, cliquez sur la colonne *% terminé(s)* pour la trier dans l'ordre numérique croissant. Vous pouvez ainsi identifier les champs où de grandes quantités de données manquent. Dans notre exemple, le seul champ à modifier est *logtoll*, qui est complet à moins de 50 %.

▶ Dans la colonne *Attribuer une entrée manquante* du champ *logtoll*, cliquez sur Spécifier.

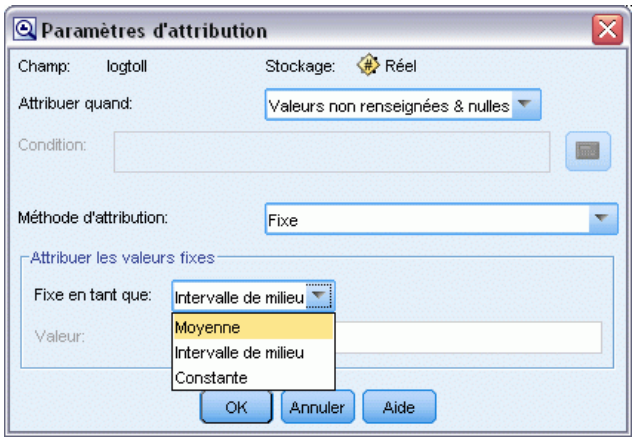
Figure 13-6
Attribution des valeurs manquantes au champ logtoll



▶ Dans le champ Attribuer quand, sélectionnez Valeurs nulles et non renseignées. Dans le champ Fixe en tant que, sélectionnez Moyenne et cliquez sur OK.

La sélection de Moyenne garantit que les valeurs attribuées n'ont pas d'impact négatif sur la moyenne de toutes les valeurs dans les données globales.

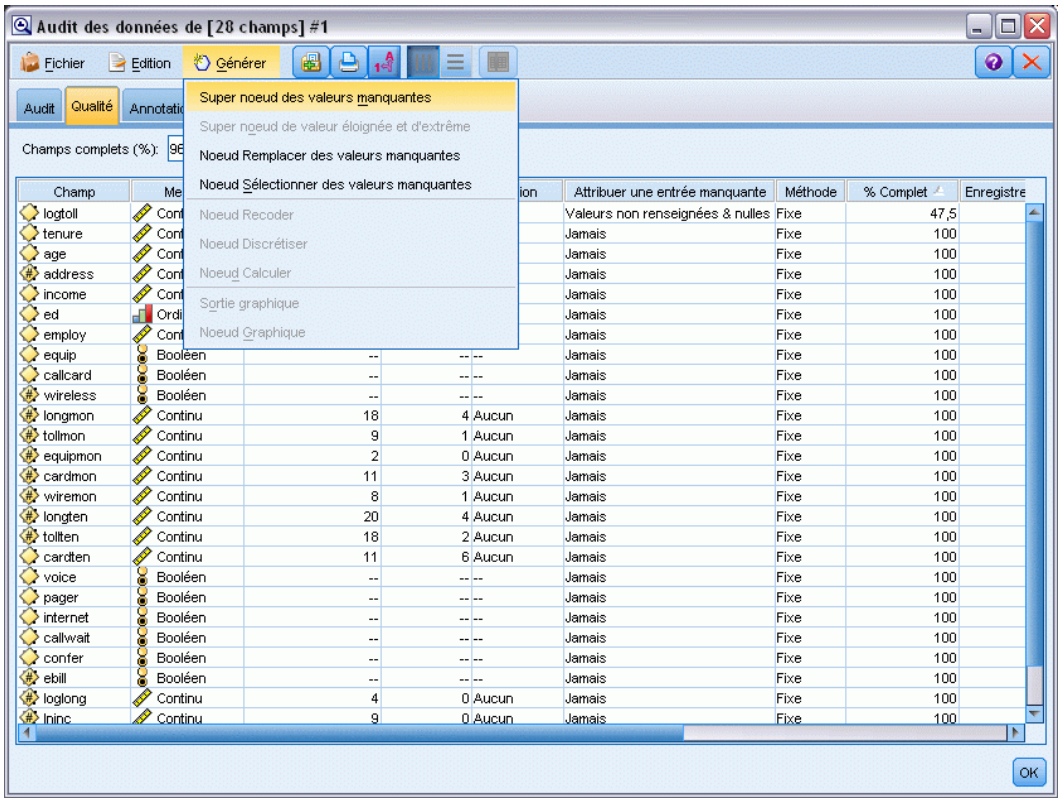
Figure 13-7
Sélection des paramètres d'attribution



- Dans l'onglet Qualité du navigateur Audit données, générez le super noeud Valeurs manquantes. Pour ce faire, dans les menus, choisissez :

  Générer > Super noeud des valeurs manquantes

Figure 13-8
Génération d'un super noeud des valeurs manquantes

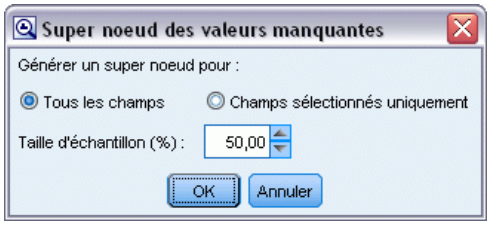


Dans la boîte de dialogue Super noeud des valeurs manquantes, augmentez le paramètre Taille d'échantillon (%) à 50 %, puis cliquez sur OK.

Le super noeud apparaît dans l'espace de travail de flux, avec l'intitulé : *Attribution de valeur manquante*.

▶ Reliez le super noeud au noeud Filtrer.

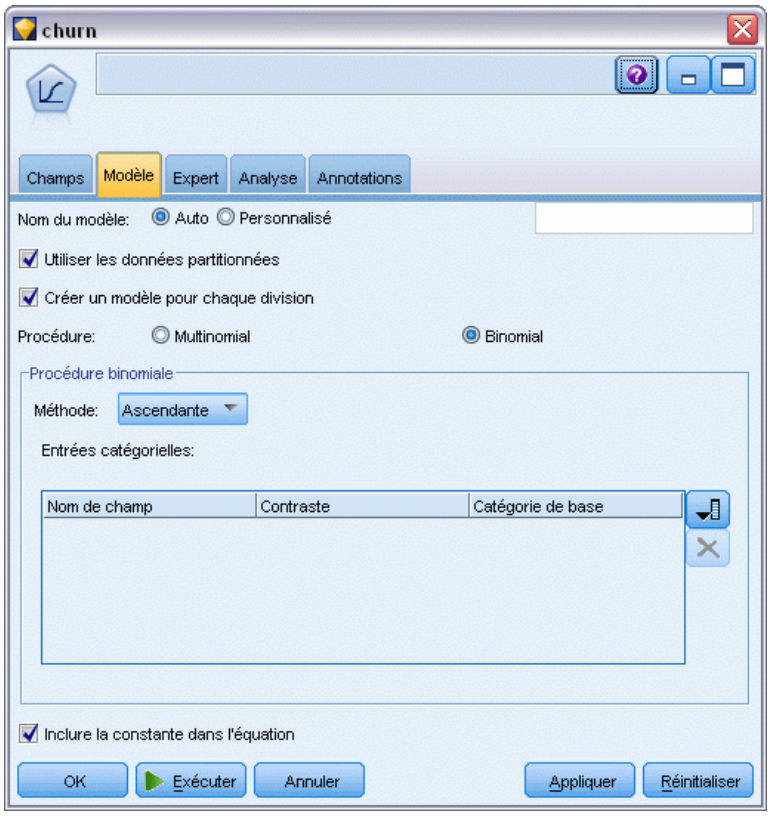
Figure 13-9
Définition de la taille d'échantillon



▶ Ajoutez un noeud Logistique au super noeud.

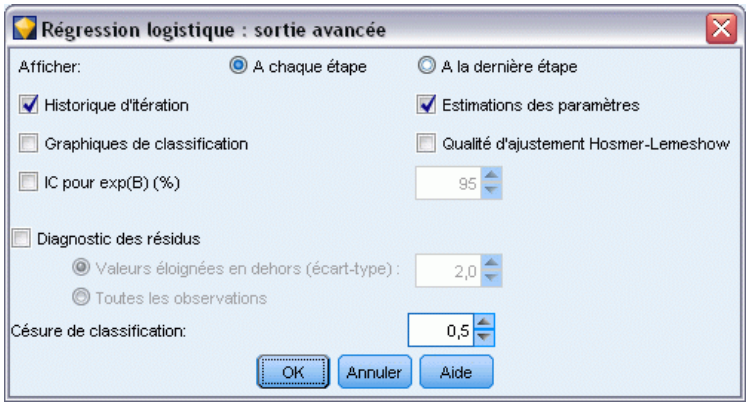
▶ Dans le noeud Logistique, cliquez sur l'onglet Modèle et sélectionnez la procédure Binomial. Dans la zone *Procédure binomiale*, sélectionnez la méthode Ascendante.

Figure 13-10
Choix des options de modèle



- Dans l'onglet Expert, sélectionnez le mode Expert, puis cliquez sur Sortie. La boîte de dialogue Sorties avancées apparaît.
- Dans la boîte de dialogue Sorties avancées, sélectionnez A chaque étape en tant que type *Afficher*. Sélectionnez Historique d'itération et Estimations des paramètres, puis cliquez sur OK.

Figure 13-11
Choix des options de sortie



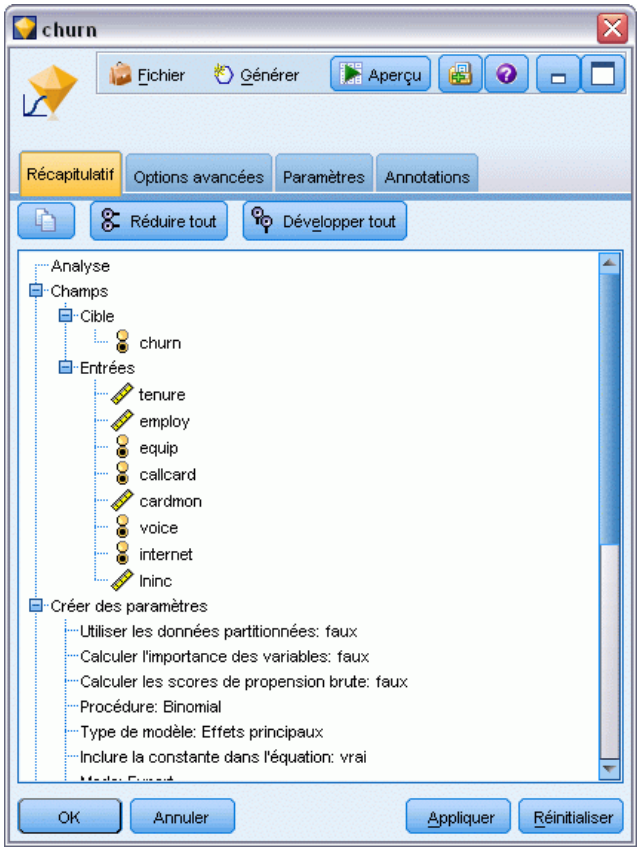
Navigation dans le modèle

- Dans le noeud Logistique, cliquez sur Exécuter pour créer le modèle.

  Le nugget de modèle est ajouté à l'espace de travail du flux et également à la palette Modèles en haut à droite. Pour afficher ses détails, cliquez avec le bouton droit de la souris sur le nugget de modèle et sélectionnez Editer ou Parcourir.

L'onglet Récapitulatif affiche (entre autres) la cible et les entrées (champs variables indépendantes) utilisées par le modèle. Ces champs sont ceux qui ont été réellement choisis sur la base de la méthode Ascendante, et non la liste complète soumise.

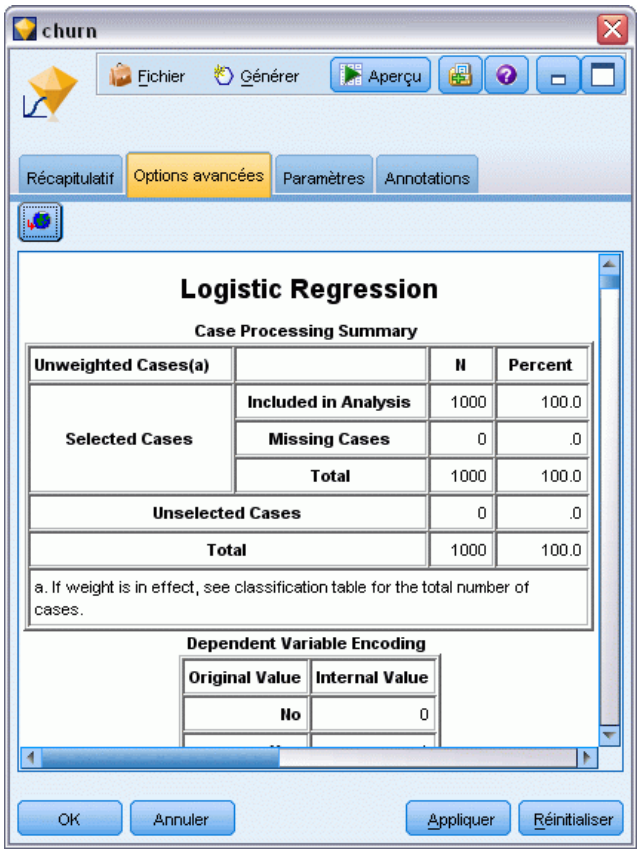
Figure 13-12
Récapitulatif du modèle avec champs cible et champs d'entrée



Les éléments affichés dans l'onglet Options avancées dépendent des options sélectionnées dans la boîte de dialogue Sorties avancées, dans le noeud Logistique. L'élément Récapitulatif du traitement des observations est systématiquement affiché. Il indique le nombre et le pourcentage d'enregistrements inclus dans l'analyse. Par ailleurs, il indique le nombre d'observations

manquantes (s'il y a lieu) où un ou plusieurs des champs d'entrée ne sont pas disponibles, ainsi que toutes les observations qui n'ont pas été sélectionnées.

Figure 13-13
Récapitulatif du traitement des observations

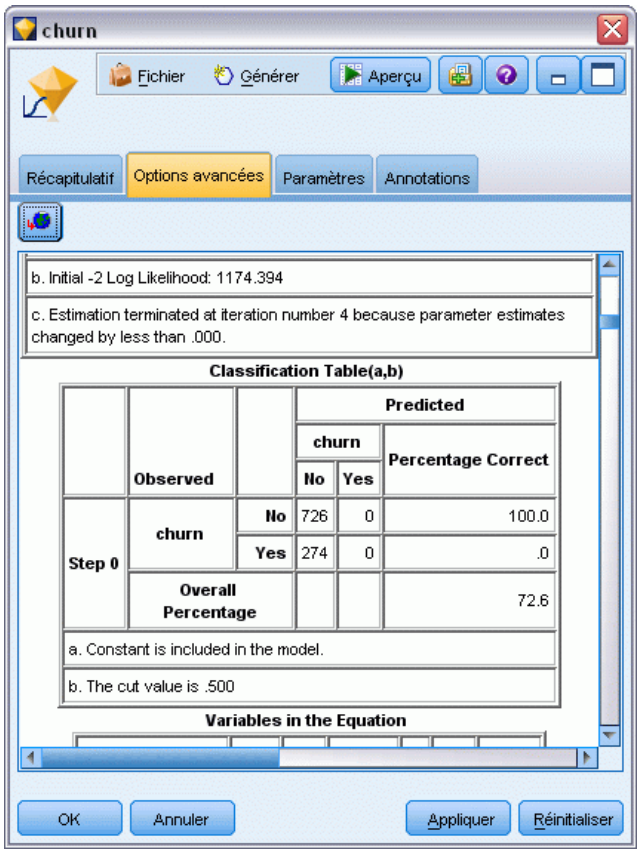


- Faites défiler la fenêtre vers le bas à partir de l'élément Récapitulatif du traitement des observations pour afficher la table de classification, sous Bloc 0 : Bloc de début.

  La méthode Pas à pas ascendante commence par un modèle nul, c'est-à-dire un modèle sans variable indépendante, qui peut servir de base à la comparaison avec le modèle final créé. Le modèle nul, par convention, donne systématiquement la valeur de prévision 0. Par conséquent, le modèle nul est précis à 72,6 %, tout simplement parce que les 726 clients n'ayant pas changé de

fournisseur font l'objet d'une prévision correcte. A l'inverse, la prévision concernant les clients ayant changé de fournisseur n'est pas du tout correcte.

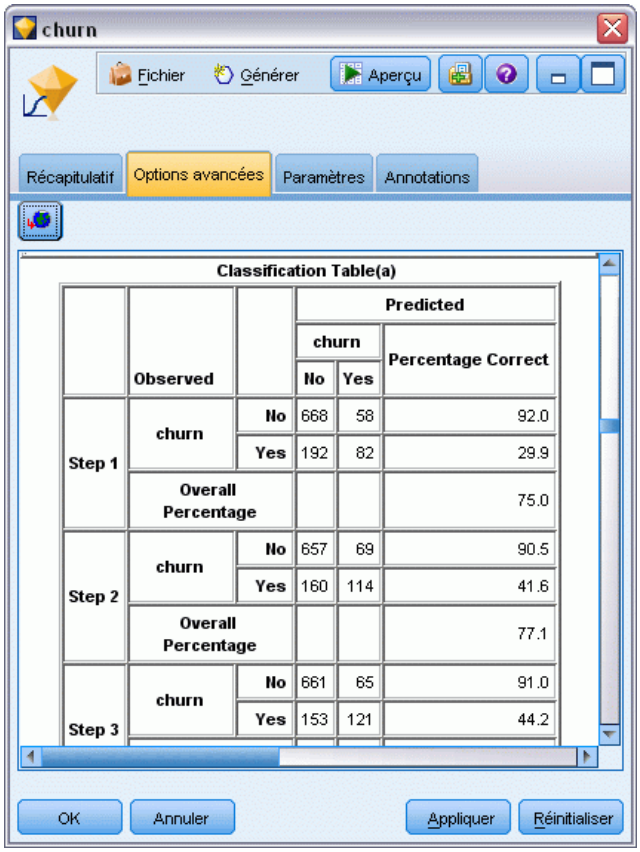
Figure 13-14
Début de la table de classification supervisée - Bloc 0



► Faites maintenant défiler la fenêtre vers le bas pour afficher la table de classification supervisée, sous Bloc 1 : Méthode = Pas à pas ascendante.

Cette table de classification supervisée affiche les résultats du modèle lors de l'ajout d'une variable indépendante à chacune des étapes. Dès la première étape (alors qu'une seule variable indépendante a été ajoutée), le modèle a permis d'augmenter la précision de la prévision d'attrition de 0,0 % à 29,9 %

Figure 13-15
Table de classification supervisée - Bloc 1

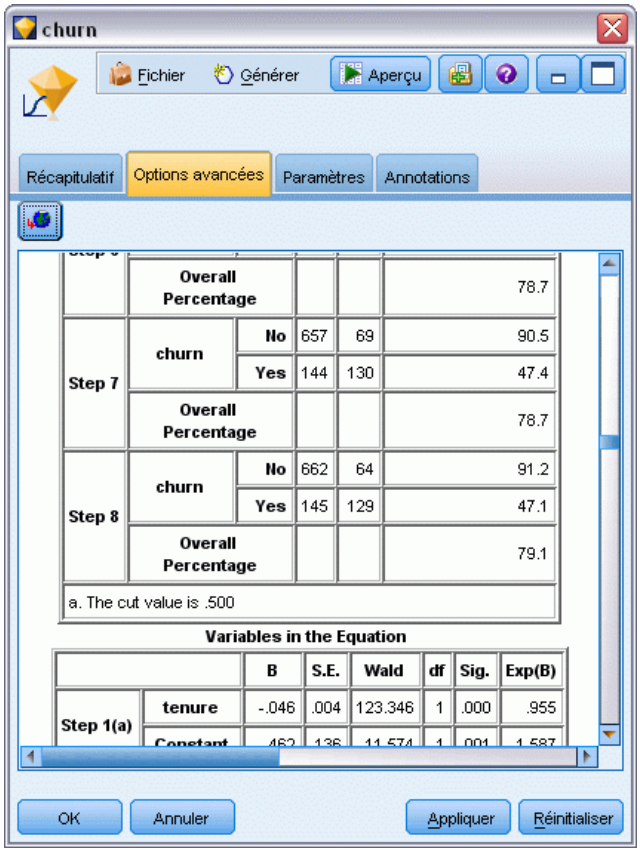


► Faites défiler la table de classification supervisée jusqu'en bas.

La table de classification supervisée montre que la dernière étape est l'étape 8. A ce stade, l'algorithme a décidé qu'il est inutile d'ajouter d'autres variables indépendantes au modèle. Bien que la précision des clients n'ayant pas changé de fournisseur ait diminué quelque peu jusqu'à 91,2%, la précision de la prévision des clients ayant changé de fournisseur a augmenté de la valeur

d'origine 0 %, à 47,1 %. Cela représente une amélioration significative par rapport au modèle nul d'origine sans variable indépendante.

Figure 13-16
Table de classification supervisée - Bloc 1



Pour un client qui souhaite réduire l'attrition, le fait de pouvoir la réduire presque de moitié est une étape majeure de protection des flux de revenus.

Remarque : cet exemple montre également que le fait de prendre le pourcentage global comme guide de précision d'un modèle peut, dans certains cas, être trompeur. Le modèle nul d'origine avait une précision globale de 72,6 %, alors que le modèle de prévision final présente une précision globale de 79.1%. Toutefois, comme nous l'avons vu, la précision des prévisions réelles par catégorie était très différente.

Pour évaluer le niveau d'adéquation du modèle aux données, divers diagnostics sont disponibles dans la boîte de dialogue Sorties avancées lorsque vous créez le modèle. Pour plus d'informations, reportez-vous à la section Nugget de modèle Logistique - Sorties avancées dans le chapitre 10 dans *Noeuds de modélisation de IBM SPSS Modeler 15*. Des explications sur les fondements mathématiques des méthodes de modélisation utilisées dans IBM® SPSS® Modeler sont présentées dans le *Guide des algorithmes SPSS Modeler*, disponible dans le répertoire *\Documentation* du disque d'installation.

Sachez également que ces résultats sont établis uniquement d'après les données d'apprentissage. Pour évaluer comment le modèle peut se généraliser à d'autres données dans le monde réel, vous devez utiliser un noeud Partitionner destiné à contenir un sous-ensemble d'enregistrements à des fins de test et de validation. Pour plus d'informations, reportez-vous à la section Noeud Partitionner dans le chapitre 4 dans *Noeuds source, exécution et de sortie de IBM SPSS Modeler 15*.

Chapitre
14

Prévision de l'utilisation de la bande passante (Séries temporelles)

Prévision avec le noeud Séries temporelles

Un analyste d'un fournisseur national de connexions haut débit doit réaliser des prévisions des abonnements utilisateurs afin de prédire le développement de l'utilisation du haut débit. Les prévisions sont requises pour chacun des marchés locaux qui constitue la base nationale de l'abonné. Vous utiliserez la modélisation des séries temporelles afin de produire des prévisions sur un sous-ensemble des marchés locaux pour les trois prochains mois. Un second exemple montre comment convertir des données source si leur format ne permet pas de les intégrer au noeud Séries temporelles.

Ces exemples utilisent le flux intitulé *broadband_create_models.str*, qui fait référence au fichier de données *broadband_1.sav*. Ces fichiers sont disponibles dans le dossier *Demos* de toutes les installations de IBM® SPSS® Modeler. Ce dossier est accessible à partir du groupe de programmes IBM® SPSS® Modeler dans le menu Démarrer de Windows. Le fichier *broadband_create_models.str* se trouve dans le dossier des *flux*.

Le dernier exemple montre comment appliquer les modèles enregistrés à un ensemble de données mis à jour afin de prolonger les prévisions d'une nouvelle période de trois mois.

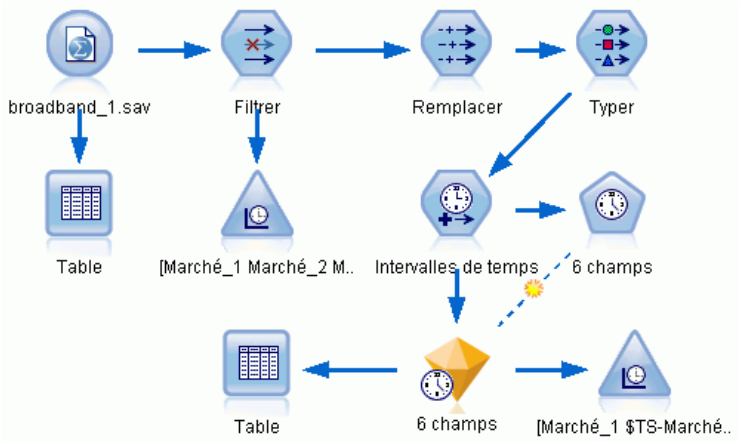
Dans SPSS Modeler, vous pouvez générer plusieurs modèles de séries temporelles simultanément. Le fichier source que vous utiliserez comporte les séries temporelles de 85 marchés différents. Toutefois, pour des raisons de simplicité, vous ne modélisez que cinq de ces marchés, ainsi que le total de tous les marchés.

Le fichier de données *broadband_1.sav* comporte les données d'utilisation mensuelle de chacun des 85 marchés locaux. Dans le cadre de cet exemple, seules les cinq premières séries seront utilisées ; un modèle distinct sera créé pour chacune de ces cinq séries, ainsi que pour un total.

En outre, le fichier comprend un champ de date qui indique le mois et l'année de chaque enregistrement. Ce champ sera utilisé dans un noeud Intervalles de temps pour l'étiquetage des enregistrements. Dans SPSS Modeler, le champ de date est lu en tant que chaîne. Cependant,

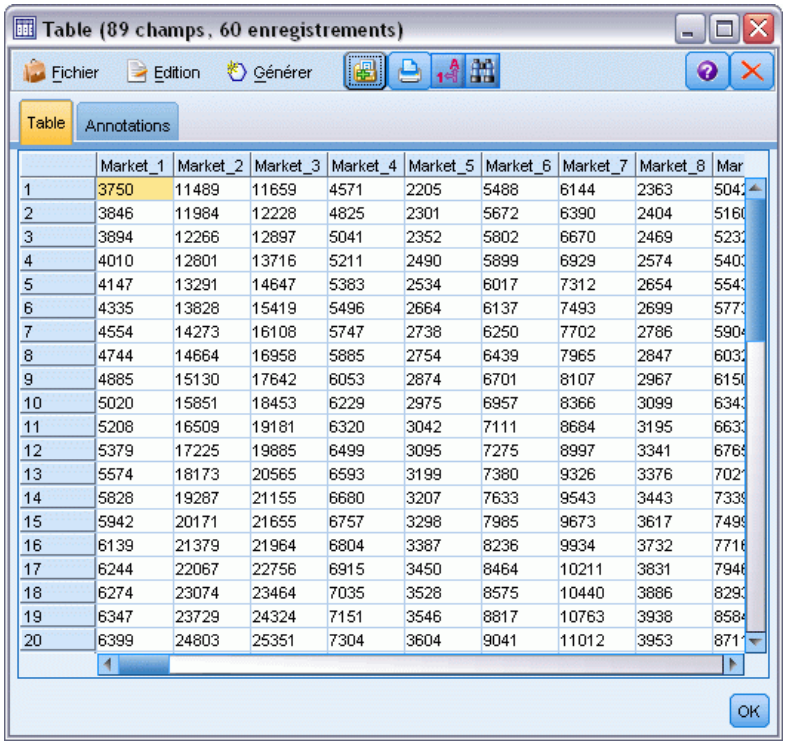
pour utiliser ce champ dans SPSS Modeler, vous convertirez le type de stockage au format de date numérique à l'aide d'un noeud Remplacer.

Figure 14-1
Exemple de flux illustrant la modélisation des séries temporelles



Le noeud Série temporelle requiert que chaque série se trouve dans une colonne distincte, avec une ligne par intervalle. SPSS Modeler propose des méthodes de transformation des données pour qu'elles correspondent à ce format, le cas échéant.

Figure 14-2
Données d'abonnement mensuelles pour les marchés locaux large bande



	Market_1	Market_2	Market_3	Market_4	Market_5	Market_6	Market_7	Market_8	Mar
1	3750	11489	11659	4571	2205	5488	6144	2363	504
2	3846	11984	12228	4825	2301	5672	6390	2404	516
3	3894	12266	12897	5041	2352	5802	6670	2469	523
4	4010	12801	13716	5211	2490	5899	6929	2574	540
5	4147	13291	14647	5383	2534	6017	7312	2654	554
6	4335	13828	15419	5496	2664	6137	7493	2699	577
7	4554	14273	16108	5747	2738	6250	7702	2786	590
8	4744	14664	16958	5885	2754	6439	7965	2847	603
9	4885	15130	17642	6053	2874	6701	8107	2967	615
10	5020	15851	18453	6229	2975	6957	8366	3099	634
11	5208	16509	19181	6320	3042	7111	8684	3195	663
12	5379	17225	19885	6499	3095	7275	8997	3341	676
13	5574	18173	20565	6593	3199	7380	9326	3376	702
14	5828	19287	21155	6680	3207	7633	9543	3443	733
15	5942	20171	21655	6757	3298	7985	9673	3617	749
16	6139	21379	21964	6804	3387	8236	9934	3732	771
17	6244	22067	22756	6915	3450	8464	10211	3831	794
18	6274	23074	23464	7035	3528	8575	10440	3886	829
19	6347	23729	24324	7151	3546	8817	10763	3938	858
20	6399	24803	25351	7304	3604	9041	11012	3953	871

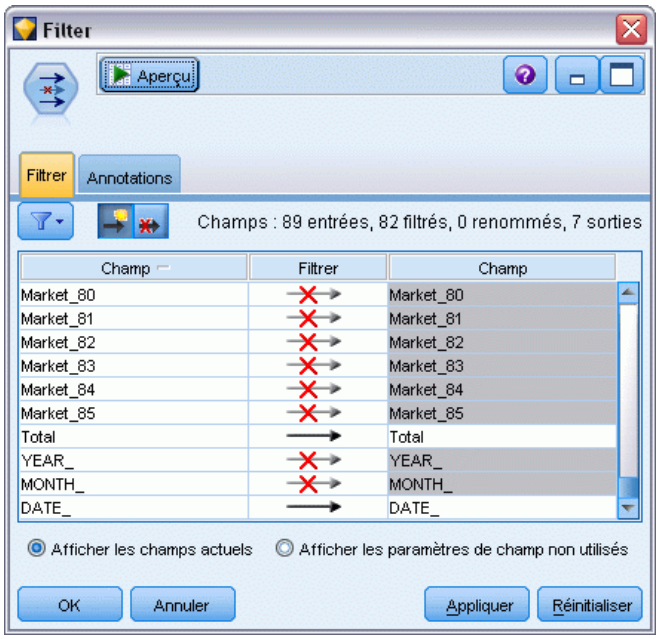
Création du flux

- Créez un nouveau flux, puis ajoutez un noeud source Statistics pointant vers le fichier *broadband_1.sav*.

- Utilisez un noeud Filtrer pour filtrer les champs *Market_6* à *Market_85*, ainsi que les champs *MONTH_* et *YEAR_*, afin de simplifier le modèle.

  Astuce : Pour sélectionner simultanément plusieurs champs adjacents, cliquez sur le champ *Market_6*, maintenez le bouton gauche de la souris enfoncé et faites glisser le curseur vers le bas jusqu'au champ *Market_85*. Les champs sélectionnés sont surlignés en bleu. Pour ajouter les autres champs, maintenez la touche Ctrl enfoncée et cliquez sur les champs *MONTH_* et *YEAR_*.

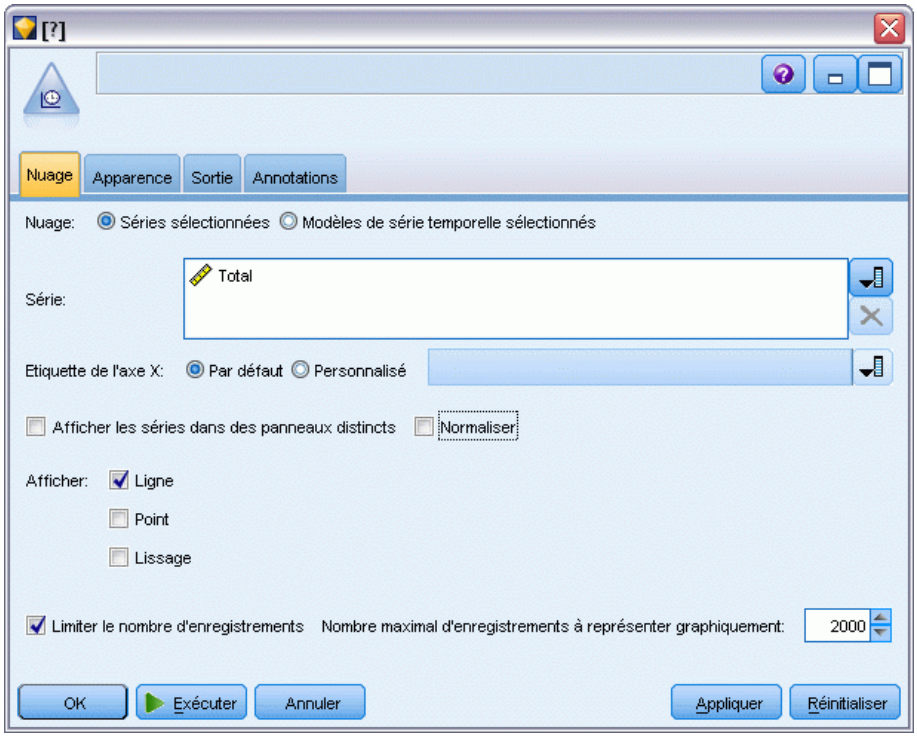
Figure 14-3
Simplification du modèle



Examen des données

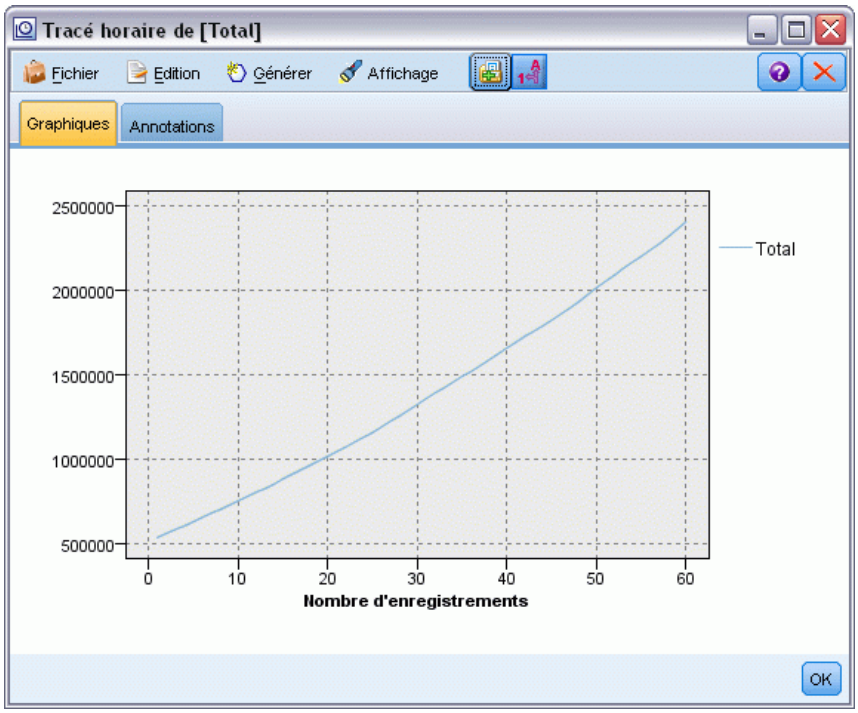
Il s'avère toujours utile d'examiner la nature de vos données avant de construire un modèle. Les données présentent-elles des variations saisonnières ? Bien que le modélisateur expert puisse rechercher automatiquement le meilleur modèle saisonnier ou non saisonnier pour chaque série, vous pouvez souvent obtenir des résultats plus rapidement en limitant la recherche aux modèles non saisonniers en l'absence d'effets saisonniers dans les données. Sans examiner les données de chacun des marchés locaux, nous pouvons obtenir une image approximative de la présence ou de l'absence d'effets saisonniers en traçant le nombre total d'abonnés pour l'ensemble des cinq marchés.

Figure 14-4
Traçage du nombre total d'abonnés



- Dans la palette Graphiques, reliez un noeud Tracé horaire au noeud Filtrer.
- Ajoutez le champ *Total* à la liste Série.
- Désélectionnez les cases Afficher les séries dans des panneaux distincts et Normaliser.
- Cliquez sur Exécuter.

Figure 14-5
Tracé horaire du champ Total

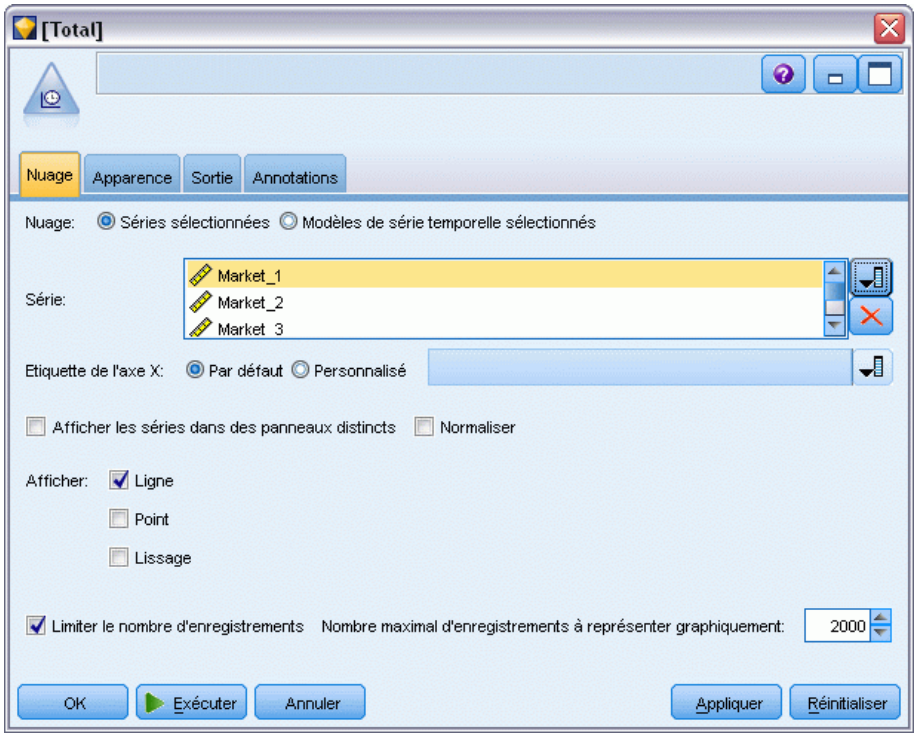


La série montre une tendance à la hausse très lisse avec aucun pic de variation saisonnière. Il se peut qu'il y ait des séries individuelles présentant une composante saisonnière mais cette composante n'est pas une particularité dominante dans les données en général.

Vous devez bien sûr inspecter chaque série avant d'exclure les modèles saisonniers. Vous pouvez alors séparer les séries présentant une composante saisonnière et les modéliser à part.

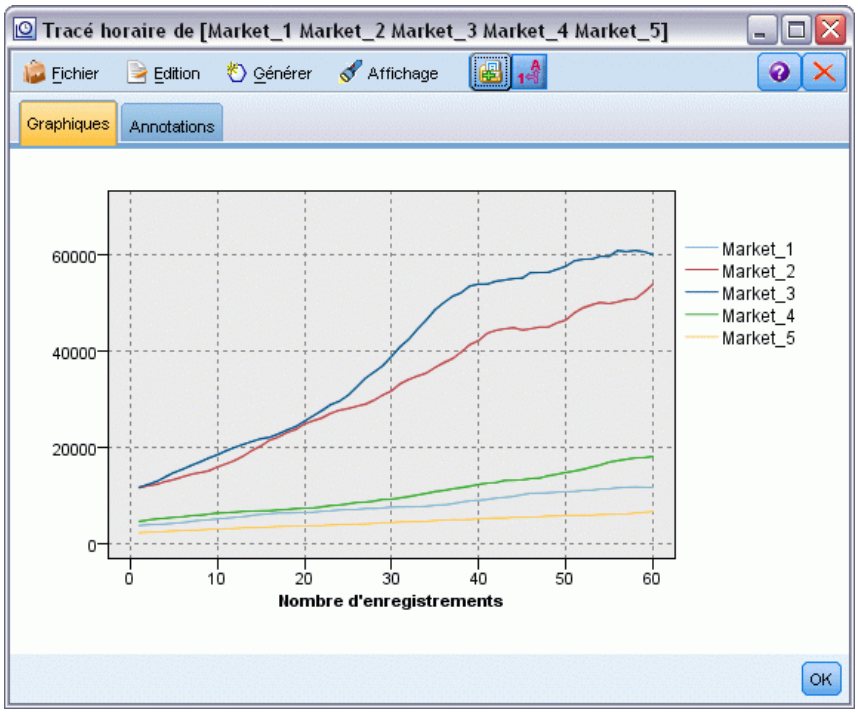
IBM® SPSS® Modeler facilite le traçage de plusieurs séries simultanément.

Figure 14-6
Traçage de plusieurs séries temporelles



- Rouvrez le noeud Tracé horaire.
- Supprimez le champ *Total* de la liste Série (sélectionnez-le, puis cliquez sur le bouton X rouge).
- Ajoutez les champs *Market_1* à *Market_5* à la liste.
- Cliquez sur Exécuter.

Figure 14-7
Tracé horaire de plusieurs champs



L'inspection de chacun des marchés met en évidence une tendance ascendante régulière dans chaque cas. Bien que certains marchés soient un peu plus imprévisibles que d'autres, rien n'atteste la présence d'effets saisonniers.

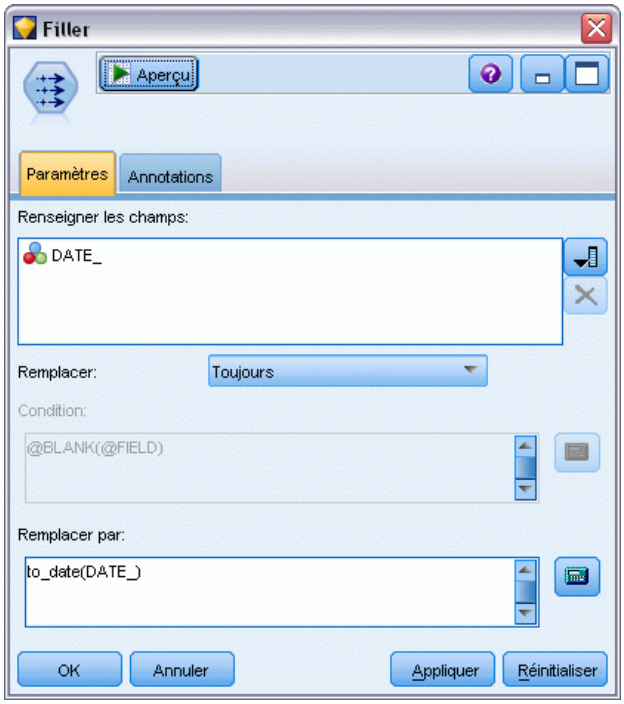
Définition des dates

Maintenant, vous devez attribuer le type de stockage de format Date au champ *DATE_*.

- Reliez un noeud Remplacer au noeud Filtrer.
- Ouvrez le noeud Remplacer, puis cliquez sur le bouton de sélection de champ.
- Sélectionnez le champ DATE_ afin de l'ajouter à la zone Renseigner les champs.
- Attribuez à la condition Remplacer la valeur Toujours.

▶ Attribuez à l'option Remplacer par la valeur to_date(DATE_).

Figure 14-8
Définition du type de stockage de date

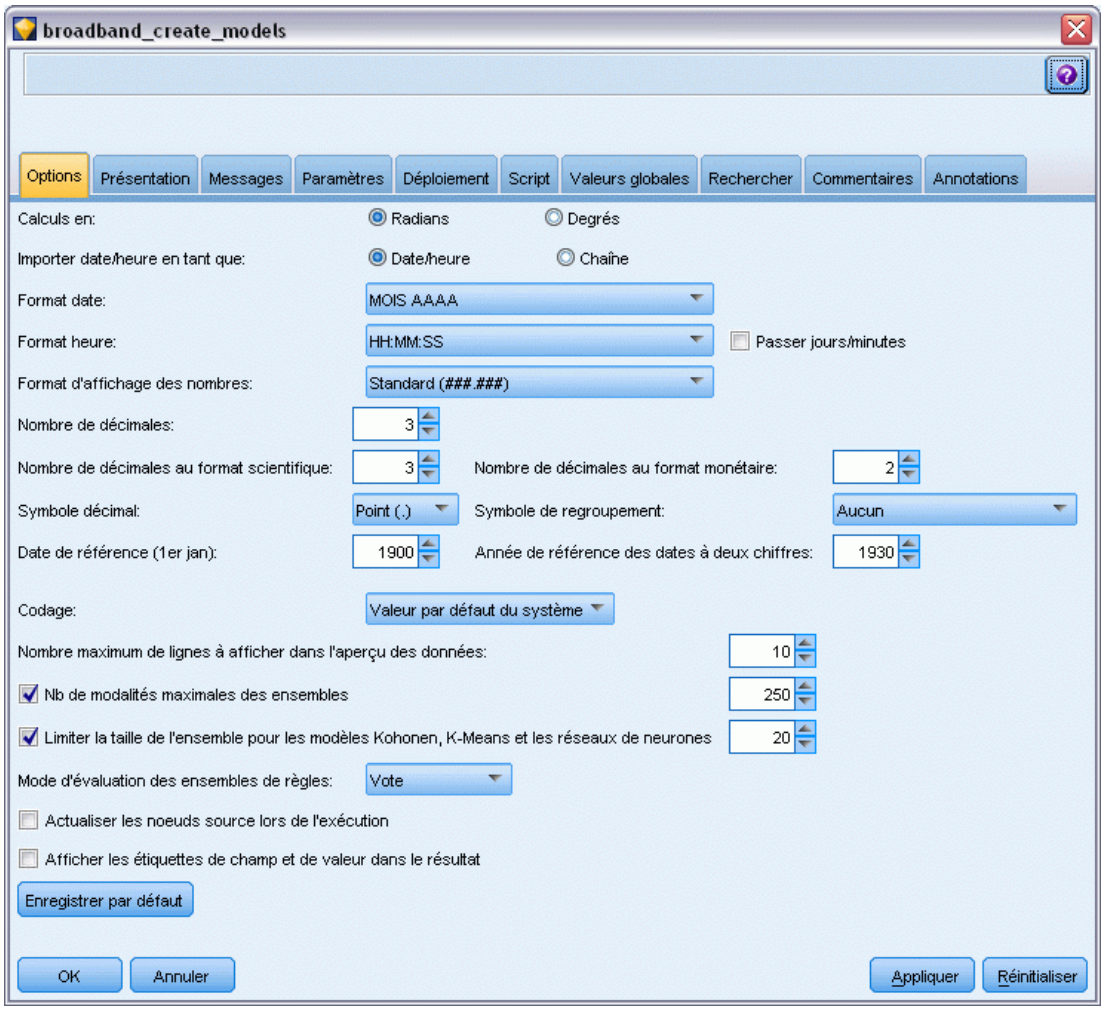


Modifiez le format de date par défaut afin qu'il corresponde au format du champ Date. Cette opération permet de convertir le champ Date correctement.

▶ Dans le menu, choisissez l'option Outils > Propriétés du flux > Options pour afficher la boîte de dialogue des options de flux.

▶ Attribuez à l'option Format de date par défaut la valeur MOIS AAAA .

Figure 14-9
Définition du format de date

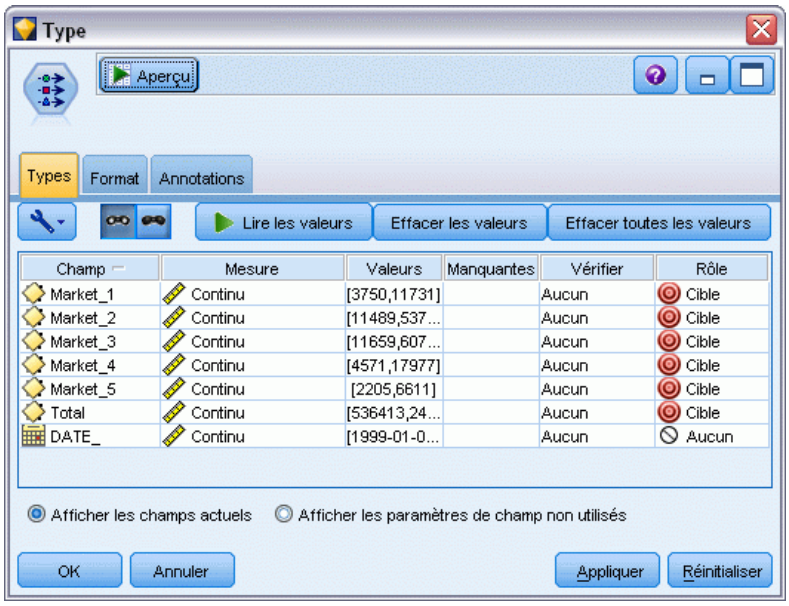


Définition des cibles

▶ Ajoutez un noeud Typer afin de définir le rôle sur Aucun pour le champ *DATE_*. Définissez le rôle sur Cible pour tous les autres champs (les champs *Market_n* ainsi que le champ *Total*).

▶ Cliquez sur le bouton Lire les valeurs pour remplir la colonne Valeurs.

Figure 14-10
Définition du rôle pour plusieurs champs



Définition des intervalles de temps

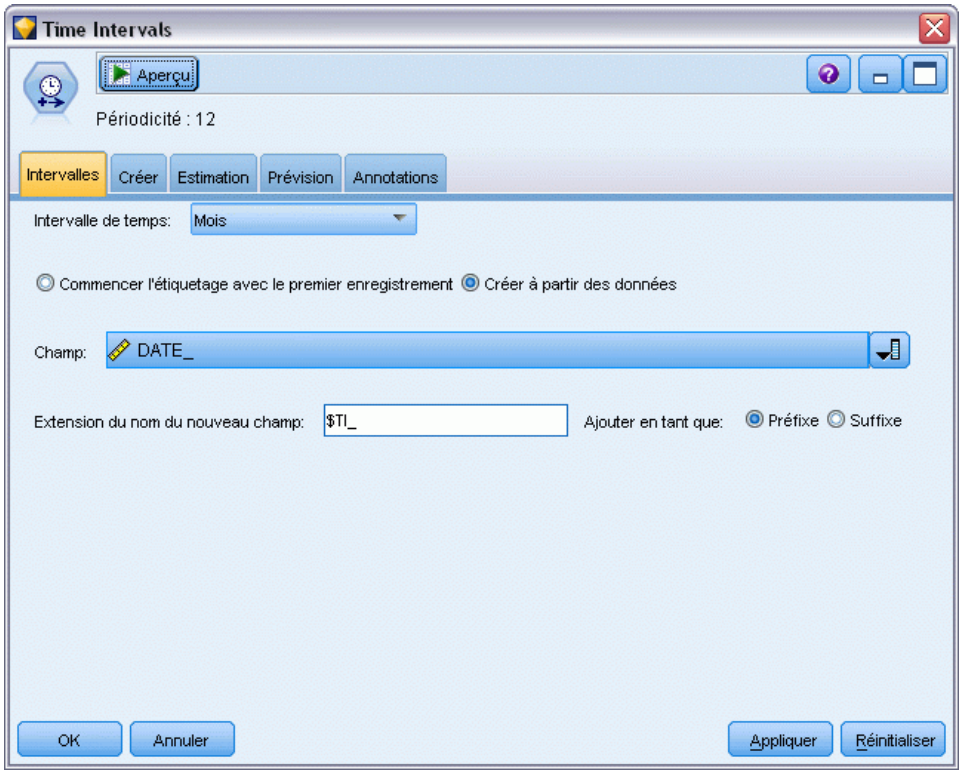
▶ Ajoutez un noeud Intervalles de temps (à partir de la palette Opérations sur les champs).

▶ Dans l'onglet Intervalles, sélectionnez l'option Mois comme intervalle de temps.

▶ Sélectionnez l'option Créer à partir des données.

▶ Sélectionnez l'option DATE_ comme champ à créer.

Figure 14-11
Définition de l'intervalle de temps

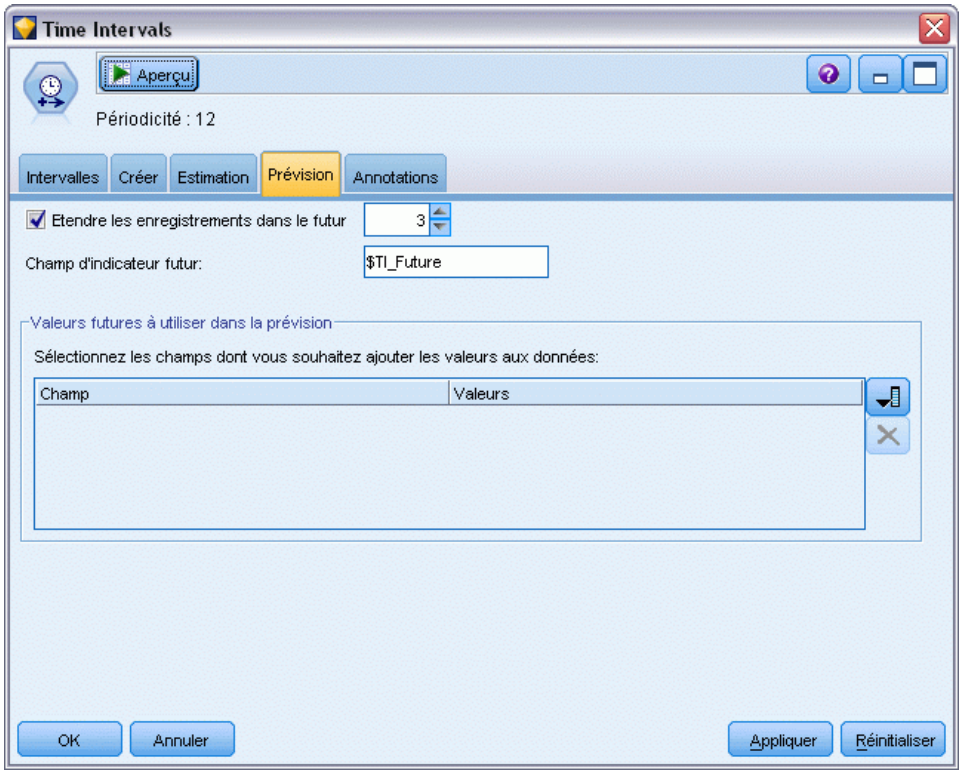


▶ Dans l'onglet Prévision, cochez la case Etendre les enregistrements dans le futur.

▶ Paramétrez la valeur sur 3.

▶ Cliquez sur OK.

Figure 14-12
Définition de la période de prévision

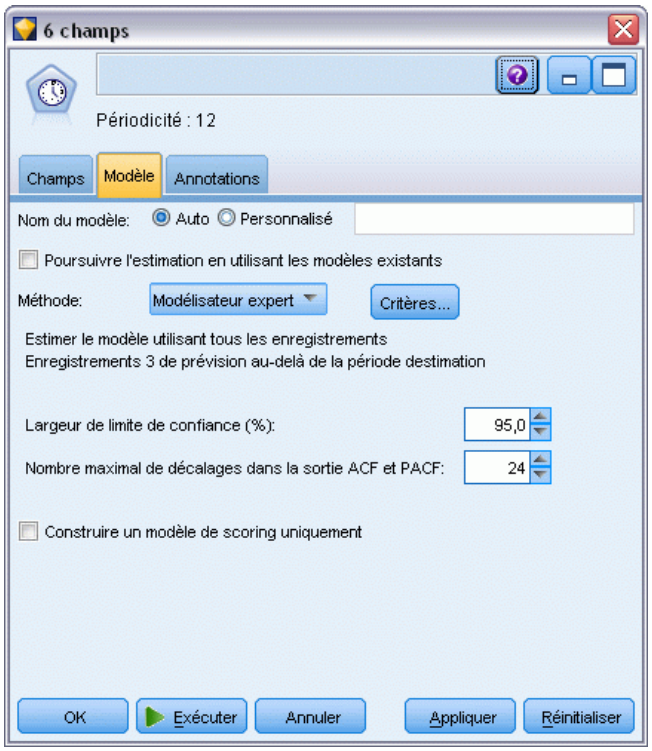


Création du modèle

▶ A partir de la palette Modélisation, ajoutez un noeud Séries temporelles au flux et reliez-le au noeud Intervalles de temps.

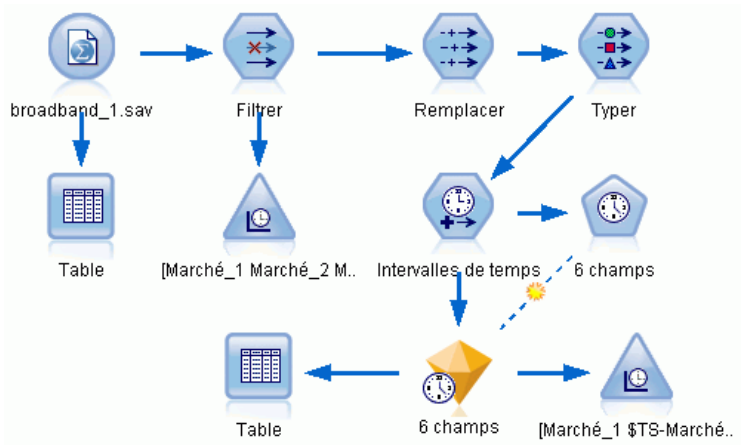
- Cliquez sur Exécuter sur le noeud Séries temporelles en utilisant tous les paramètres par défaut. Ainsi, le modélisateur expert peut décider du meilleur modèle à utiliser pour chaque série temporelle.

Figure 14-13
Choix du modélisateur expert pour les séries temporelles



- Reliez le nugget de modèle de séries temporelles au noeud Intervalles de temps.
- Reliez un noeud Table au modèle de séries temporelles et cliquez sur Exécuter.

Figure 14-14
Exemple de flux illustrant la modélisation des séries temporelles



Trois nouvelles lignes (61 à 63) sont désormais ajoutées aux données d'origine. Il s'agit des lignes relatives à la période de prévision, en l'occurrence janvier à mars 2004.

Plusieurs nouvelles colonnes sont également présentes maintenant : une série de colonnes *$TI_* ajoutées par le noeud Intervalles de temps et les colonnes *$TS-* ajoutées par le noeud Séries temporelles. Les colonnes indiquent les informations suivantes pour chaque ligne (c'est-à-dire, pour chaque intervalle dans les séries temporelles) :

Column	Description
$TI_TimeIndex	Valeur d'index de l'intervalle de temps pour cette ligne.
$TI_TimeLabel	Etiquette de l'intervalle de temps pour cette ligne.
$TI_Year	Indicateurs d'année et de mois pour les données générées dans cette ligne.
$TI_Month	
$TI_Count	Nombre d'enregistrements impliqués dans la détermination des nouvelles données pour cette ligne.
$TI_Future	Indique si cette ligne contient des données prévisionnelles.
$TS-*colname*	The generated model data for each column of the original data.
$TSLCI-*colname*	Valeur inférieure de l'intervalle de confiance pour chaque colonne de données du modèle généré.
$TSUCI-*colname*	Valeur supérieure de l'intervalle de confiance pour chaque colonne de données du modèle généré.
$TS-Total	Total des valeurs de $TS-*colname* pour cette ligne.
$TSLCI-Total	Total des valeurs de $TSLCI-*nom_colonne* pour cette ligne.
$TSUCI-Total	Total des valeurs de $TSUCI-*nom_colonne* pour cette ligne.

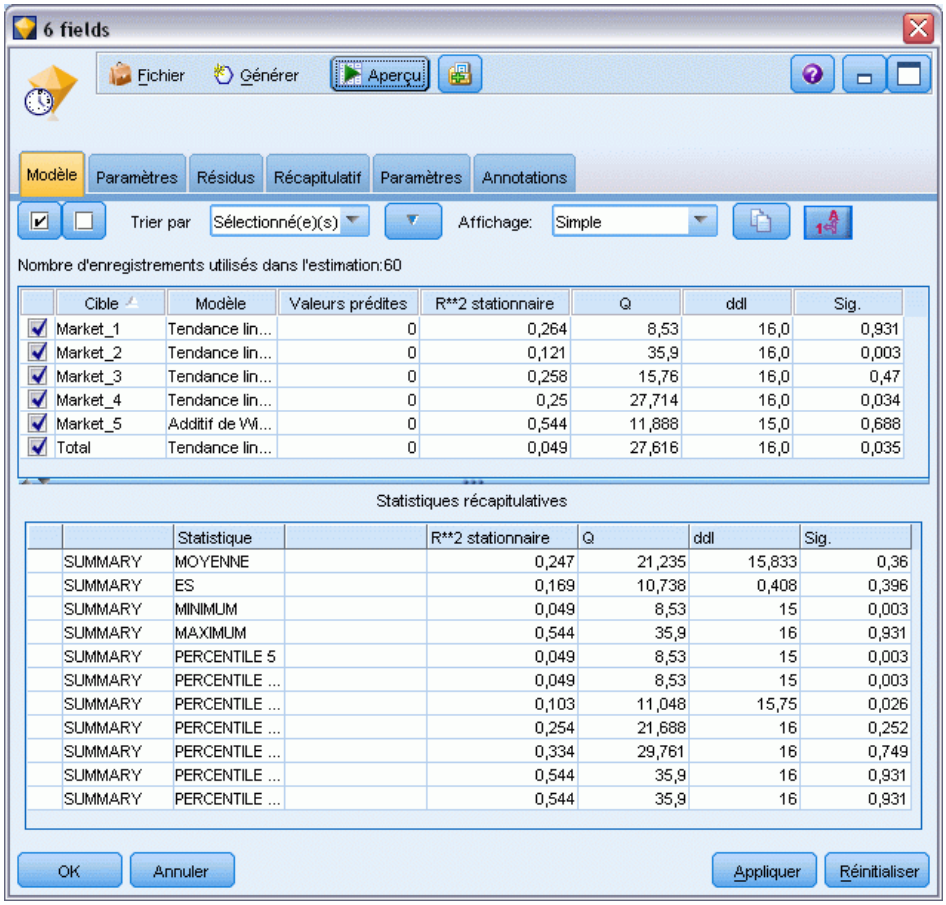
Les colonnes les plus significatives pour la prévision sont les colonnes *$TS-Market_n*, *$TSLCI-Market_n* et *$TSUCI-Market_n*. En particulier, dans les lignes 61 à 63, ces colonnes contiennent les données prévisionnelles sur les abonnements des utilisateurs et les intervalles de confiance de chacun des marchés locaux.

Examen du modèle

- Double-cliquez sur le nugget de modèle de séries temporelles afin d'afficher les données relatives aux modèles créés pour chacun des marchés.

Comme vous pouvez le constater, le modélisateur expert a choisi de générer pour le marché 5 un type de modèle différent de celui qu'il a créé pour les autres marchés.

Figure 14-15
Modèles de séries temporelles générés pour les marchés



Cible	Modèle	Valeurs prédites	R**2 stationnaire	Q	ddl	Sig.
Market_1	Tendance lin...	0	0,264	8,53	16,0	0,931
Market_2	Tendance lin...	0	0,121	35,9	16,0	0,003
Market_3	Tendance lin...	0	0,258	15,76	16,0	0,47
Market_4	Tendance lin...	0	0,25	27,714	16,0	0,034
Market_5	Additif de Wi...	0	0,544	11,888	15,0	0,688
Total	Tendance lin...	0	0,049	27,616	16,0	0,035

	Statistique		R**2 stationnaire	Q	ddl	Sig.
SUMMARY	MOYENNE		0,247	21,235	15,833	0,36
SUMMARY	ES		0,169	10,738	0,408	0,396
SUMMARY	MINIMUM		0,049	8,53	15	0,003
SUMMARY	MAXIMUM		0,544	35,9	16	0,931
SUMMARY	PERCENTILE 5		0,049	8,53	15	0,003
SUMMARY	PERCENTILE ...		0,049	8,53	15	0,003
SUMMARY	PERCENTILE ...		0,103	11,048	15,75	0,026
SUMMARY	PERCENTILE ...		0,254	21,688	16	0,252
SUMMARY	PERCENTILE ...		0,334	29,761	16	0,749
SUMMARY	PERCENTILE ...		0,544	35,9	16	0,931
SUMMARY	PERCENTILE ...		0,544	35,9	16	0,931

La colonne Variables indépendantes indique le nombre de champs utilisés comme variables indépendantes pour chaque cible, soit zéro en l'occurrence.

Les autres colonnes de cet affichage montrent différentes mesures de qualité d'ajustement pour chaque modèle. La colonne R**2 stationnaire indique la valeur *R*-deux stationnaire. Cette statistique estime dans quelle proportion le modèle explique la variation totale de la série. Plus la valeur est élevée (avec un maximum de 1,0), meilleur est l'ajustement du modèle.

Les colonnes Q, df et Sig. sont associées à la statistique Ljung-Box qui est un test de l'aspect aléatoire des erreurs résiduelles du modèle : plus les erreurs sont aléatoires, meilleur peut être le modèle. Q est la statistique Ljung-Box elle-même alors que df (degrés de liberté) indique le nombre de paramètres de modèles qui sont libres de varier lors de l'estimation d'une cible spécifique.

La colonne Sig affiche la valeur de signification de la statistique Ljung-Box, qui indique si le modèle est correctement spécifié. Une valeur de signification inférieure à 0,05 indique que les erreurs résiduelles ne sont pas aléatoires ce qui implique l'existence, dans la série observée, d'une structure inexpliquée par le modèle.

Si l'on prend en compte la valeur *R*-deux stationnaire et la valeur de signification, les modèles que Expert Modeler a choisi pour *Market_1*, *Market_3* et *Market_5* sont acceptables. Les valeurs de Sig. pour *Market_2* et *market_4* sont toutes deux inférieures à 0,05, ce qui signifie qu'une expérimentation avec des modèles mieux adaptés pour ces marchés pourrait s'avérer nécessaire.

Les valeurs récapitulatives dans la partie inférieure de l'affichage fournissent des informations sur la distribution des statistiques dans l'ensemble des modèles. Par exemple, la valeur *R*-deux stationnaire moyenne pour l'ensemble des modèles est de 0,247, avec un minimum de 0,049 (modèle *Total*) et un maximum de 0,544 (valeur de *Market_5*).

ES désigne l'erreur standard de chaque statistique pour la totalité des modèles. Par exemple, l'erreur standard du *R*-deux stationnaire pour l'ensemble des modèles est de 0,169.

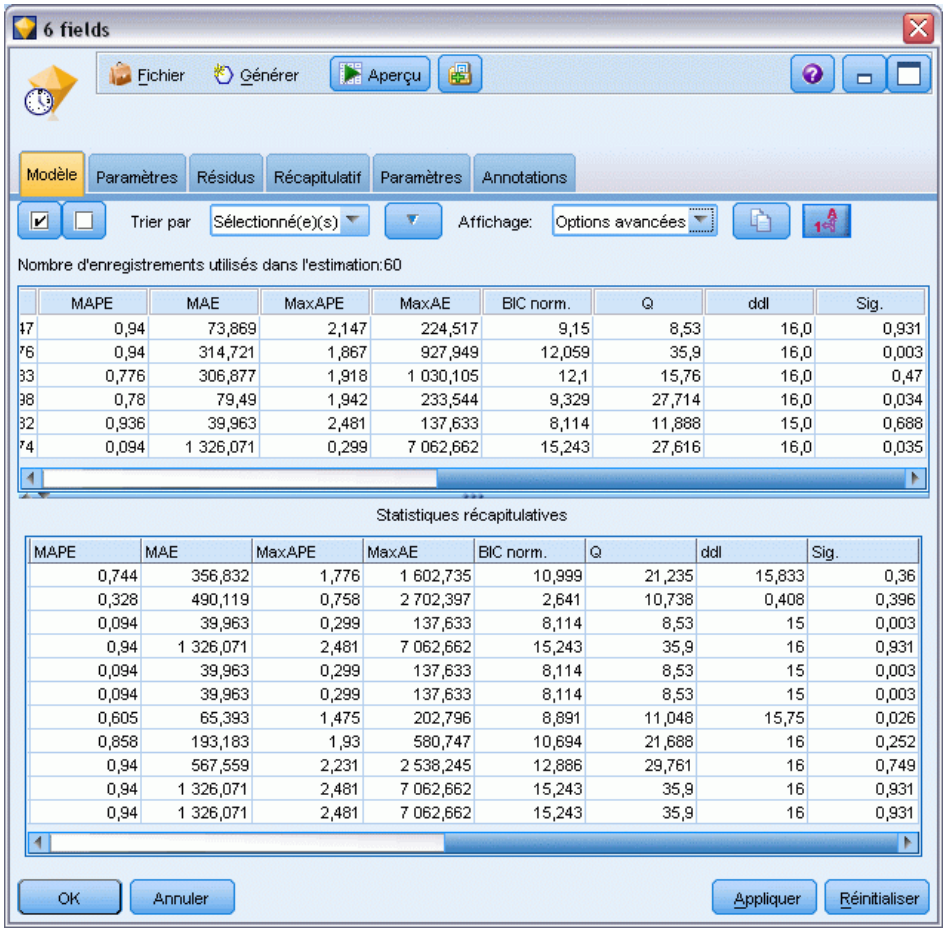
La section récapitulative comprend également des valeurs de centile qui fournissent des informations sur la distribution des statistiques dans les modèles. Une valeur de centile indique que le pourcentage de modèles correspondant comporte une valeur de statistique de l'ajustement inférieure à la valeur annoncée.

Ainsi, par exemple, seuls 25 % des modèles possèdent une valeur *R*-deux stationnaire inférieure à 0,121.

► Cliquez sur la liste déroulante Affichage et sélectionnez l'option Options avancées.

Une série de mesures de qualité d'ajustement supplémentaires apparaît. R**2 est la valeur *R*-deux, qui est une estimation de la variation totale de la série temporelle qui peut être expliquée par le modèle. Avec une valeur maximum de 1,0 pour cette statistique, nos modèles sont acceptables à cet égard.

Figure 14-16
Affichage avancé des modèles de séries temporelles



RMSE est l'erreur moyenne quadratique, une mesure de la différence entre les valeurs réelles d'une série et les valeurs prédites par le modèle et est exprimée dans la même unité que celle utilisée pour la série elle-même. En tant que mesure d'une erreur, cette valeur doit être ausi basse que possible. A première vue, il semble que les modèles de *Market_2* et *Market_3*, tout en étant acceptables selon les statistiques produites jusque là, sont moins réussis que ceux des trois autres marchés.

Ces mesures de qualité d'ajustement supplémentaires comprennent l'erreur moyenne absolue en pourcentage (MAPE) et sa valeur maximale (MaxAPE). L'erreur absolue en pourcentage mesure la proportion de la variation d'une série cible par rapport à son niveau prévu par le modèle et elle est exprimée en valeur de pourcentage. En examinant la moyenne et le maximum parmi tous les modèles, vous obtiendrez une indication de l'incertitude de vos prévisions.

La valeur MAPE indique que tous les modèles affichent une incertitude moyenne inférieure à 1%, ce qui est très bas. La valeur MaxAPE affiche l'erreur maximale absolue en pourcentage et permet d'imaginer le pire des scénarios pour vos prévisions. Elle indique que l'erreur maximale en pourcentage pour chacun des modèles se situe approximativement entre 1,8 et 2,5 % qui, de nouveau, sont des valeurs très basses.

La valeur MAE (erreur moyenne absolue) indique la moyenne des valeurs absolues des erreurs de prévision. Comme la valeur RMSE, elle est exprimée dans les mêmes unités que celles utilisées pour la série même. MaxAE indique l'erreur maximale de prévision dans les mêmes unités et le pire scénario de prévisions possible.

Bien que ces valeurs absolues soient intéressantes, il est préférable d'observer les valeurs des erreurs en pourcentage (MAPE et MaxAPE), car les séries cibles représentent les quantités d'abonnés appartenant à des marchés de taille variable.

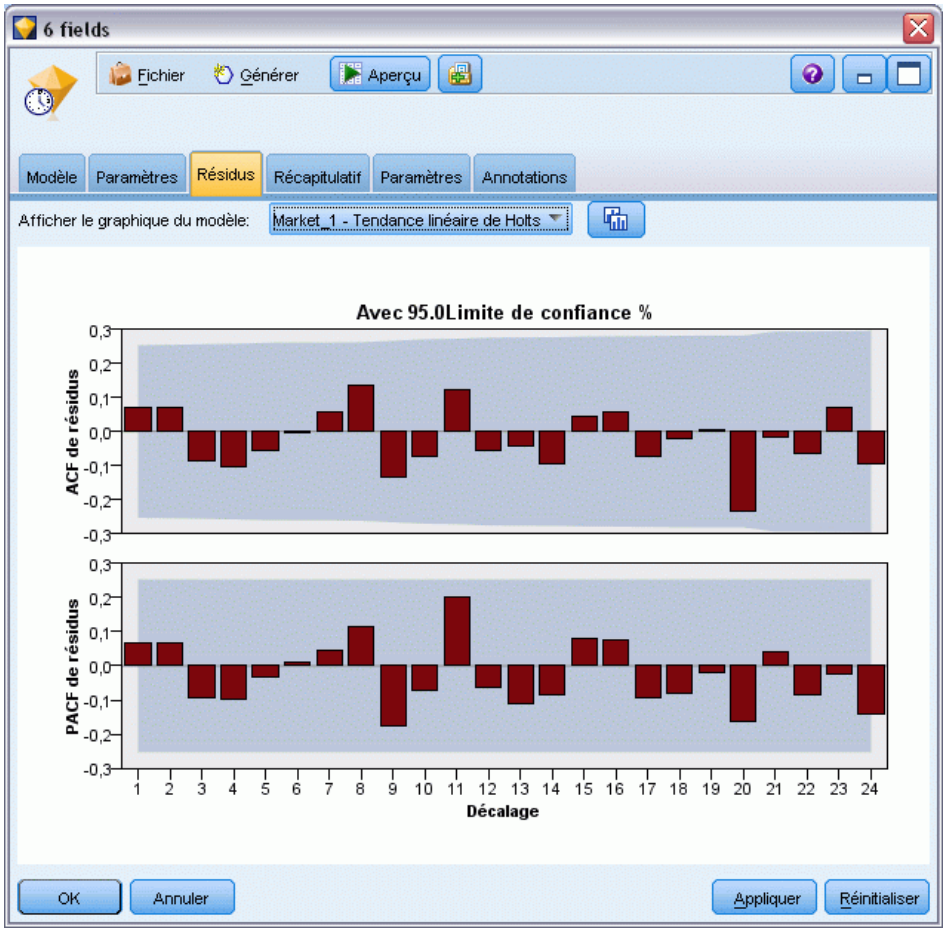
Les valeurs MAPE et MaxAPE représentent-elles un niveau d'incertitude acceptable pour les modèles ? Elles sont certainement très basses. Dans ce genre de situation, les facteurs relatifs au développement de l'entreprise entrent en jeu car le niveau de risque acceptable change d'un problème à l'autre. Nous supposons que les qualités d'ajustement des statistiques se trouvent dans des limites acceptables et examinons à présent les erreurs résiduelles.

L'examen des valeurs de la fonction des autocorrélations (ACF) et de la fonction des autocorrélations partielles (PACF) pour les résidus des modèles donne un aperçu plus quantitatif sur les modèles que la simple consultation des statistiques de qualité d'ajustement.

Un modèle de série temporelle bien défini capturera toutes les variations non-aléatoires, y compris l'effet des saisons, la tendance, la nature cyclique et les autres facteurs importants. Le cas échéant, aucune erreur ne devra être corrélée avec elle-même (autocorrélation). Toute structure significative dans l'une de ces fonctions d'autocorrélation impliquerait que le modèle sous-jacent soit incomplet.

▶ Cliquez sur l'onglet Résidus pour afficher les valeurs de la fonction d'autocorrélation (ACF) et de la fonction d'autocorrélation partielle (PACF) pour les erreurs résiduelles du modèle pour le premier marché local.

Figure 14-17
Valeurs des fonctions d'autocorrélation (ACF) et d'autocorrélation partielle (PACF) pour les marchés



Dans ces graphiques, les valeurs d'origine de la variable d'erreur ont été découpées en 24 périodes et comparées avec la valeur d'origine afin de savoir s'il existera des corrélations. Pour que le modèle soit acceptable, aucune des barres dans le graphique supérieur (ACF) ne doit dépasser la zone grisée, que ce soit dans les plus (vers le haut) ou dans les moins (vers le bas).

Si cela se produisait, il vous faudrait vérifier le graphique inférieur (PACF) pour savoir si la structure y est confirmée. Le graphique PACF examine les corrélations après avoir contrôlé les valeurs des séries aux points temporels intermédiaires.

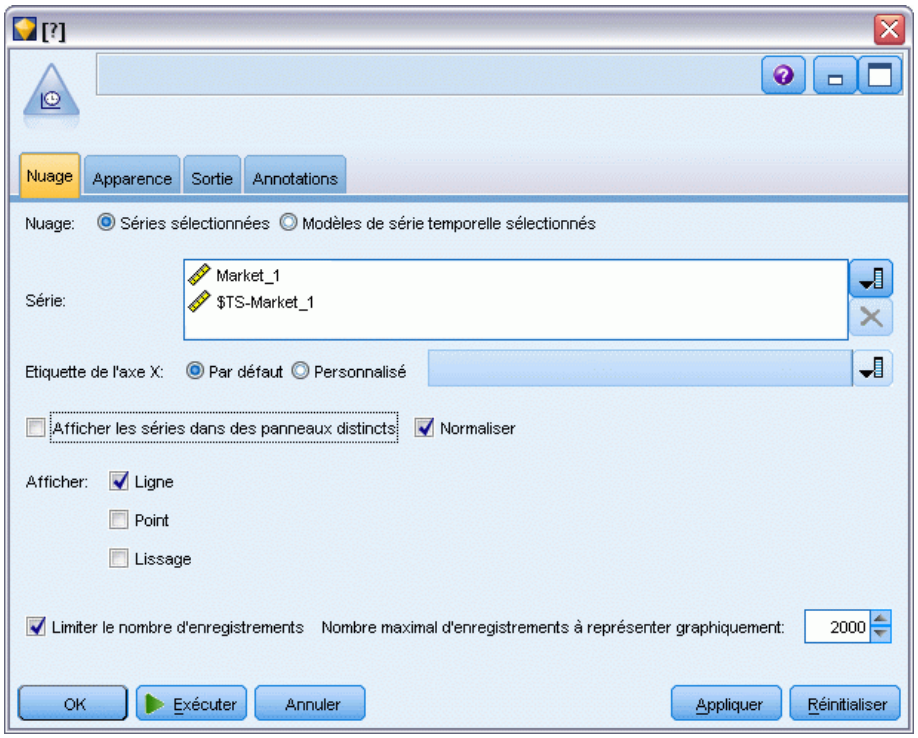
Les valeurs de *Market_1* étant toutes à l'intérieur de la zone grisée, nous pouvons continuer et vérifier les valeurs des autres marchés.

▶ Cliquez sur la liste déroulante Afficher le graphique du modèle afin d'afficher ces valeurs pour les autres marchés, ainsi que les totaux.

Les valeurs de *Market_2* et de *Market_4* sont un peu inquiétantes et confirment ce que nous avions suspecté auparavant au vue des valeurs Sig.. Il nous faudra utiliser d'autres modèles pour ces marchés afin de savoir s'il en existe un plus adapté, mais pour le reste de cet exemple, nous nous concentrerons sur ce que nous pouvons encore apprendre du modèle *Market_1*.

- Dans la palette Graphiques, reliez un noeud Tracé horaire au nugget de modèle Séries temporelles.
- Dans l'onglet Nuage, désélectionnez la case à cocher Afficher les séries dans des panneaux distincts.
- Dans la liste Série, cliquez sur le bouton de sélection de champ, sélectionnez les champs *Market_1* et *$TS-Market_1*, puis cliquez sur OK pour les ajouter à la liste.
- Cliquez sur Exécuter pour afficher un graphique linéaire des données réelles et prévisionnelles du premier marché local.

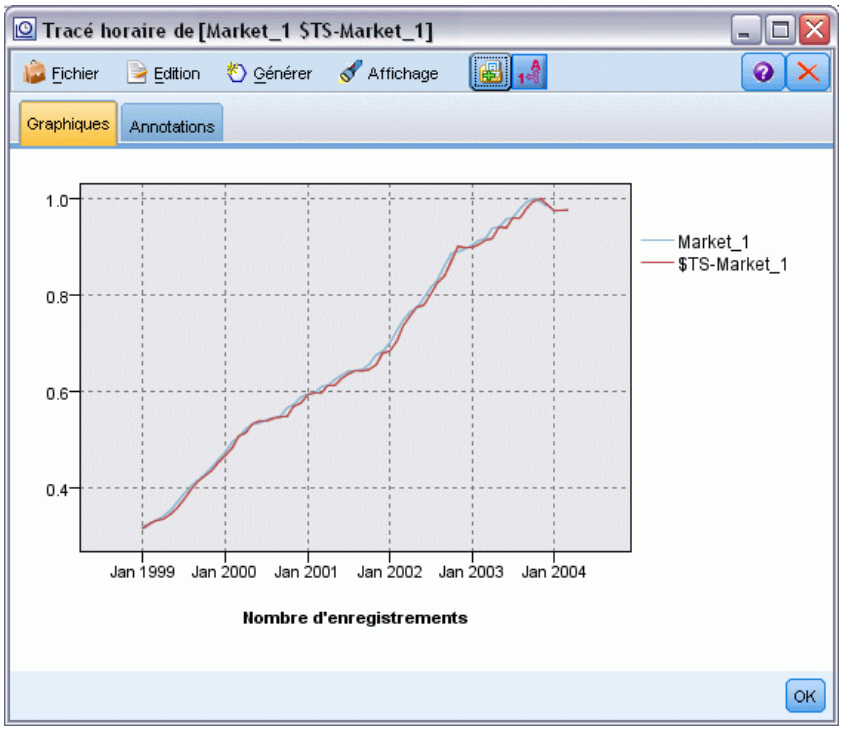
Figure 14-18
Sélection des champs à tracer



Comme vous pouvez le constater, la ligne prévisionnelle (*$TS-Market_1*) s'étend au-delà de la fin des données réelles. Vous disposez désormais d'une prévision de la demande pour les trois prochains mois dans ce marché.

Dans le graphique, les lignes des données réelles et prévisionnelles sont très proches l'une de l'autre sur la totalité de la série temporelle, ce qui indique que ce modèle est fiable pour cette série.

Figure 14-19
Tracé horaire des données réelles et prévisionnelles pour Market_1



Enregistrez le modèle dans un fichier pour l'utiliser ultérieurement dans un autre exemple :

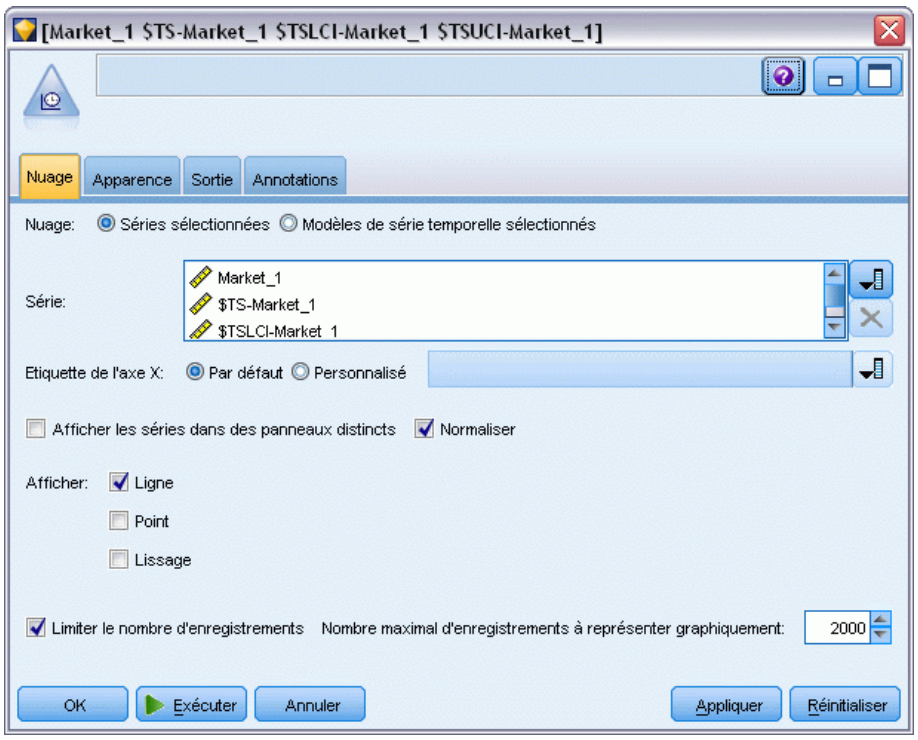
- Cliquez sur OK pour fermer le graphique actuel.
- Ouvrez le nugget de modèle Séries temporelles.
- Choisissez l'option Fichier > Enregistrer le noeud et spécifiez l'emplacement du fichier.
- Cliquez sur Enregistrer.

Vous disposez d'un modèle fiable pour ce marché en particulier, mais quelle est la marge d'erreur de la prévision ? Vous pouvez obtenir une indication en examinant l'intervalle de confiance.

- Double-cliquez sur le dernier noeud Tracé horaire dans le flux (celui nommé Market_1 $TS-Market_1) pour rouvrir sa boîte de dialogue.
- Cliquez sur le bouton de sélection de champ et ajoutez les champs *$TSLCI-Market_1* et *$TSUCI-Market_1* à la liste Série.

▶ Cliquez sur Exécuter.

Figure 14-20
Ajout d'autres champs à tracer



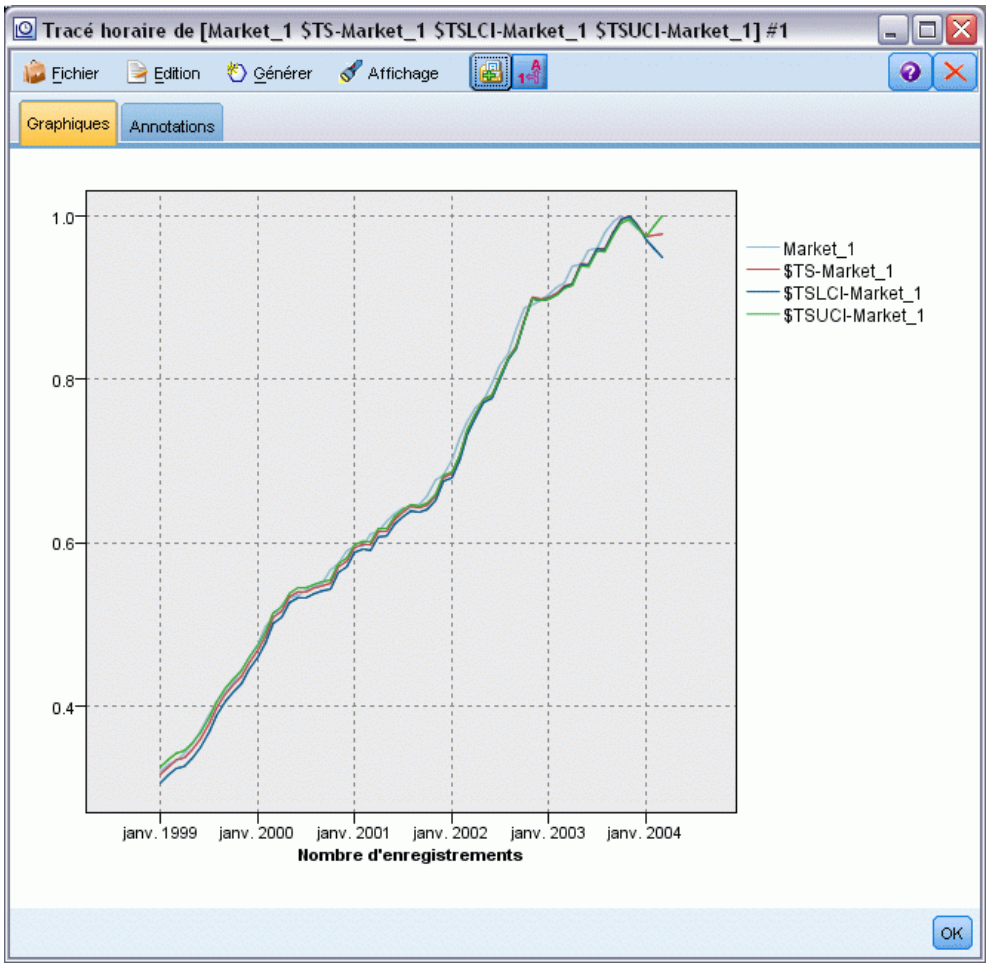
Vous obtenez le même graphique qu'auparavant avec, en plus, les limites supérieure (*$TSUCI*) et inférieure (*$TSLCI*) de l'intervalle de confiance.

Comme vous pouvez le constater, les limites de l'intervalle de confiance divergent sur l'ensemble de la période de prévision, ce qui traduit une incertitude croissante dès lors que la prévision porte sur une période plus longue.

Toutefois, au terme de chaque période, vous disposez, en l'occurrence, d'un mois supplémentaire de données d'utilisation réelles sur lesquelles vous pouvez baser vos prévisions. Vous pouvez lire les nouvelles données du flux et réappliquer le modèle, puisque vous savez

que celui-ci est fiable. Pour plus d'informations, reportez-vous à la section Réapplication d'un modèle de séries temporelles sur p. 201.

Figure 14-21
Tracé horaire comportant l'intervalle de confiance



Récapitulatif

Vous avez appris à utiliser le modélisateur expert afin de produire des prévisions pour plusieurs séries temporelles et avez enregistré les modèles obtenus dans un fichier externe.

Dans l'exemple suivant, vous allez apprendre à convertir les séries temporelles non standard dans un format permettant de les intégrer à un noeud Séries temporelles.

Réapplication d'un modèle de séries temporelles

Cet exemple applique les modèles de séries temporelles issus du premier exemple de séries temporelles, mais il peut être utilisé indépendamment. Pour plus d'informations, reportez-vous à la section Prévision avec le noeud Séries temporelles sur p. 178.

Comme dans le scénario d'origine, un analyste pour un fournisseur de large bande national doit établir des prévisions mensuelles sur les abonnements des utilisateurs pour chaque marché d'une série de marchés locaux, afin de prédire les exigences en matière de bande passante. Vous avez déjà utilisé le modélisateur expert pour créer des modèles et effectuer des prévisions à trois mois.

Votre entrepôt de données ayant été mis à jour avec les données réelles correspondant à la période prévisionnelle d'origine, vous souhaitez utiliser ces données pour étendre l'horizon de prévision d'une nouvelle période de trois mois.

Cet exemple utilise le flux intitulé *broadband_apply_models.str*, qui référence le fichier de données *broadband_2.sav*. Ces fichiers sont disponibles dans le dossier *Demos* de toutes les installations de IBM® SPSS® Modeler. Ce dossier est accessible à partir du groupe de programmes IBM® SPSS® Modeler dans le menu Démarrer de Windows. Le fichier *broadband_apply_models.str* se trouve dans le dossier des *flux* .

Récupération du flux

Dans cet exemple, vous allez recréer un noeud Séries temporelles à partir du modèle de séries temporelles enregistré dans le premier exemple. Ne vous inquiétez pas si vous ne disposez d'aucun modèle enregistré ; le dossier *Demos* en contient déjà un.

- Ouvrez le flux *broadband_apply_models.str* à partir du dossier des *flux* du dossier *Demos*.

Figure 14-22
Ouverture du flux

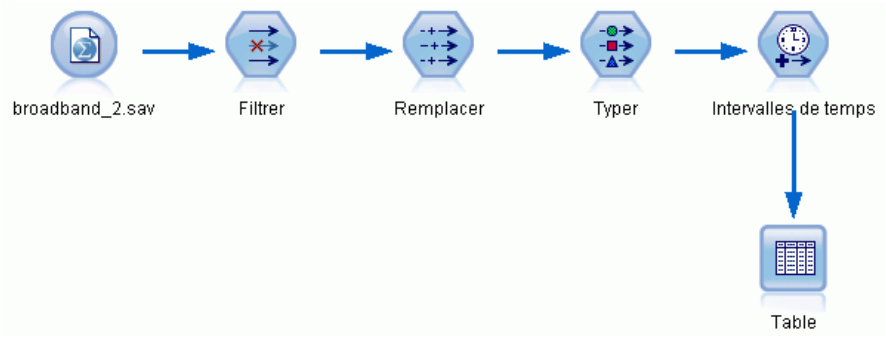
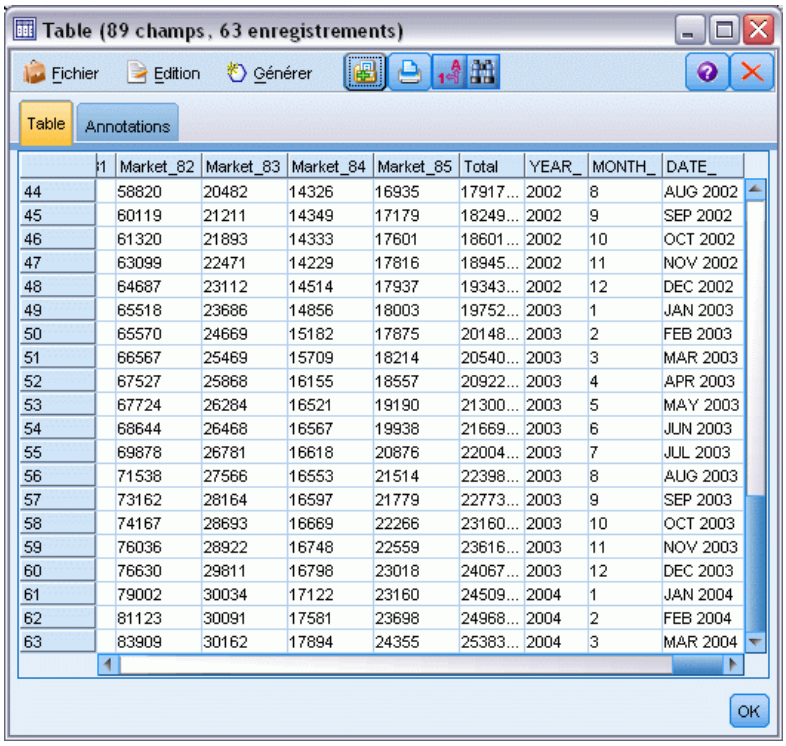


Figure 14-23
Données de vente mises à jour



	Market_82	Market_83	Market_84	Market_85	Total	YEAR_	MONTH_	DATE_
44	58820	20482	14326	16935	17917...	2002	8	AUG 2002
45	60119	21211	14349	17179	18249...	2002	9	SEP 2002
46	61320	21893	14333	17601	18601...	2002	10	OCT 2002
47	63099	22471	14229	17816	18945...	2002	11	NOV 2002
48	64687	23112	14514	17937	19343...	2002	12	DEC 2002
49	65518	23686	14856	18003	19752...	2003	1	JAN 2003
50	65570	24669	15182	17875	20148...	2003	2	FEB 2003
51	66567	25469	15709	18214	20540...	2003	3	MAR 2003
52	67527	25868	16155	18557	20922...	2003	4	APR 2003
53	67724	26284	16521	19190	21300...	2003	5	MAY 2003
54	68644	26468	16567	19938	21669...	2003	6	JUN 2003
55	69878	26781	16618	20876	22004...	2003	7	JUL 2003
56	71538	27566	16553	21514	22398...	2003	8	AUG 2003
57	73162	28164	16597	21779	22773...	2003	9	SEP 2003
58	74167	28693	16669	22266	23160...	2003	10	OCT 2003
59	76036	28922	16748	22559	23616...	2003	11	NOV 2003
60	76630	29811	16798	23018	24067...	2003	12	DEC 2003
61	79002	30034	17122	23160	24509...	2004	1	JAN 2004
62	81123	30091	17581	23698	24968...	2004	2	FEB 2004
63	83909	30162	17894	24355	25383...	2004	3	MAR 2004

Les données mensuelles mises à jour sont collectées dans le fichier *broadband_2.sav*.

- Reliez un noeud Table au noeud source du fichier IBM® SPSS® Statistics, ouvrez le noeud Table et cliquez sur Exécuter.

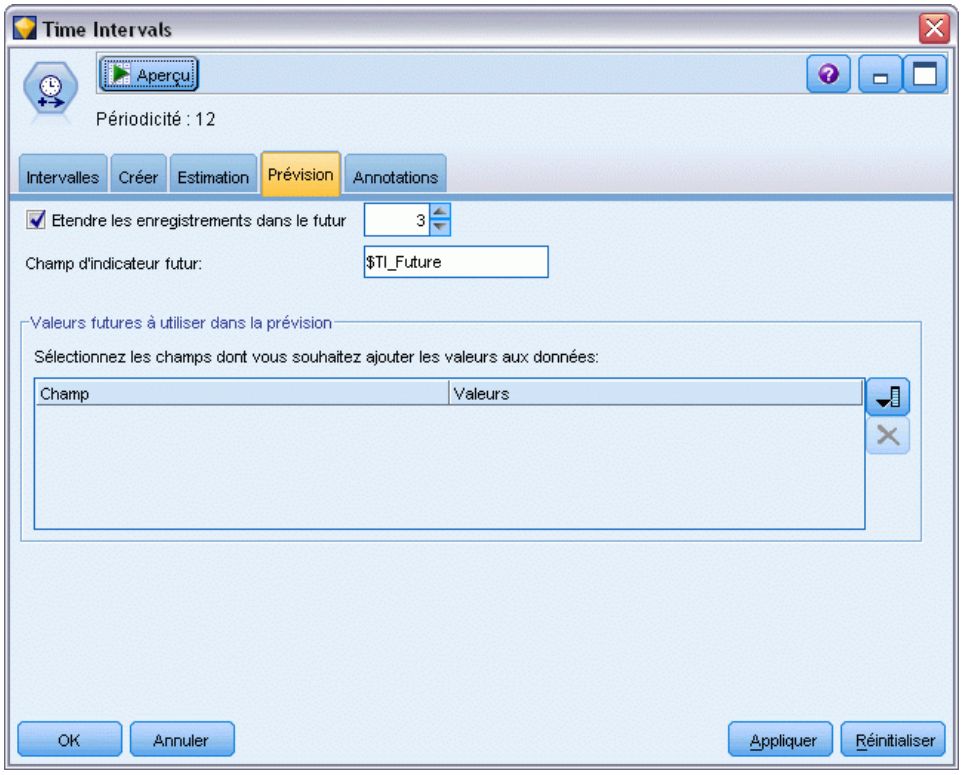
  Remarque : Le fichier de données a été mis à jour avec les données réelles relatives aux ventes réalisées de janvier à mars 2004, dans les lignes 61 à 63.

- Dans le flux, ouvrez le noeud Intervalles de temps.

- Cliquez sur l'onglet Prévision.

▶ Vérifiez que l'option Etendre les enregistrements dans le futur a pour valeur 3.

Figure 14-24
Vérification du paramétrage de la période prévisionnelle

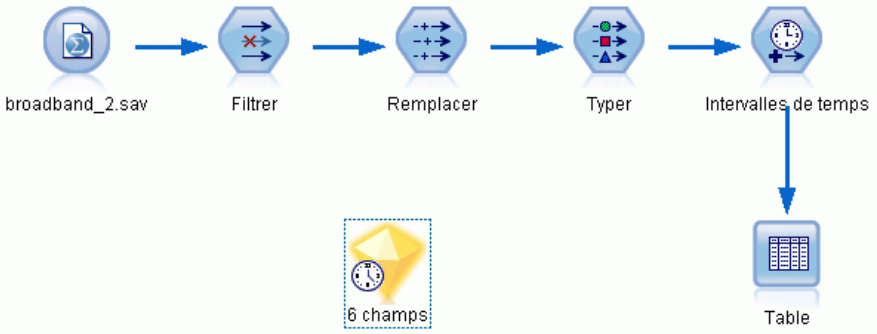


Extraction du modèle sauvegardé

▶ Dans le menu IBM® SPSS® Modeler, choisissez l'option Insérer > Noeud depuis le fichier, puis sélectionnez le fichier *TSmodel.nod* dans le dossier *Demos* (ou utilisez le modèle de séries temporelles que vous avez enregistré dans le premier exemple de séries temporelles).

Ce fichier contient les modèles de séries temporelles issus de l'exemple précédent. L'opération d'insertion place le nugget de modèle Séries temporelles correspondant sur l'espace de travail.

Figure 14-25
Ajout du nugget de modèle

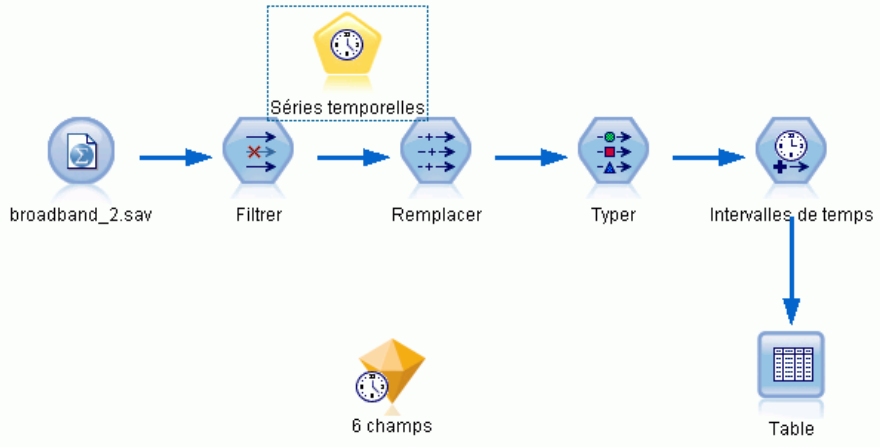


Génération d'un noeud de modélisation

- Ouvrez le nugget de modèle Séries temporelles et sélectionnez l'option Générer > Générer le noeud de modélisation.

Cette opération place un noeud de modélisation Séries temporelles sur l'espace de travail.

Figure 14-26
Génération d'un noeud de modélisation à partir du nugget de modèle



Génération d'un nouveau modèle

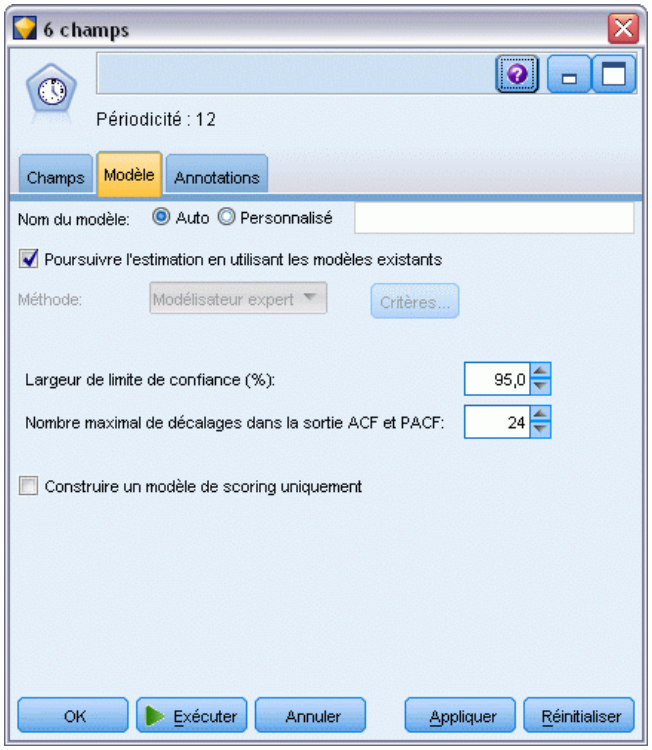
- Fermez le nugget de modèle Séries temporelles puis supprimez-le de l'espace de travail.

  L'ancien modèle a été construit à partir de 60 lignes de données. Vous devez générer un nouveau modèle à partir des données de vente mises à jour (63 lignes).

- Reliez au flux le noeud de construction Séries temporelles nouvellement généré.

Figure 14-27
Association du noeud de modélisation au flux

Figure 14-28
Réutilisation des paramètres stockés pour le modèle de séries temporelles



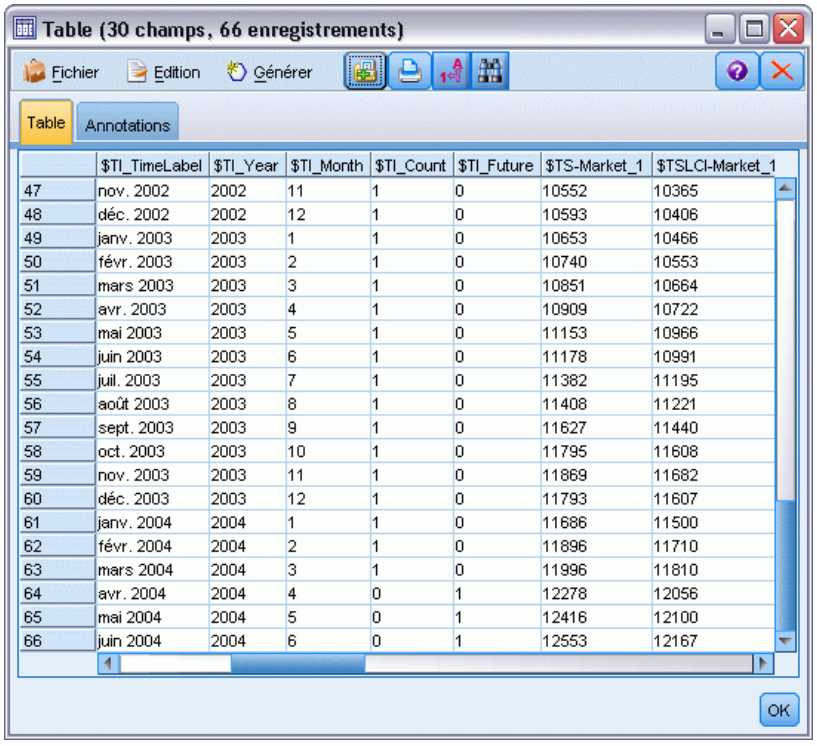
- Ouvrez le noeud Séries temporelles.

- Dans l'onglet Modèle, vérifiez que l'option Poursuivre l'estimation à l'aide des modèles existants est activée.

- Cliquez sur Exécuter pour mettre un nouveau nugget de modèle sur l'espace de travail et dans la palette Modèles.

Examen du nouveau modèle

Figure 14-29
Table indiquant la nouvelle prévision

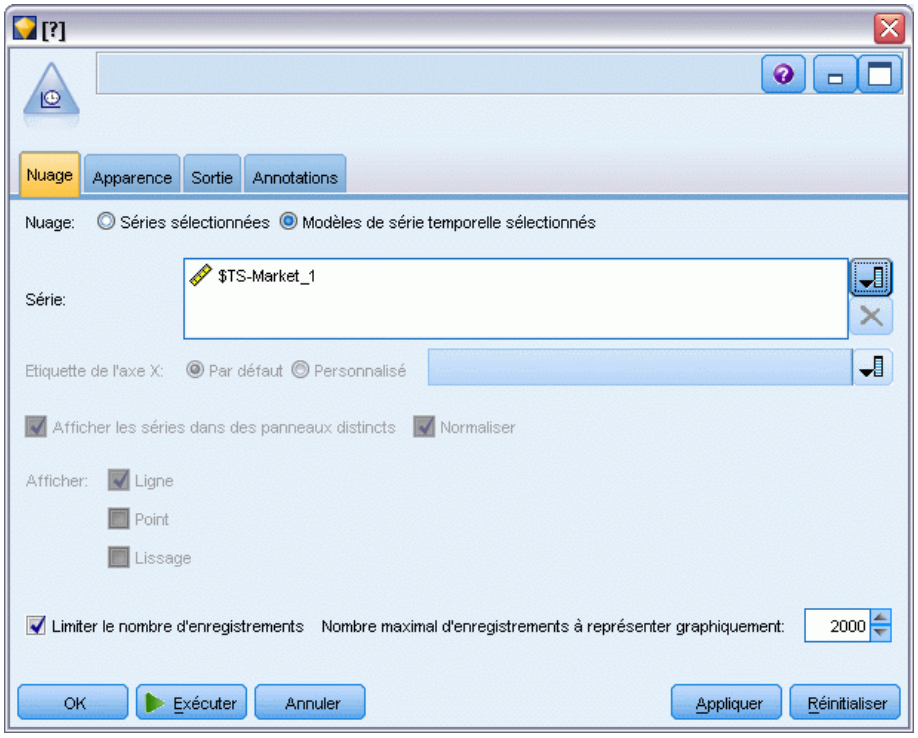


	$TI_TimeLabel	$TI_Year	$TI_Month	$TI_Count	$TI_Future	$TS-Market_1	$TSLCI-Market_1
47	nov. 2002	2002	11	1	0	10552	10365
48	déc. 2002	2002	12	1	0	10593	10406
49	janv. 2003	2003	1	1	0	10653	10466
50	févr. 2003	2003	2	1	0	10740	10553
51	mars 2003	2003	3	1	0	10851	10664
52	avr. 2003	2003	4	1	0	10909	10722
53	mai 2003	2003	5	1	0	11153	10966
54	juin 2003	2003	6	1	0	11178	10991
55	juil. 2003	2003	7	1	0	11382	11195
56	août 2003	2003	8	1	0	11408	11221
57	sept. 2003	2003	9	1	0	11627	11440
58	oct. 2003	2003	10	1	0	11795	11608
59	nov. 2003	2003	11	1	0	11869	11682
60	déc. 2003	2003	12	1	0	11793	11607
61	janv. 2004	2004	1	1	0	11686	11500
62	févr. 2004	2004	2	1	0	11896	11710
63	mars 2004	2004	3	1	0	11996	11810
64	avr. 2004	2004	4	0	1	12278	12056
65	mai 2004	2004	5	0	1	12416	12100
66	juin 2004	2004	6	0	1	12553	12167

- Reliez un noeud Table au nouveau noeud de modèle Séries temporelles sur l'espace de travail.
- Ouvrez le noeud Table, puis cliquez sur Exécuter.

  Le nouveau modèle effectue toujours des prévisions sur trois mois car vous réutilisez les paramètres stockés. Néanmoins, cette fois, la prévision porte sur la période d'avril à juin car la période d'estimation (spécifiée dans le noeud Intervalles de temps) ne s'achève plus en janvier mais en mars.

Figure 14-30
Spécification des champs à tracer



- Reliez un noeud de graphique Tracé horaire au nugget de modèle de séries temporelles.

  Cette fois, nous allons utiliser l'affichage Tracé horaire conçu spécifiquement pour les modèles de séries temporelles.

- Dans l'onglet Nuage, sélectionnez l'option Modèles de série temporelle sélectionnés.

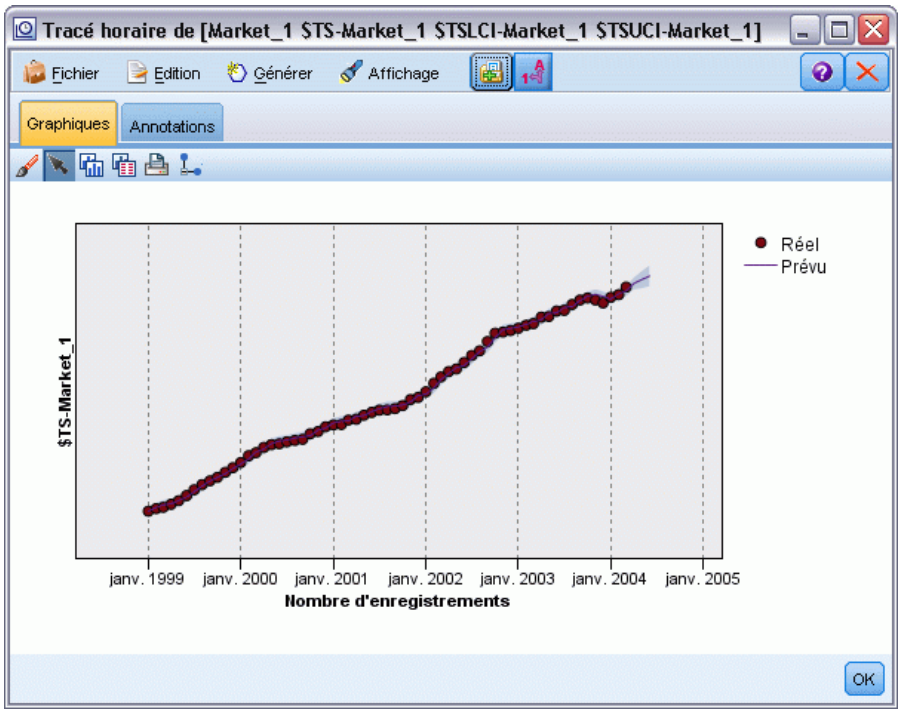
- Dans la liste Série, cliquez sur le bouton de sélection de champ, sélectionnez le champ *$TS-Market_1*, puis cliquez sur OK pour l'ajouter à la liste.

- Cliquez sur Exécuter.

  Vous disposez maintenant d'un graphique qui illustre les ventes réelles pour *Market_1* jusqu'à mars 2004, ainsi que les ventes prévisionnelles (Prévisions) et l'intervalle de confiance (indiqué par la zone ombrée bleue) jusqu'à juin 2004.

Comme dans le premier exemple, les valeurs prévisionnelles suivent étroitement les données réelles sur toute la période, ce qui indique une fois encore que vous disposez d'un modèle adéquat.

Figure 14-31
Prévision étendue jusqu'à juin



Récapitulatif

Vous avez appris à appliquer des modèles enregistrés afin d'étendre les prévisions antérieures lorsque de nouvelles données actuelles sont disponibles, et cela sans recréer vos modèles. Bien entendu, si vous avez une bonne raison de penser qu'un modèle a évolué, vous devez le recréer.

Chapitre
15

Prévision des ventes sur catalogue (séries temporelles)

Une société de vente sur catalogue souhaite prévoir les ventes mensuelles de sa ligne de vêtements pour hommes, en fonction des données des ventes réalisées au cours des 10 dernières années.

Cet exemple utilise le flux intitulé *catalog_forecast.str*, qui référence le fichier de données *catalog_seasfac.sav*. Ces fichiers sont disponibles dans le répertoire *Demos* de n'importe quelle installation de IBM® SPSS® Modeler. Ce dossier est accessible à partir du groupe de programmes IBM® SPSS® Modeler dans le menu Démarrer de Windows. Le fichier *catalog_forecast.str* se trouve dans le répertoire des *flux*.

Dans un exemple précédent, nous avons vu comment vous pouvez laisser le modélisateur expert choisir à votre place le modèle le plus approprié pour les séries temporelles. L'heure est venue d'examiner de plus près les deux méthodes disponibles vous permettant de choisir vous-même un modèle, à savoir le lissage exponentiel et l'ARIMA.

Pour choisir un modèle approprié, il est d'abord recommandé de tracer les séries temporelles. L'inspection visuelle d'une série temporelle facilite souvent le choix à réaliser. Vous devez notamment vous poser les questions suivantes :

- La série présente-t-elle une tendance globale ? Si tel est le cas, la tendance est-elle constante ou s'atténue-t-elle avec le temps ?
- La série présente-t-elle des effets saisonniers ? Si tel est le cas, les fluctuations saisonnières croissent-elles avec le temps ou apparaissent-elles constantes sur des périodes successives ?

Création du flux

▶ Créez un nouveau flux, puis ajoutez un noeud source Statistics pointant vers le fichier *catalog_seasfac.sav*.

Figure 15-1
Prévision des ventes sur catalogue

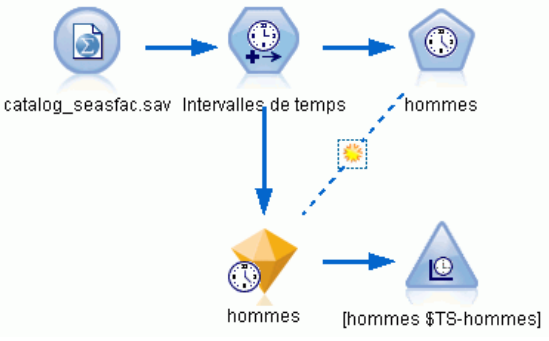
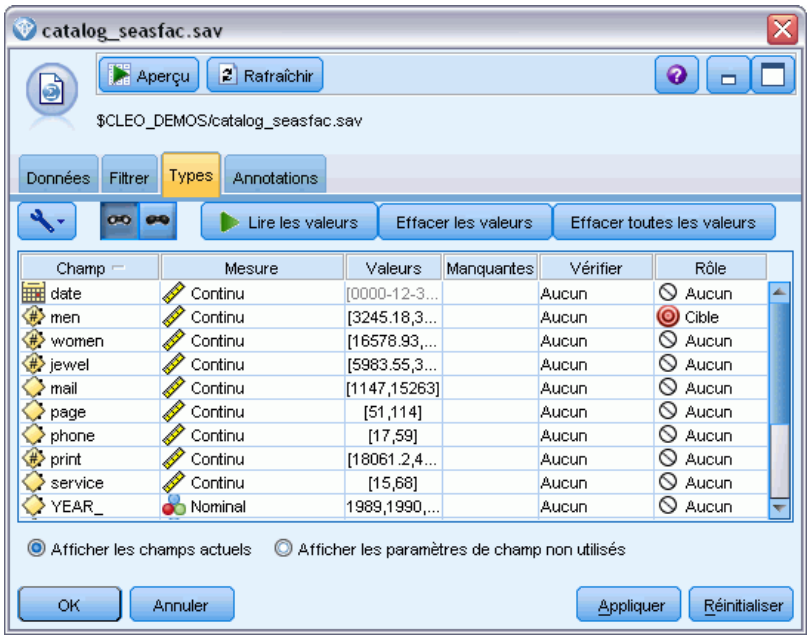
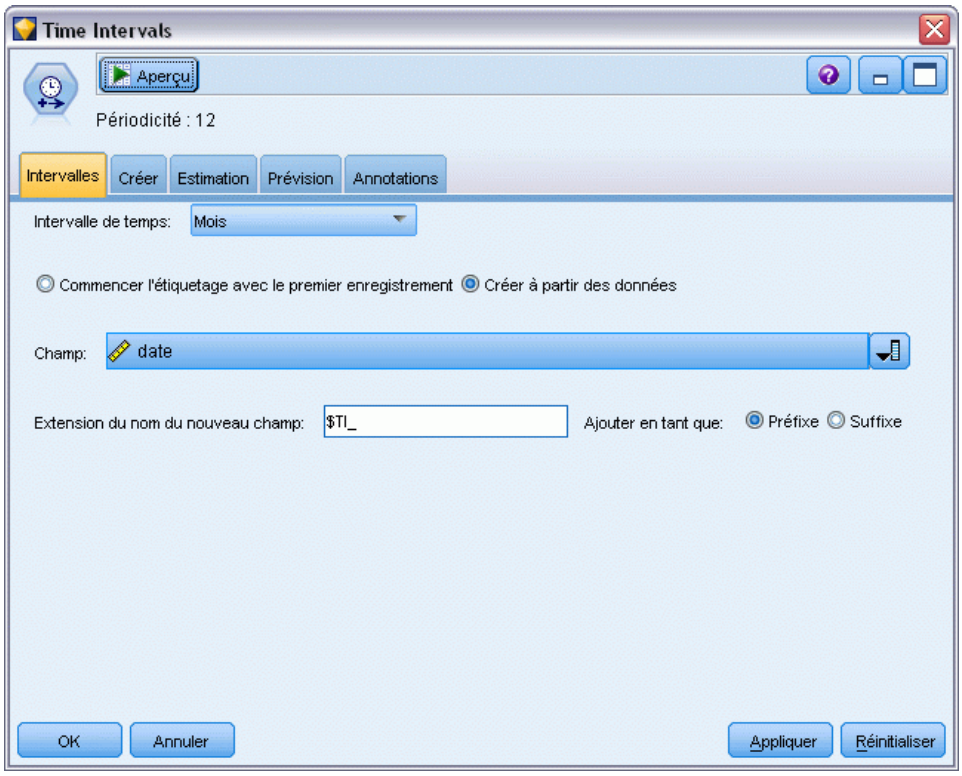


Figure 15-2
Spécification du champ cible



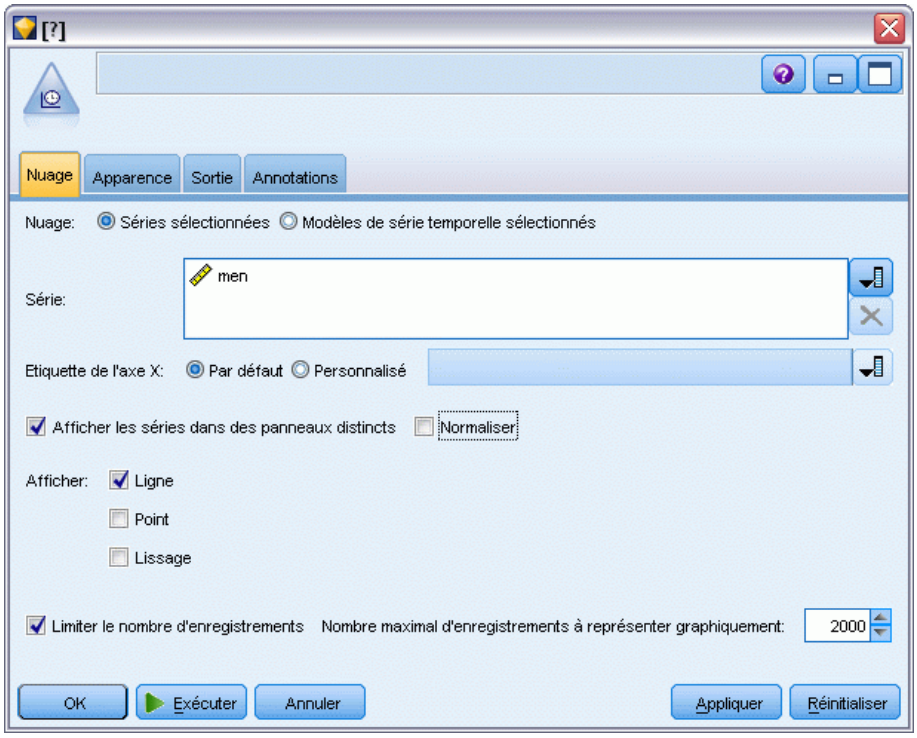
- Ouvrez le noeud source IBM® SPSS® Statistics, puis cliquez sur l'onglet Types.
- Cliquez sur Lire les valeurs, puis sur OK.
- Cliquez sur la colonne *Rôle* du champ *men*, puis définissez le rôle sur Cible.
- Définissez le rôle de tous les autres champs sur Aucun, puis cliquez sur OK.

Figure 15-3
Définition de l'intervalle de temps



- Reliez un noeud Intervalles de temps au noeud source SPSS Statistics.
- Ouvrez le noeud Intervalles de temps et définissez l'option Intervalle de temps sur Mois.
- Sélectionnez Créer à partir des données.
- Définissez l'option Champ sur Date, puis cliquez sur OK.

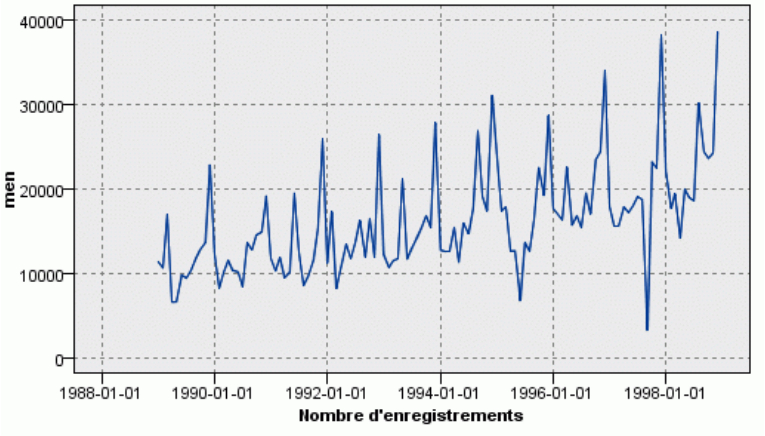
Figure 15-4
Traçage de la série temporelle



- Reliez un noeud Tracé horaire au noeud Intervalles de temps.
- Dans l'onglet Nuage, ajoutez men à la liste Série.
- Désélectionnez la case Normaliser.
- Cliquez sur Exécuter.

Examen des données

Figure 15-5
Ventes réelles des vêtements pour hommes



La série indique une tendance ascendante générale ; en d'autres termes, les valeurs de la série ont tendance à augmenter dans le temps. La tendance ascendante semble constante, ce qui indique une tendance linéaire.

En outre, la série présente un schéma saisonnier caractérisé par des ventes élevées en décembre, comme l'indiquent les lignes verticales du graphique. Les variations saisonnières semblent croître avec la tendance ascendante de la série, ce qui suggère la présence d'effets saisonniers multiplicatifs plutôt qu'additifs.

▶ Cliquez sur OK pour fermer le graphique.

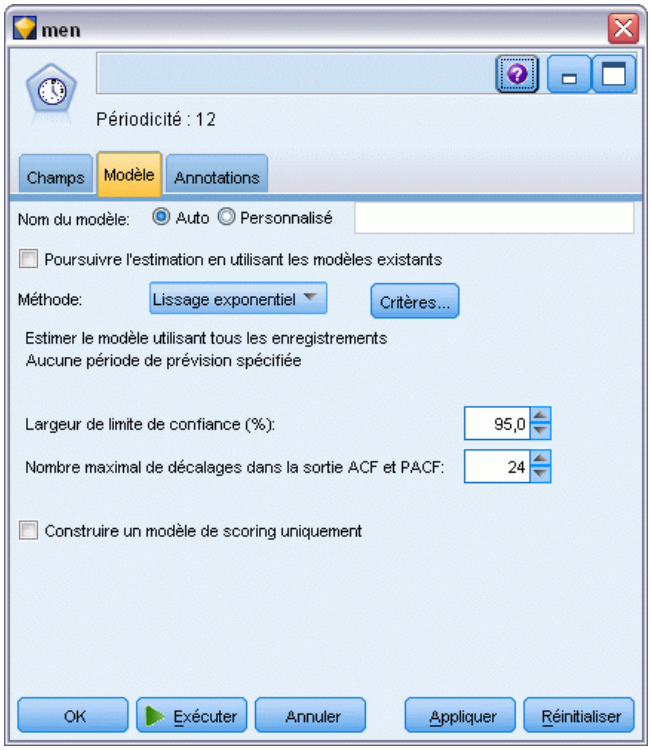
Les caractéristiques de la série étant identifiées, vous pouvez essayer de la modéliser. La méthode de lissage exponentiel permet de prévoir les séries qui présentent une tendance et/ou des effets saisonniers. Comme nous l'avons vu, les données présentent les deux caractéristiques.

Lissage exponentiel

La création d'un modèle de lissage exponentiel offrant le meilleur ajustement implique la détermination du type de modèle (à savoir, si le modèle doit inclure la tendance et/ou les effets saisonniers), puis l'obtention des paramètres offrant le meilleur ajustement pour le modèle choisi.

Le graphique des ventes de vêtements pour hommes dans le temps suggérait un modèle avec à la fois une composante de tendance linéaire et une composante d'effets saisonniers multiplicatifs. Cela implique un modèle de Winters. Toutefois, dans un premier temps, nous explorerons un modèle simple (sans tendance, ni effets saisonniers), puis un modèle de Holt (incorporant une tendance linéaire, mais aucun effet saisonnier). Au terme de cette opération, vous saurez identifier un modèle qui ne constitue pas un ajustement aux données adéquat, compétence essentielle pour la création réussie de modèles.

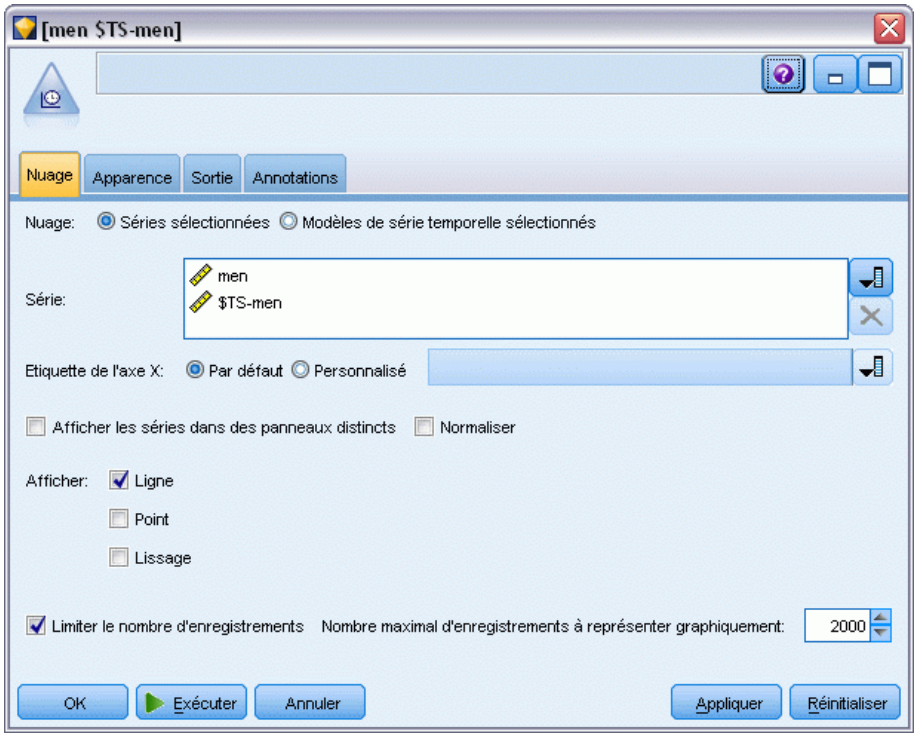
Figure 15-6
Spécification du lissage exponentiel



Nous allons commencer avec un modèle de lissage exponentiel simple.

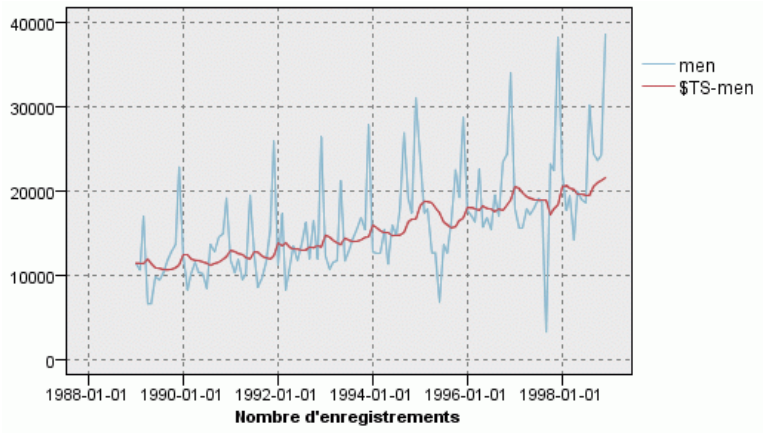
- Reliez un noeud Séries temporelles au noeud Intervalles de temps.
- Dans l'onglet Modèle, définissez l'option Méthode sur Lissage exponentiel.
- Cliquez sur Exécuter pour créer le nugget de modèle.

Figure 15-7
Traçage du modèle de séries temporelles



- Reliez un noeud Tracé horaire au nugget de modèle.
- Dans l'onglet Nuage, ajoutez *men* et *$TS-men* à la liste Série.
- Désélectionnez les cases Afficher les séries dans des panneaux distincts et Normaliser.
- Cliquez sur Exécuter.

Figure 15-8
Modèle de lissage exponentiel simple

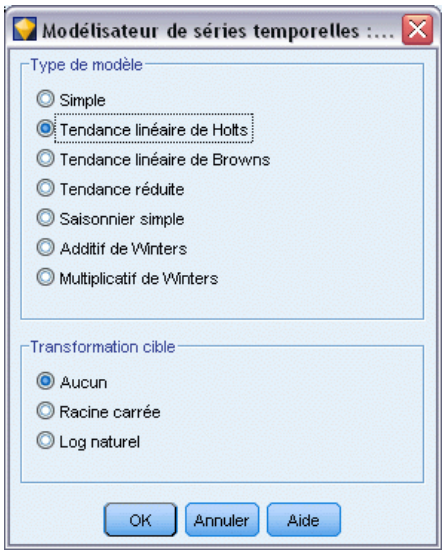


Le diagramme men représente les données réelles, tandis que $TS-men désigne le modèle de séries temporelles.

Bien que le modèle simple présente, en fait, une tendance ascendante progressive (et plutôt lourde), il ne prend aucunement en compte les effets saisonniers. Vous pouvez exclure ce modèle en toute sécurité.

► Cliquez sur OK pour fermer la fenêtre du tracé horaire.

Figure 15-9
Sélection du modèle de Holt

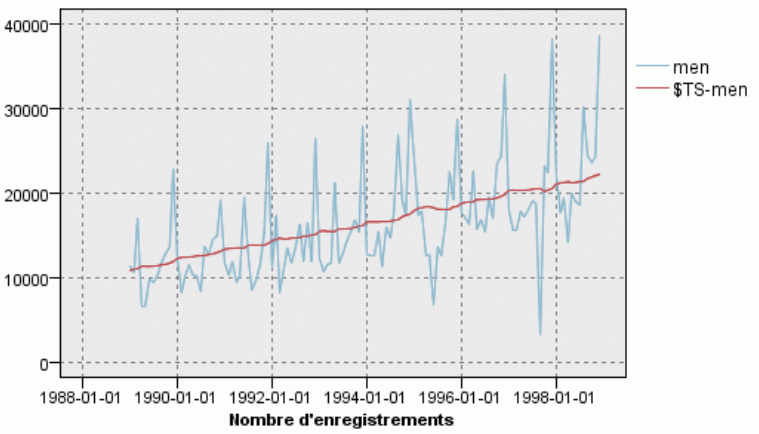


Essayons le modèle linéaire de Holt. Celui-ci devrait au moins mieux modéliser la tendance que le modèle simple, bien que lui aussi ne soit probablement pas en mesure de capturer les effets saisonniers.

► Rouvrez le noeud Séries temporelles.

- Dans l'onglet Modèle, l'option Lissage exponentiel étant toujours sélectionnée en guise de méthode, cliquez sur Critères.
- Dans la boîte de dialogue Critères de lissage exponentiel, sélectionnez l'option Tendance linéaire de Holts.
- Cliquez sur OK pour fermer la boîte de dialogue.
- Cliquez sur Exécuter pour recréer le nugget de modèle.
- Réouvrez le noeud Tracé horaire et cliquez sur Exécuter.

Figure 15-10
Modèle de tendance linéaire de Holt

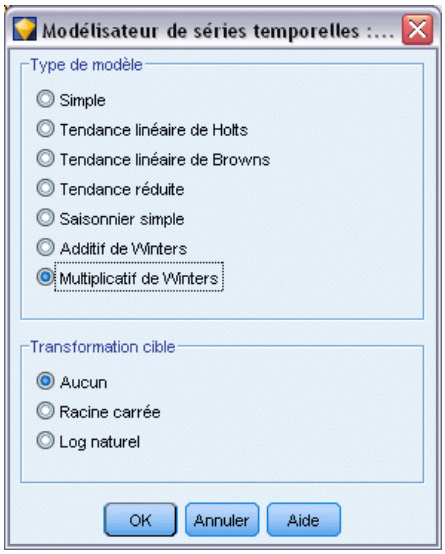


Le modèle de Holt affiche une tendance ascendante plus lisse que le modèle simple, mais ne prend aucunement en compte les effets saisonniers ; par conséquent, vous pouvez également le supprimer.

- Fermez la fenêtre du tracé horaire.

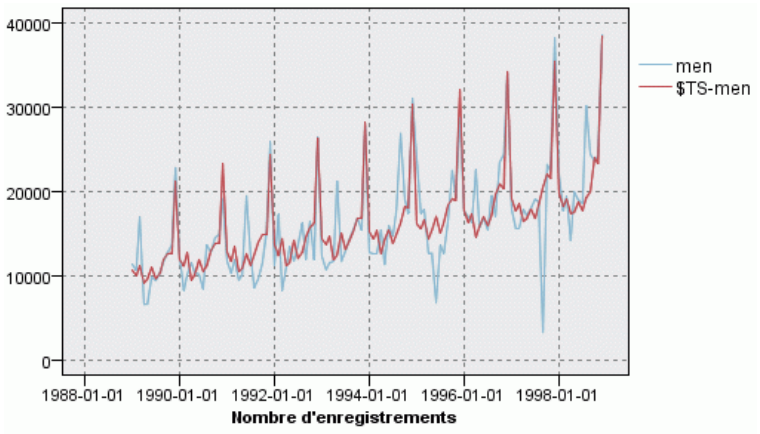
Vous vous souvenez peut-être que le graphique initial des ventes de vêtements pour hommes dans le temps suggérait un modèle incorporant une tendance linéaire et des effets saisonniers multiplicatifs. Par conséquent, le modèle de Winters pourrait être plus approprié.

Figure 15-11
Sélection du modèle de Winters



- Rouvrez le noeud Séries temporelles.
- Dans l'onglet Modèle, l'option Lissage exponentiel étant toujours sélectionnée en guise de méthode, cliquez sur Critères.
- Dans la boîte de dialogue Critères de lissage exponentiel, sélectionnez l'option Multiplicatif de Winters.
- Cliquez sur OK pour fermer la boîte de dialogue.
- Cliquez sur Exécuter pour recréer le nugget de modèle.
- Ouvrez le noeud Tracé horaire et cliquez sur Exécuter.

Figure 15-12
Modèle multiplicatif de Winters



Ce modèle semble plus approprié car il reflète à la fois la tendance et les effets saisonniers des données.

L'ensemble de données couvre une période de 10 années et comprend 10 pics saisonniers, se produisant au mois de décembre de chaque année. Les 10 pics présents dans les résultats prévus correspondent parfaitement aux 10 pics annuels dont font état les données réelles.

Toutefois, les résultats soulignent les limites de la procédure de lissage exponentiel. L'observation des pointes ascendantes et descendantes révèle l'existence d'une structure significative non expliquée.

Si vous êtes essentiellement intéressé par la modélisation d'une tendance à long terme avec variation saisonnière, le lissage exponentiel peut s'avérer un choix adéquat. Pour modéliser une structure plus complexe telle que celle-ci, nous devons envisager l'utilisation de la procédure ARIMA.

ARIMA

Grâce à la procédure ARIMA, vous pouvez créer un modèle ARIMA (AutoRegressive Integrated Moving Average) qui vous permet d'affiner la modélisation des séries temporelles. Les modèles ARIMA mettent à votre disposition des méthodes de modélisation des composantes de tendance et saisonnières plus élaborées que celles proposées par les modèles de lissage exponentiel et ils permettent d'inclure des variables de prévision dans le modèle.

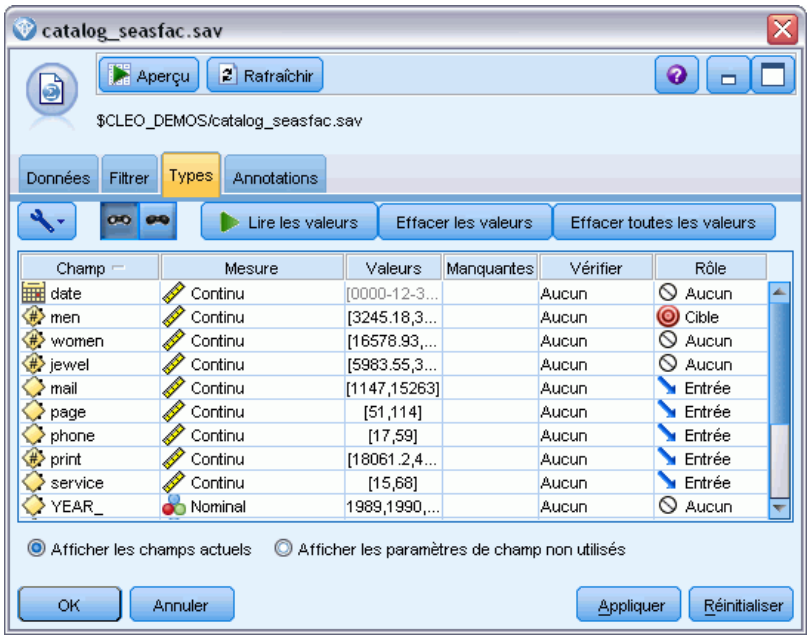
A travers l'exemple de la société de vente sur catalogue qui souhaite développer un modèle de prévision, nous avons vu comment celle-ci a collecté des données sur les ventes mensuelles de vêtements pour hommes, ainsi que plusieurs séries susceptibles de faciliter l'explication d'une partie de la variation des ventes. Parmi les variables indépendantes possibles figurent le nombre de catalogues envoyés par messagerie électronique, le nombre de pages dans le catalogue, le nombre de lignes téléphoniques dédiées à la prise de commandes, les frais de publicité imprimée et le nombre de conseillers du service à la clientèle.

L'une des variables indépendantes est-elle utile pour la prévision ? Un modèle comportant des variables indépendantes est-il réellement meilleur qu'un modèle qui en est dépourvu ? A l'aide de la procédure ARIMA, nous pouvons créer un modèle de prévision comportant des variables indépendantes et déterminer s'il existe une différence significative en matière de capacité prédictive par rapport au modèle de lissage exponentiel dépourvu de variable indépendante.

La méthode ARIMA vous permet d'affiner le modèle en spécifiant les ordres d'autorégression, de différenciation et de moyenne mobile, ainsi que les équivalents saisonniers de ces composantes. Etant donné que la détermination manuelle des meilleures valeurs pour ces composantes peut être un processus de longue durée impliquant une série d'essais et d'erreurs, nous allons, pour cet exemple, laisser le modélisateur expert choisir un modèle ARIMA à notre place.

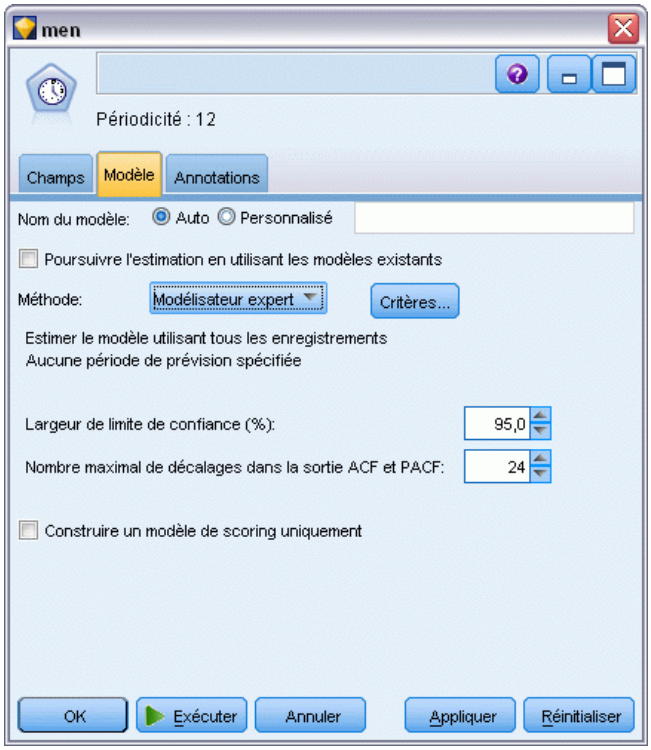
Nous allons essayer de créer un meilleur modèle en traitant certaines des autres variables de l'ensemble de données en tant que variables de prévision. Celles qui semblent le plus utile à inclure en tant que variables indépendantes sont le nombre de catalogues envoyés par messagerie électronique (*mail*), le nombre de pages dans le catalogue (*page*), le nombre de lignes téléphoniques dédiées à la prise des commandes (*phone*), les frais de publicité imprimée (*print*) et le nombre de conseillers du service à la clientèle (*service*).

Figure 15-13
Définition des champs variables indépendantes



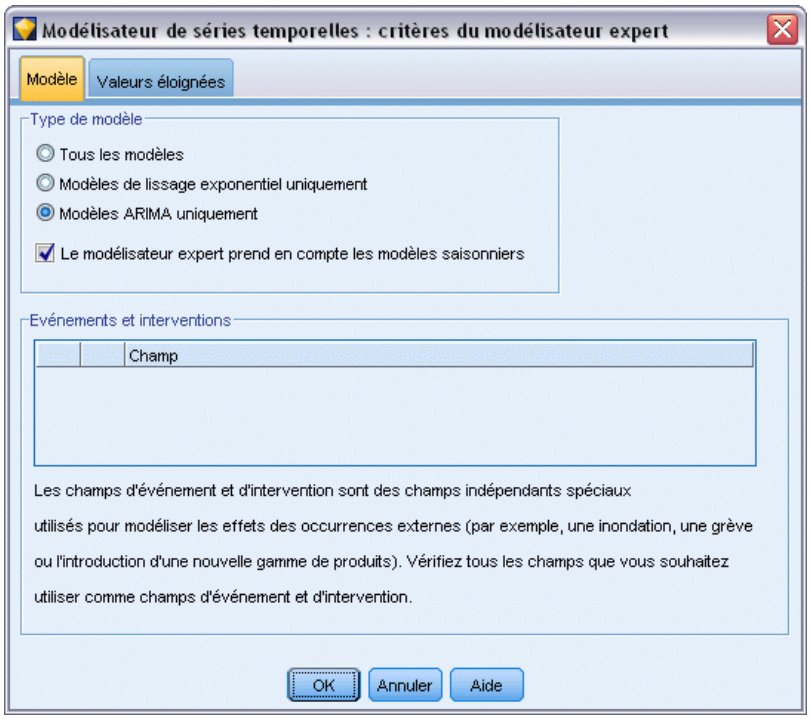
- Ouvrez le noeud source de fichier IBM® SPSS® Statistics.
- Dans l'onglet Types, définissez l'option *Rôle* des variables *mail*, *page*, *phone*, *print* et *service* sur Entrée.
- Vérifiez que la direction pour men est définie sur Cible et que tous les champs restants sont définis sur Aucun.
- Cliquez sur OK.

Figure 15-14
Sélection du modélisateur expert



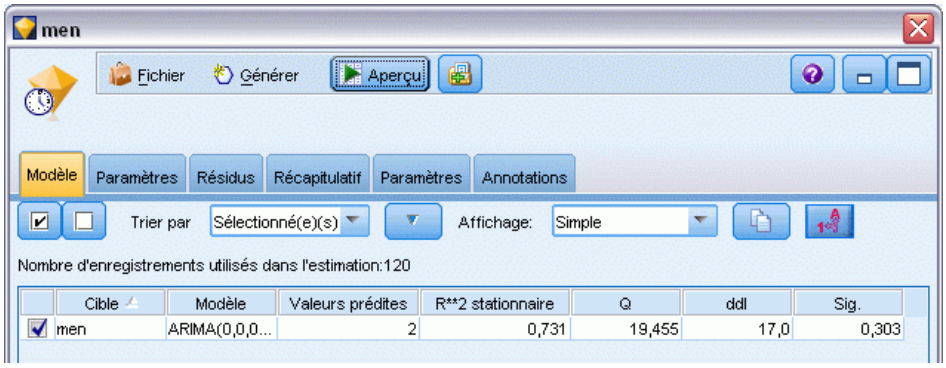
- Ouvrez le noeud Séries temporelles.
- Dans l'onglet Modèle, définissez l'option Méthode sur Modélisateur expert, puis cliquez sur Critères.

Figure 15-15
Sélection des modèles ARIMA uniquement



- Dans la boîte de dialogue Critères du modélisateur expert, sélectionnez l'option Modèles ARIMA uniquement, puis vérifiez que la case Le modélisateur expert prend en compte les modèles saisonniers est cochée.
- Cliquez sur OK pour fermer la boîte de dialogue.
- Cliquez sur Exécuter dans l'onglet Modèle pour recréer le nugget de modèle.

Figure 15-16
Le modélisateur expert choisit deux variables indépendantes



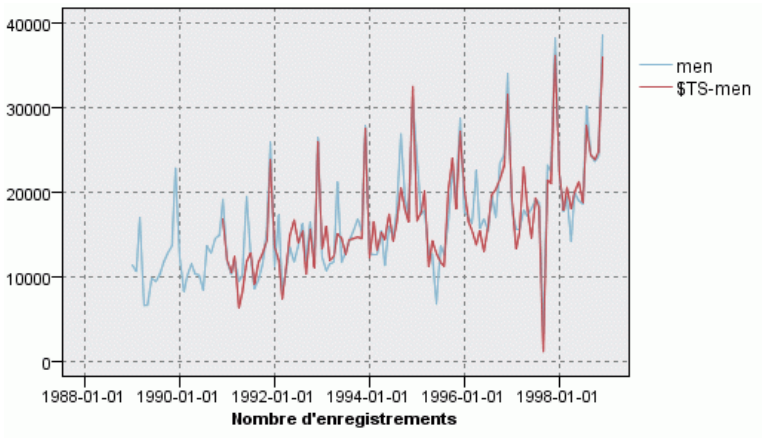
▶ Ouvrez le nugget de modèle.

Comme vous pouvez le constater, le modélisateur expert n'a choisi, parmi les cinq variables indépendantes spécifiées, que deux variables comme étant significatives pour le modèle.

▶ Cliquez sur OK pour fermer le nugget du modèle.

▶ Ouvrez le noeud Tracé horaire et cliquez sur Exécuter.

Figure 15-17
Modèle ARIMA avec variables indépendantes spécifiées



Ce modèle est meilleur que le précédent car il capture également la grande pointe descendante, ce qui en fait le meilleur ajustement constaté jusqu'à présent.

Nous pourrions essayer d'affiner davantage le modèle, mais toute amélioration à partir de ce point est susceptible d'être minimale. Nous avons démontré que le modèle ARIMA avec variables indépendantes est préférable ; par conséquent, utilisons le modèle que nous venons de créer. Dans le cadre de cet exemple, nous allons prévoir les ventes de l'année à venir.

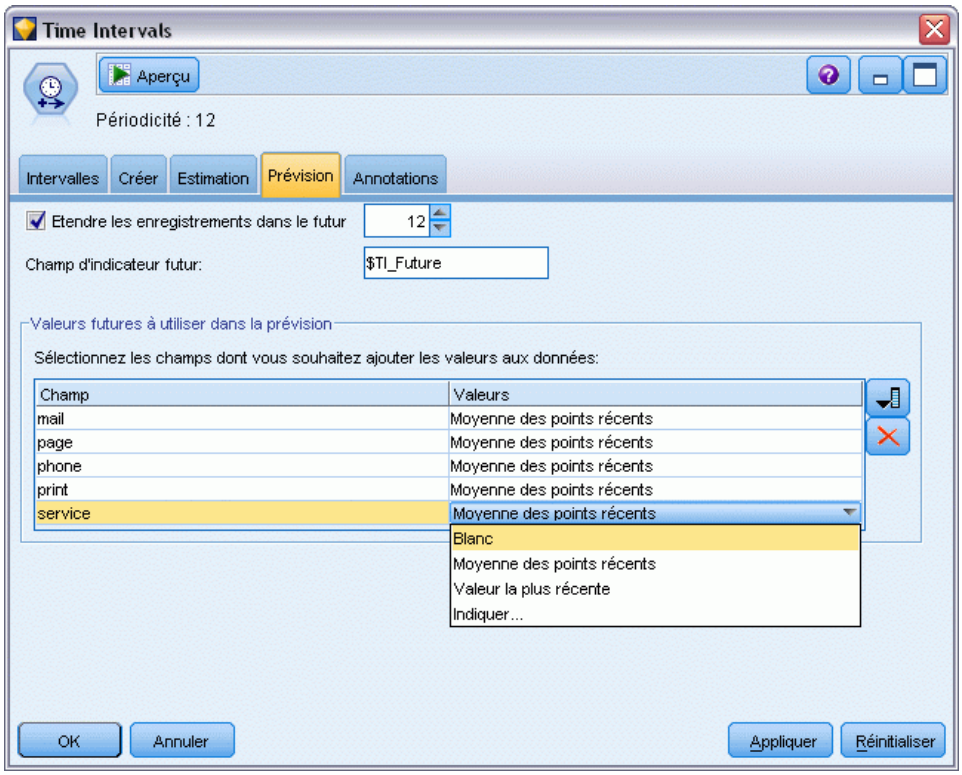
▶ Cliquez sur OK pour fermer la fenêtre du tracé horaire.

▶ Ouvrez le noeud Intervalles de temps, puis sélectionnez l'onglet *Prévision*.

▶ Cochez la case *Etendre les enregistrements dans le futur*, puis attribuez-lui la valeur 12.

L'utilisation de variables indépendantes pour la prévision nécessite la spécification de valeurs estimées pour les champs concernés par la période prévisionnelle, afin que le modélisateur puisse prévoir le champ cible avec davantage de précision.

Figure 15-18
Spécification des valeurs futures des champs variables indépendantes



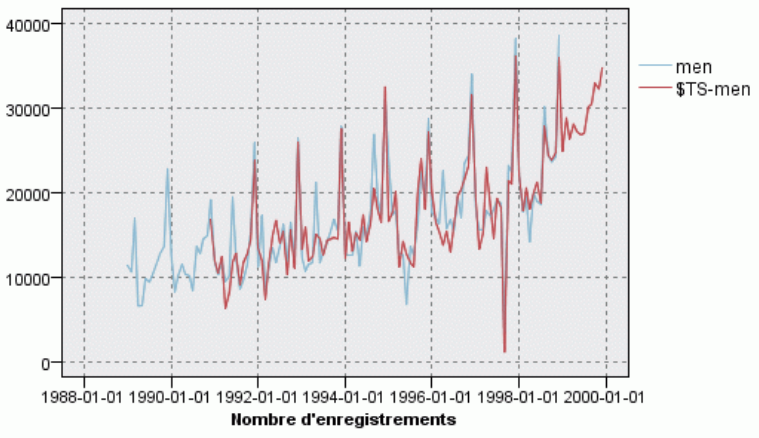
- Dans le groupe Valeurs futures à utiliser dans la prévision, cliquez sur le bouton de sélection de champ situé à droite de la colonne Valeurs.
- Dans la boîte de dialogue Sélectionner les champs, sélectionnez les variables mail à service, puis cliquez sur OK.

  Dans la réalité, vous spécifieriez les valeurs futures manuellement à ce stade, dans la mesure où ces cinq variables indépendantes concernent toutes des éléments dont vous avez le contrôle. Dans le cadre de cet exemple, nous allons utiliser l'une des fonctions prédéfinies, afin de ne pas avoir à spécifier 12 valeurs pour chaque variable indépendante. (Lorsque vous maîtriserez davantage cet exemple, vous pourrez recourir à différentes valeurs futures afin de déterminer leur impact sur le modèle.)
- Pour chaque champ tour à tour, cliquez sur le champ Valeurs afin d'afficher la liste des valeurs possibles, puis sélectionnez l'option Moyenne des points récents. Cette option calcule la moyenne des trois derniers points de données pour ce champ et elle utilise cette moyenne comme valeur estimée dans chaque cas.
- Cliquez sur OK.
- Ouvrez le noeud Séries temporelles et cliquez sur Exécuter pour recréer le nugget de modèle.

▶ Ouvrez le noeud Tracé horaire et cliquez sur Exécuter.

La prévision pour 1999 semble correcte : comme prévu, il existe un retour aux niveaux normaux de ventes après le pic de décembre, ainsi qu'une tendance ascendante régulière dans la seconde moitié de l'année, dont les ventes sont globalement très supérieures à celles de l'année précédente.

Figure 15-19
Prévision des ventes avec variables indépendantes spécifiées



Récapitulatif

Vous avez correctement modélisé une série temporelle complexe, en intégrant non seulement une tendance ascendante, mais aussi des effets saisonniers et d'autres variations. Vous avez également vu comment, à l'aide d'une série d'essais et d'erreurs, vous avez pu approcher petit à petit un modèle précis, que vous avez ensuite utilisé pour prévoir les ventes à venir.

Dans la pratique, vous devriez réappliquer le modèle car les données de ventes réelles sont mises à jour, par exemple, chaque mois ou chaque trimestre, et générer des prévisions actualisées. Pour plus d'informations, reportez-vous à la section Réapplication d'un modèle de séries temporelles dans le chapitre 14 sur p. 201.

Chapitre
16

Propositions aux clients (auto-apprentissage)

Le noeud de réponse Auto-formation permet de générer et de mettre à jour un modèle grâce auquel vous pouvez prévoir les offres les plus appropriées pour les clients et la probabilité d'acceptation des offres. Ces types de modèle sont les plus utiles dans la gestion de la relation client, notamment dans les applications marketing ou les centres d'appels.

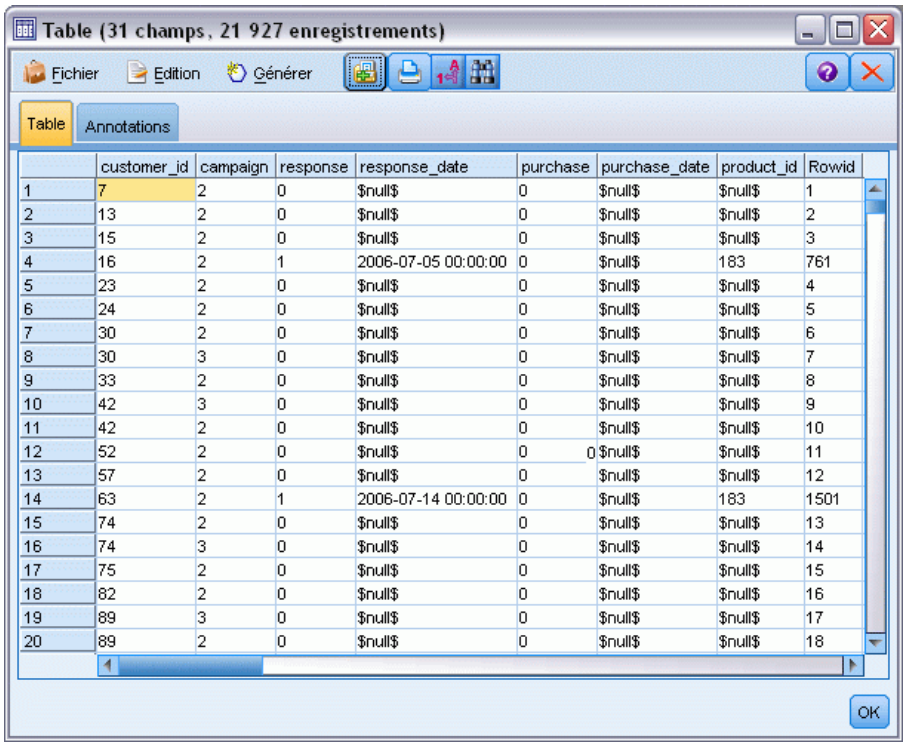
Cet exemple est basé sur un établissement bancaire fictif. Le service marketing souhaite obtenir des résultats plus rentables lors des prochaines campagnes en présentant à chaque client une offre de service financier adaptée. Plus précisément, l'exemple utilise un modèle de réponse d'auto-apprentissage pour identifier les caractéristiques des clients les plus susceptibles de répondre favorablement, en fonction d'offres et de réponses précédentes, et de promouvoir la meilleure offre existant à partir des résultats.

Cet exemple utilise le flux *pm_selflearn.str* qui fait référence aux fichiers de données *pm_customer_train1.sav*, *pm_customer_train2.sav* et *pm_customer_train3.sav*. Ces fichiers sont disponibles dans le dossier *Demos* de toutes les installations de IBM® SPSS® Modeler. Ce dossier est accessible à partir du groupe de programmes IBM® SPSS® Modeler dans le menu Démarrer de Windows. Le fichier *pm_selflearn.str* se trouve dans le dossier des *flux*.

Données existantes

L'entreprise dispose de données historiques retraçant les offres faites aux clients lors des campagnes précédentes, ainsi que les réponses à ces offres. Ces données incluent également des informations démographiques et financières qui peuvent être utilisées pour prévoir les taux de réponse pour différents clients.

Figure 16-1
Réponses aux offres précédentes

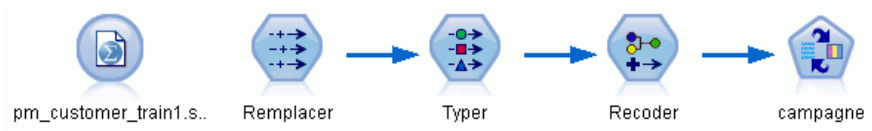


	customer_id	campaign	response	response_date	purchase	purchase_date	product_id	Rowid
1	7	2	0	$null$	0	$null$	$null$	1
2	13	2	0	$null$	0	$null$	$null$	2
3	15	2	0	$null$	0	$null$	$null$	3
4	16	2	1	2006-07-05 00:00:00	0	$null$	183	761
5	23	2	0	$null$	0	$null$	$null$	4
6	24	2	0	$null$	0	$null$	$null$	5
7	30	2	0	$null$	0	$null$	$null$	6
8	30	3	0	$null$	0	$null$	$null$	7
9	33	2	0	$null$	0	$null$	$null$	8
10	42	3	0	$null$	0	$null$	$null$	9
11	42	2	0	$null$	0	$null$	$null$	10
12	52	2	0	$null$	0	0 $null$	$null$	11
13	57	2	0	$null$	0	$null$	$null$	12
14	63	2	1	2006-07-14 00:00:00	0	$null$	183	1501
15	74	2	0	$null$	0	$null$	$null$	13
16	74	3	0	$null$	0	$null$	$null$	14
17	75	2	0	$null$	0	$null$	$null$	15
18	82	2	0	$null$	0	$null$	$null$	16
19	89	3	0	$null$	0	$null$	$null$	17
20	89	2	0	$null$	0	$null$	$null$	18

Création du flux

- Ajoutez un noeud source de fichier Statistics qui pointe sur *pm_customer_train1.sav*, dans le dossier *Demos* du répertoire d'installation de IBM® SPSS® Modeler.

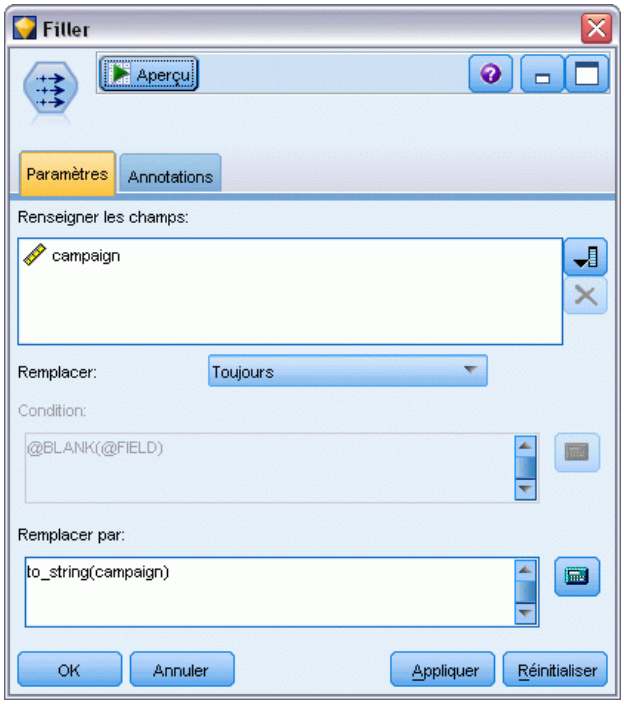
Figure 16-2
Exemple de flux MRAA



- Ajoutez un noeud Remplacer et saisissez campaign comme champ à renseigner.
- Sélectionnez Toujours comme type de champ Remplacer.

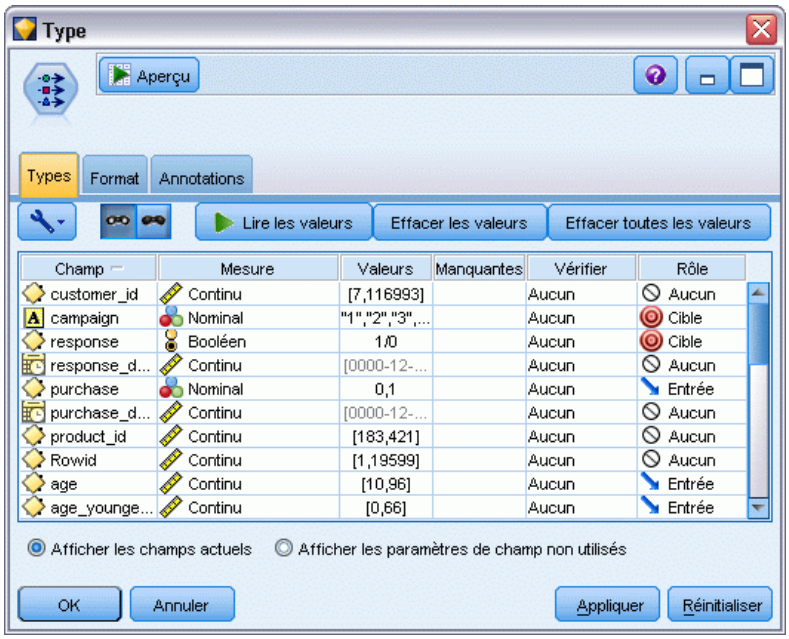
▶ Dans la zone de texte Remplacer par, entrez to_string(campaign) et cliquez sur OK.

Figure 16-3
Calcul d'un champ de campagne



▶ Ajoutez un noeud Typer et définissez l'option *Rôle* sur Aucun pour les champs *customer_id*, *response_date*, *purchase_date*, *product_id*, *Rowid*, et *X_random*.

Figure 16-4
Modification des paramètres du noeud Typer



▶ Définissez l'option *Rôle* sur Cible pour les champs *campaign* et *response*. Il s'agit des champs sur lesquels doivent reposer les prévisions.

Définissez la mesure sur Booléen pour le champ *response*.

▶ Cliquez sur Lire les valeurs, puis sur OK.

Comme les données de champ de campagne se présentent sous la forme d'une liste numérique (1, 2, 3 et 4), vous pouvez recodifier les champs pour obtenir des titres plus évocateurs.

▶ Ajoutez un noeud Recoder au noeud Typer.

▶ Dans le champ Recoder dans, sélectionnez Champ existant.

▶ Dans la liste Recoder le champ, sélectionnez campagne.

▶ Cliquez sur le bouton Obtenir ; les valeurs de campagne sont ajoutées à la colonne *Valeur d'origine*.

▶ Dans la colonne *Nouvelle valeur*, entrez les noms de campagne suivants sur les quatre premières lignes :

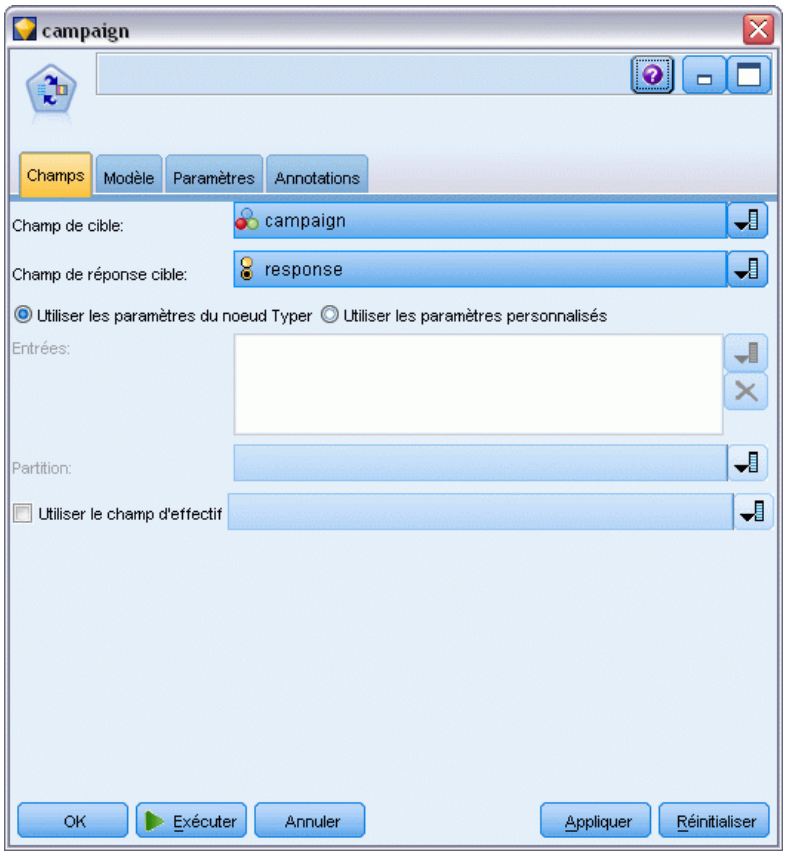
- Prêt hypothécaire
- Prêt automobile
- Epargne
- Retraite

▶ Cliquez sur OK.

Figure 16-5
Recodification des champs de campagne

▶ Reliez un noeud de modélisation MRAA au noeud Recoder. Dans l'onglet Champs, sélectionnez campaign pour le champ Cible et response pour le champ de réponse Cible.

Figure 16-6
Sélection de la cible et de la réponse cible

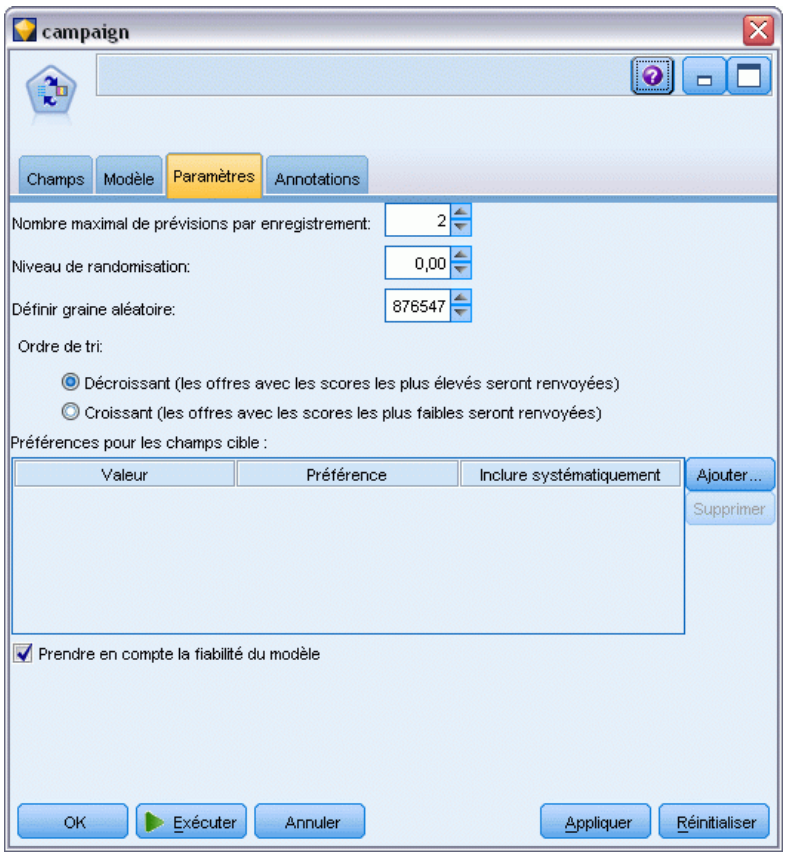


▶ Dans le champ Nombre maximal de prédictions par enregistrement de l'onglet Paramètres, réduisez la valeur à 2.

Deux offres seront ainsi identifiées comme présentant la plus forte probabilité d'acceptation pour chaque client.

▶ Assurez-vous de sélectionner Prendre en compte la fiabilité du modèle et cliquez sur Exécuter.

Figure 16-7
Paramètres du noeud MRAA



Navigation dans le modèle

- Ouvrez le nugget de modèle. Initialement, l'onglet Modèle indique l'estimation de la précision des prévisions pour chaque offre, ainsi que l'importance relative de chaque variable indépendante pour estimer le modèle.

  Pour afficher la corrélation de chaque variable indépendante avec la variable cible, choisissez Association avec réponse dans la liste Affichage du panneau de droite.

- Pour vous déplacer entre les quatre offres auxquelles les prédictions s'appliquent, sélectionnez l'offre appropriée dans la liste Affichage du panneau de gauche.

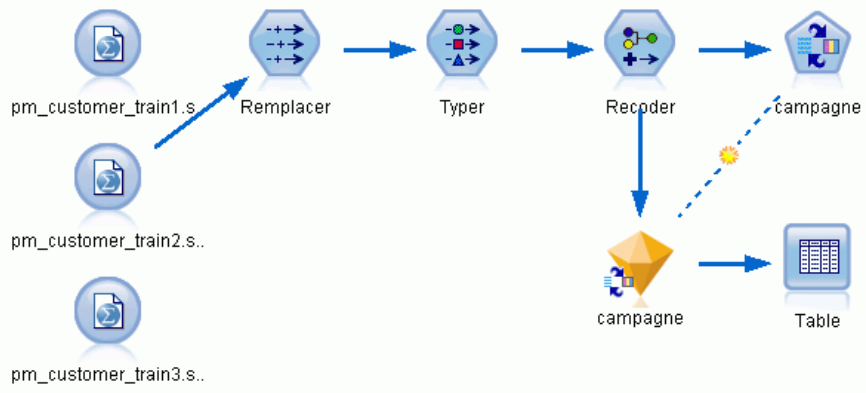
Figure 16-8
Nugget de modèle MRAA

- Fermez la fenêtre du nugget de modèle.

- Dans l'espace de travail, déconnectez le noeud source IBM® SPSS® Statistics pointant vers le fichier *pm_customer_train1.sav*.

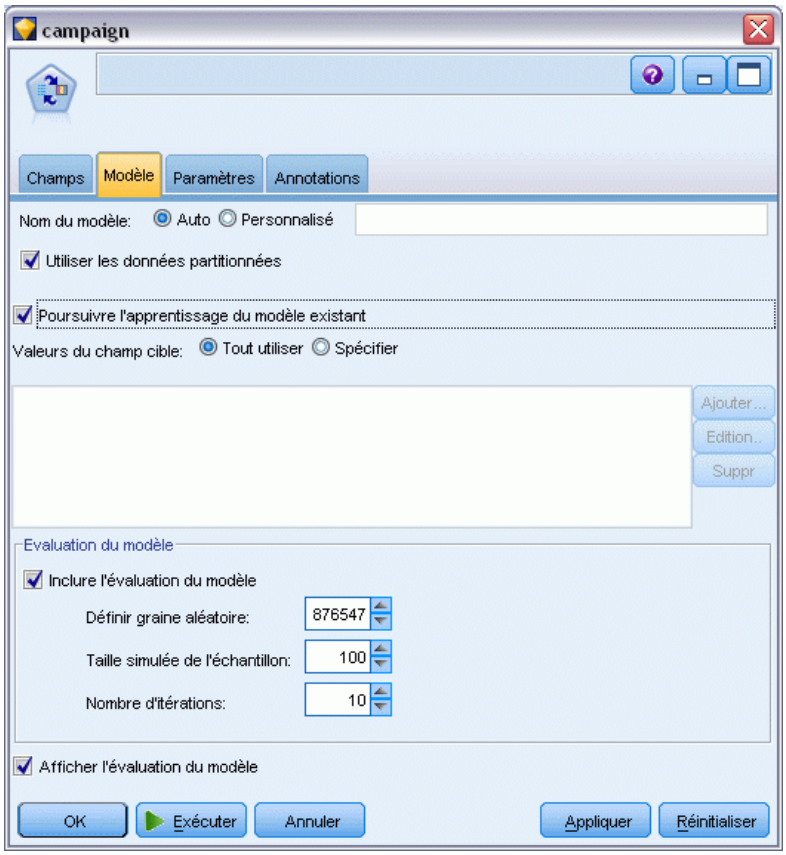
- Ajoutez un noeud source de fichier Statistics pointant vers le fichier *pm_customer_train2.sav* situé dans le dossier *Demos* du répertoire d'installation de IBM® SPSS® Modeler, puis connectez-le au noeud Remplacer.

Figure 16-9
Association d'une deuxième source de données au flux MRAA



- Dans l'onglet Modèle du noeud MRAA, sélectionnez Poursuivre le modèle d'apprentissage existant.

Figure 16-10
Poursuite de l'apprentissage du modèle

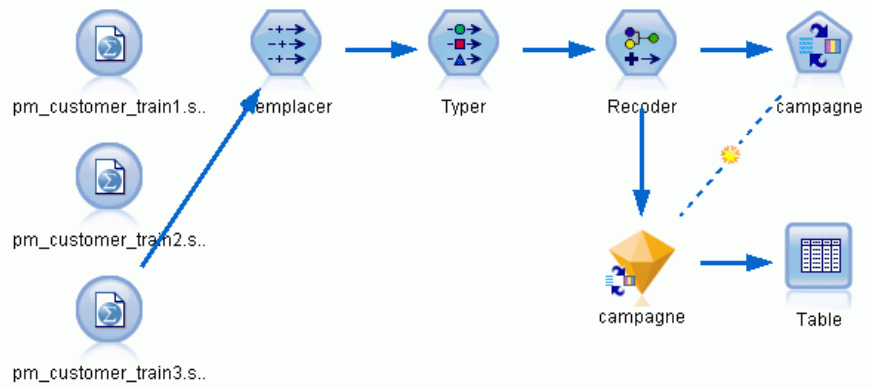


- Cliquez sur Exécuter pour recréer le nugget de modèle. Pour en afficher les détails, double-cliquez sur le nugget dans l'espace de travail.

  L'onglet Modèle affiche les estimations révisées de la précision des prévisions pour chaque offre.

- Ajoutez un noeud source de fichier Statistics pointant vers le fichier *pm_customer_train3.sav* situé dans le dossier *Demos* du répertoire d'installation de SPSS Modeler, puis connectez-le au noeud Remplacer.

Figure 16-11
Association d'une troisième source de données au flux MRAA

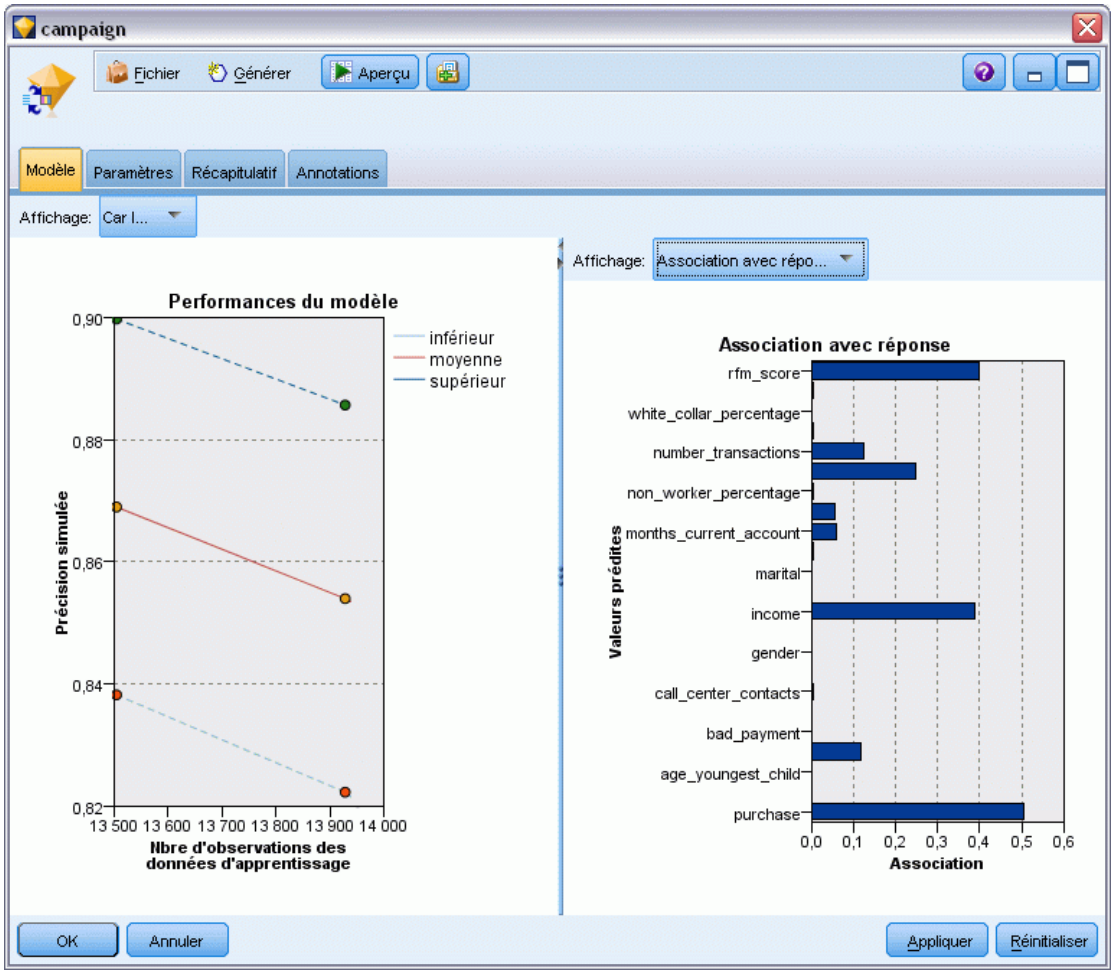


- Cliquez sur Exécuter pour recréer une nouvelle fois le nugget de modèle. Pour en afficher les détails, double-cliquez sur le nugget dans l'espace de travail.

- L'onglet Modèle affiche maintenant les estimations finales de la précision des prévisions pour chaque offre.

Comme vous pouvez le constater, la précision moyenne a légèrement baissé (de 86,9% à 85,4%) à la suite de l'ajout des sources de données supplémentaires. Cela étant, cette faible variation peut s'expliquer par la présence de petites anomalies dans les données disponibles.

Figure 16-12
Nugget de modèle MRAA mis à jour

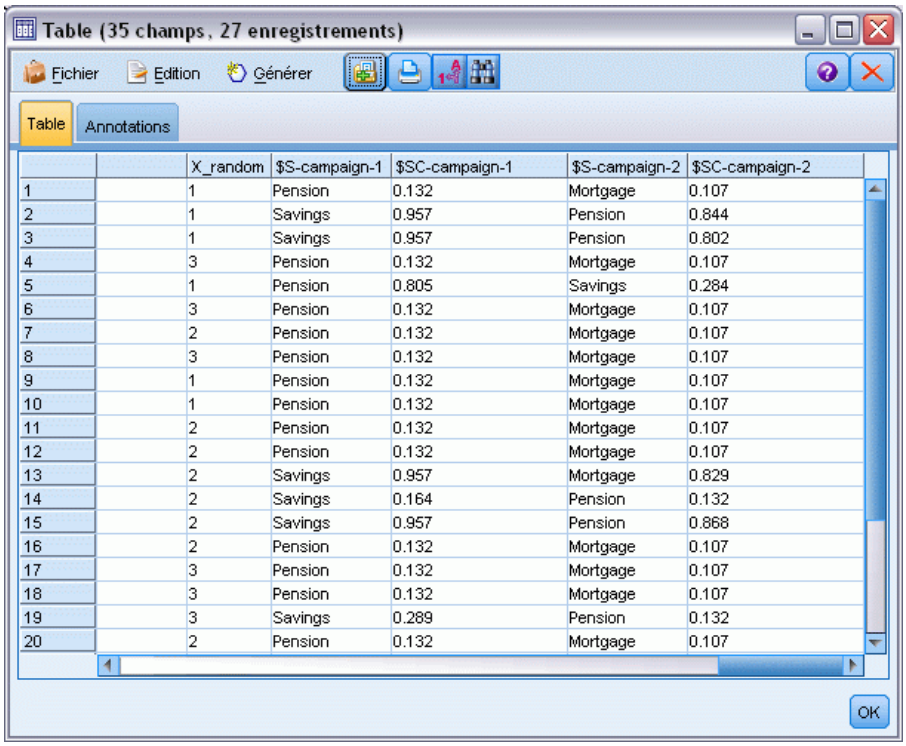


- Reliez un noeud Table au dernier (troisième) modèle généré, puis exécutez ce noeud.
- Faites défiler le tableau vers la droite. Les prévisions indiquent les offres qu'un client est le plus enclin à accepter et l'assurance de l'acceptation de ces offres, en fonction des détails concernant chaque client.

  Par exemple, selon la première ligne du tableau illustré, la certitude qu'un client ayant déjà contracté un prêt automobile accepte une éventuelle offre d'épargne retraite ne s'élève qu'à 13,2 % (indiqué par la valeur 0,132 dans la colonne *$SC-campaign-1*). Toutefois, les deuxième et troisième lignes présentent deux autres clients ayant souscrit ce même type de prêt ; dans leur cas,

il existe un taux de confiance de 95,7 % qu'ils ouvriraient un compte d'épargne suite à une offre, et un taux de confiance supérieur à 80% qu'ils accepteraient une offre d'épargne de retraite.

Figure 16-13
Résultat du modèle : prédictions d'offre et de confiance



		X_random	$S-campaign-1	$SC-campaign-1	$S-campaign-2	$SC-campaign-2
1		1	Pension	0.132	Mortgage	0.107
2		1	Savings	0.957	Pension	0.844
3		1	Savings	0.957	Pension	0.802
4		3	Pension	0.132	Mortgage	0.107
5		1	Pension	0.805	Savings	0.284
6		3	Pension	0.132	Mortgage	0.107
7		2	Pension	0.132	Mortgage	0.107
8		3	Pension	0.132	Mortgage	0.107
9		1	Pension	0.132	Mortgage	0.107
10		1	Pension	0.132	Mortgage	0.107
11		2	Pension	0.132	Mortgage	0.107
12		2	Pension	0.132	Mortgage	0.107
13		2	Savings	0.957	Mortgage	0.829
14		2	Savings	0.164	Pension	0.132
15		2	Savings	0.957	Pension	0.868
16		2	Pension	0.132	Mortgage	0.107
17		3	Pension	0.132	Mortgage	0.107
18		3	Pension	0.132	Mortgage	0.107
19		3	Savings	0.289	Pension	0.132
20		2	Pension	0.132	Mortgage	0.107

Des explications sur les fondements mathématiques des méthodes de modélisation utilisées dans SPSS Modeler sont présentées dans le *Guide des algorithmes de SPSS Modeler*, disponible dans le répertoire *\Documentation* du produit DVD.

Sachez également que ces résultats sont établis uniquement d'après les données d'apprentissage. Pour évaluer comment le modèle peut se généraliser à d'autres données dans le monde réel, vous devez utiliser un noeud Partitionner destiné à contenir un sous-ensemble d'enregistrements à des fins de test et de validation. Pour plus d'informations, reportez-vous à la section Noeud Partitionner dans le chapitre 4 dans *Noeuds source, exécution et de sortie de IBM SPSS Modeler 15*. Pour plus d'informations sur le noeud MRAA, reportez-vous au chapitre le chapitre 14 du guide Référence des noeuds.

Chapitre
17

Prévision des défauts de paiement (Réseau Bayésien)

Les réseaux Bayésiens permettent de créer un modèle de probabilité en combinant les preuves observées et enregistrées avec les connaissances réelles “de bon sens” pour établir la probabilité des occurences en utilisant des attributs apparemment sans lien.

Cet exemple utilise le flux nommé *bayes_bankloan.str*, qui fait référence au fichier de données *bankloan.sav*. Ces fichiers sont accessibles dans le répertoire *Demos* de toute installation IBM® SPSS® Modeler. Vous pouvez y accéder à partir du groupe de programmes IBM® SPSS® Modeler du menu Démarrer de Windows. Le fichier *bayes_bankloan.str* se trouve dans le répertoire des *flux*.

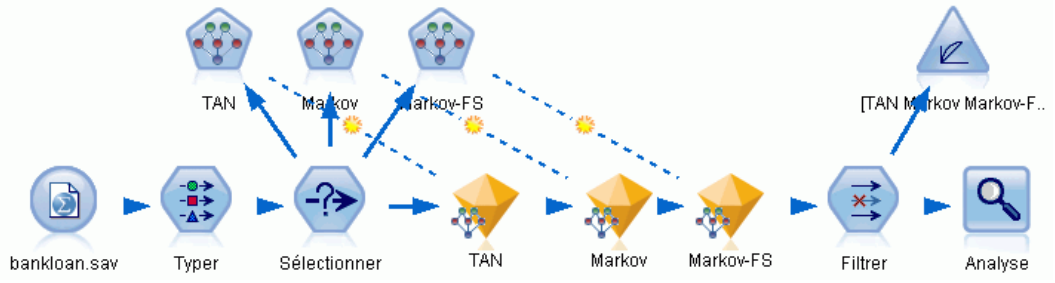
Par exemple, imaginons qu’une banque s’inquiète des prêts susceptibles de ne pas être remboursés. Si les données par défaut des prêts précédents peuvent être utilisées pour prédire quels clients potentiels risquent d’avoir des difficultés à rembourser leur prêt, il est alors possible de refuser un prêt à ces clients à “fort risque” ou de leur proposer d’autres produits.

Cet exemple se concentre sur l’utilisation de données par défaut sur des prêts existants permettant de prédire quels futurs clients sont susceptibles de ne pas pouvoir rembourser leur prêt et examine trois différents types de modèle de réseau Bayésien afin d’établir celui le plus adapté à cette situation.

Création du flux

- Ajoutez un noeud source Statistics pointant sur *bankloan.sav* dans le dossier *Demos*.

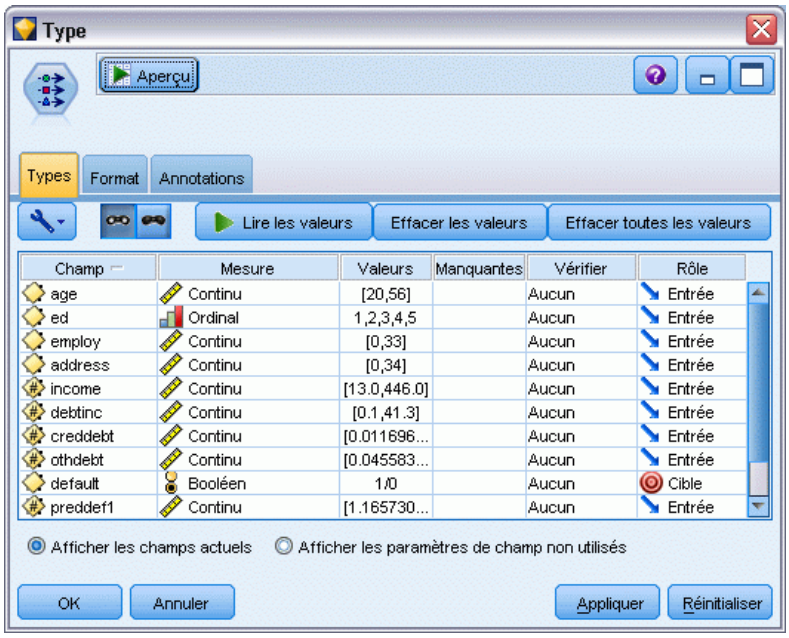
Figure 17-1
Exemple de flux de réseau Bayésien



- Ajoutez un noeud Typer au noeud source et définissez le rôle du champ par défaut sur Cible. Le rôle de tous les autres champs doit être défini sur Entrée.

▶ Cliquez sur le bouton Lire les valeurs pour remplir la colonne *Valeurs*.

Figure 17-2
Sélection du champ cible



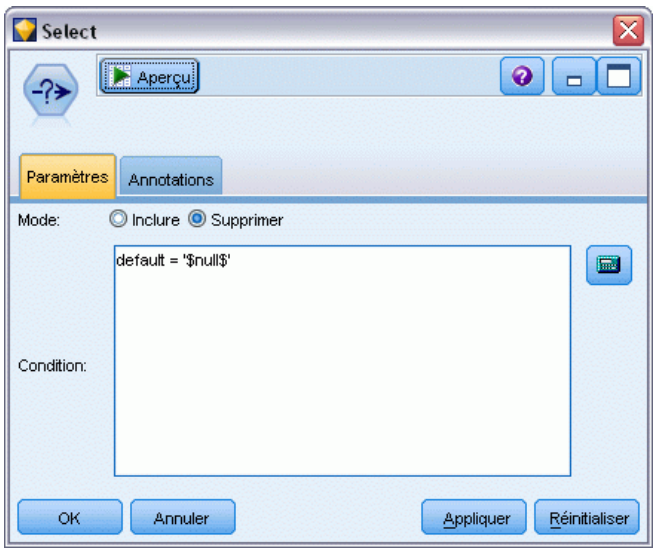
Les observations ayant une valeur de cible nulle ne servent à rien lors de la création d'un modèle. Il est possible d'exclure ces observations afin qu'elles ne soient pas utilisées lors de l'évaluation du modèle.

▶ Ajoutez un noeud Sélectionner au noeud Typer.

▶ Pour le mode, choisissez Supprimer.

▶ Dans la boîte de dialogue Condition, saisissez valeur par défaut = '$null$'.

Figure 17-3
Suppression des cibles à valeur nulle



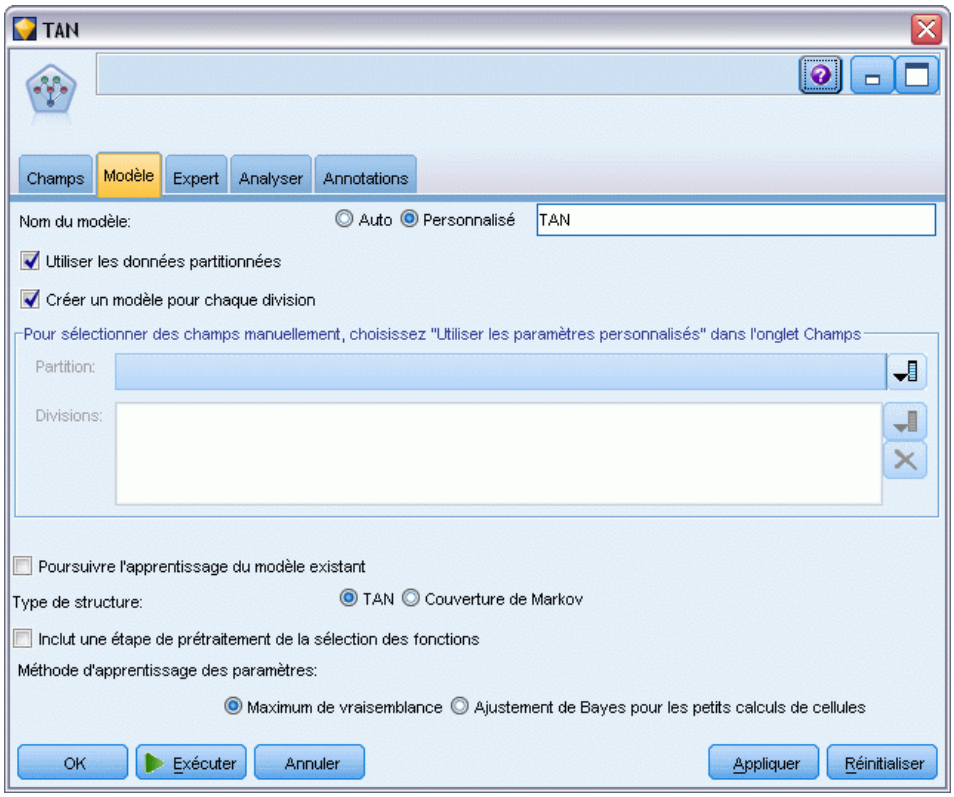
Il est possible de créer plusieurs types de réseaux Bayésiens. Par conséquent, il peut être utile d'en comparer plusieurs afin de connaître celui qui offre les meilleures prédictions. Le premier est un modèle Tree Augmented Naïve Bayes (TAN).

▶ Relier un noeud de réseau Bayésien au noeud Sélectionner.

▶ Pour choisir le nom du modèle, dans l'onglet Modèle, sélectionnez Personnalisé puis saisissez TAN dans la zone de texte.

▶ Pour le type de structure, sélectionnez TAN et cliquez sur OK.

Figure 17-4
Création d'un modèle Tree Augmented Naïve Bayes



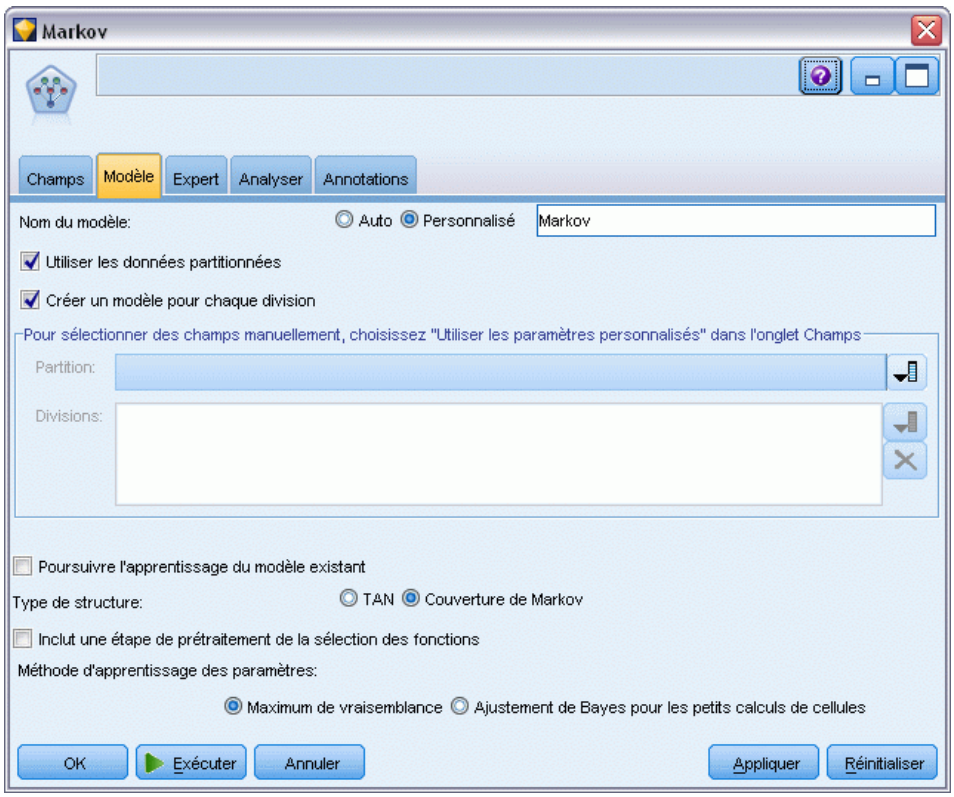
Le deuxième type de modèle à créer a une structure de couverture de Markov.

▶ Relier un deuxième noeud de réseau Bayésien au noeud Sélectionner.

▶ Pour choisir le nom du modèle, dans l'onglet Modèle, sélectionnez Personnalisé puis saisissez Markov dans la zone de texte.

▶ Pour le type de structure, sélectionnez Couverture de Markov et cliquez sur OK.

Figure 17-5
Création d'un modèle Couverture de Markov



Le troisième type de modèle a une structure de couverture de Markov et utilise également le prétraitement de la sélection de fonctions pour sélectionner les entrées importantes liées à la variable cible.

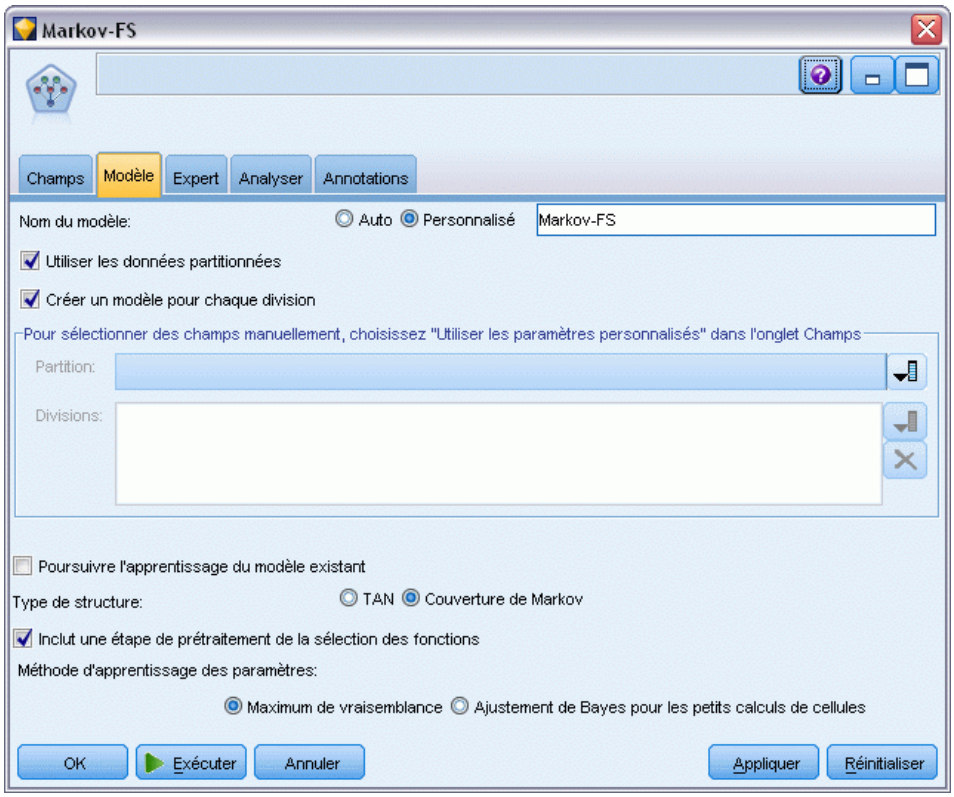
▶ Relier un troisième noeud de réseau Bayésien au noeud Sélectionner.

▶ Pour choisir le nom du modèle, dans l'onglet Modèle, sélectionnez Personnalisé puis saisissez Markov-FS dans la zone de texte.

▶ Pour le type de structure, sélectionnez Couverture de Markov.

▶ Sélectionner Inclure une étape de prétraitement de la sélection des fonctions et cliquer sur OK.

Figure 17-6
Création d'un modèle de couverture de Markov avec prétraitement de la sélection des fonctions.



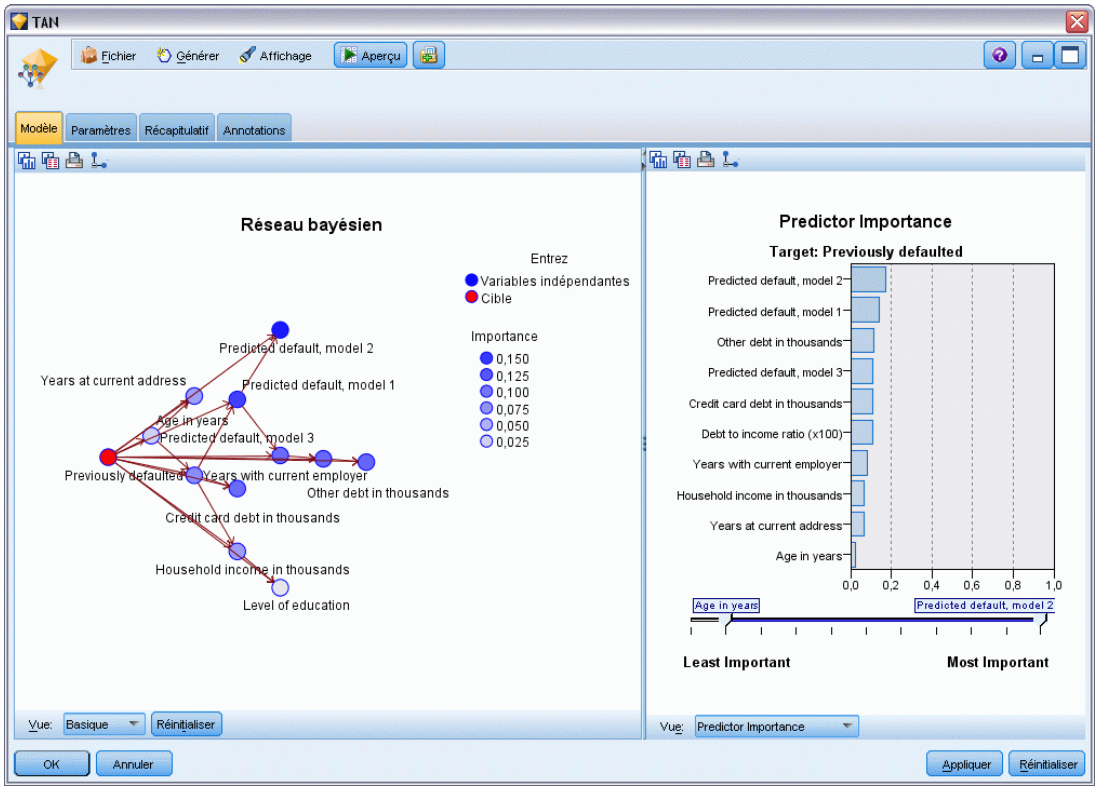
Navigation dans le modèle

▶ Exécutez le flux pour créer des nuggets de modèle, qui sont ajoutés au flux et à la palette Modèles dans l'angle supérieur droit. Pour afficher leurs détails, double-cliquez sur l'un des nuggets de modèle du flux.

L'onglet Modèle du nugget de modèle est divisé en deux panneaux. Le panneau de gauche contient un graphique de noeuds en réseau qui affiche la relation entre la cible et ses variables indépendantes les plus importantes, ainsi que la relation entre les variables indépendantes.

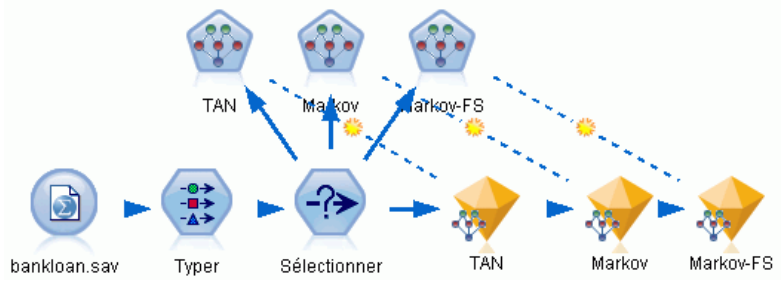
Le panneau de droite affiche soit l'*Importance des variables indépendantes*, qui indique l'importance relative de chaque variable indépendante pour l'estimation du modèle ou les *Probabilités conditionnelles*, qui contiennent la valeur de la probabilité conditionnelle de chaque valeur de nœud et pour chaque combinaison de valeurs dans ses nœuds parent.

Figure 17-7
Afichage d'un modèle Tree Augmented Naïve Bayes



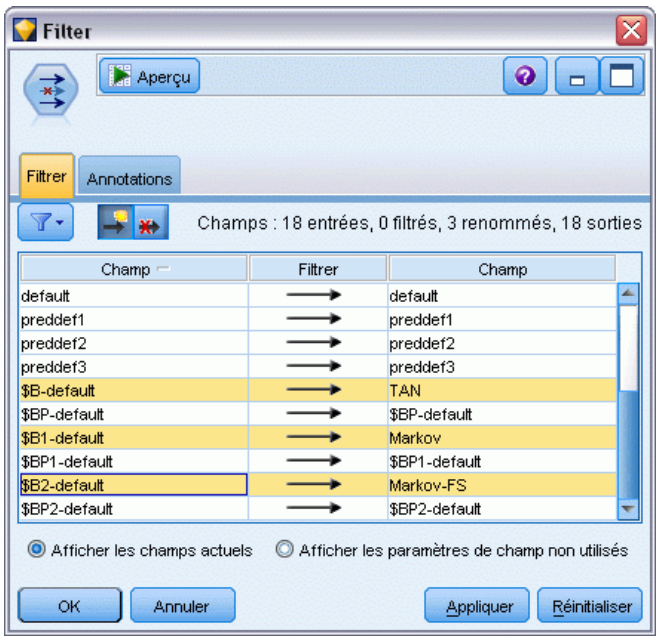
- Connectez le nugget de modèle TAN au nugget Markov (choisissez Remplacer dans la boîte de dialogue d'avertissement).
- Connectez le nugget Markov au nugget Markov-FS (choisissez Remplacer dans la boîte de dialogue d'avertissement).
- Alignez les trois nuggets avec le noeud Sélectionner pour une meilleure lecture.

Figure 17-8
Alignement des nuggets dans le flux



- Pour renommer les sorties de modèle et obtenir un graphique d'évaluation plus clair, liez le noeud Filtrer au nugget de modèle Markov-FS.
- Dans la colonne de droite *Champ* , remplacez le nom $B par défaut par TAN, le nom $B1 par défaut par Markov et le nom $B2 par défaut par Markov-FS.

Figure 17-9
Modifiez les noms des champs de modèle

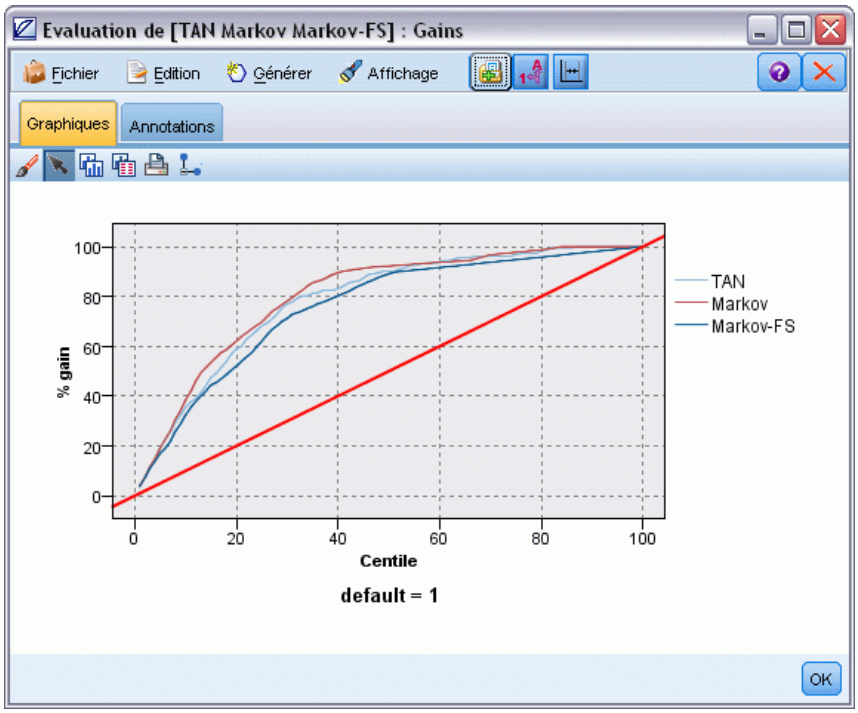


Pour comparer la précision des prédictions des modèles, vous pouvez créer un graphique de gains.

- Liez un noeud de graphique d'évaluation au noeud Filtre et exécutez le noeud Graphique en utilisant les paramètres par défaut.

Le graphique montre que chaque type de modèle produit des résultats similaires ; mais, le modèle Markov est légèrement meilleur.

Figure 17-10
Evaluation de la précision du modèle



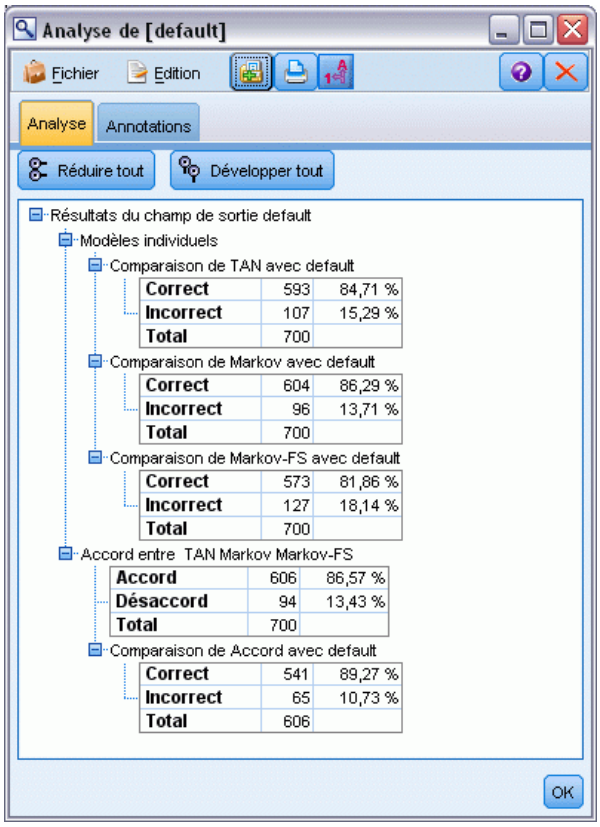
Pour vérifier la précision des prédictions de chaque modèle, vous pouvez utiliser le noeud Analyse à la place du graphique d'évaluation. Ceci affiche la précision en pourcentage à la fois pour les prédictions correctes et incorrectes.

▶ Liez un noeud Analyse au noeud Filtre et exécutez le noeud Analyse en utilisant les paramètres par défaut.

Comme pour le graphique d'évaluation, ce noeud montre que le modèle Markov est légèrement meilleur dans ses prédictions ; cependant, le modèle Markov-FS n'est qu'à quelques points de pourcentage derrière ce modèle. Cela peut indiquer qu'il est peut-être préférable d'utiliser le modèle Markov-FS car celui-ci utilise un moins grand nombre d'entrées pour calculer ses

résultats, ce qui diminue la durée de la collecte de données ainsi que la durée d'entrée et de traitement des données.

Figure 17-11
Analyse de la précision des modèles



Des explications sur les fondements mathématiques des méthodes de modélisation utilisées dans IBM® SPSS® Modeler sont présentées dans le *Guide des algorithmes de SPSS Modeler*, disponible dans le répertoire *\Documentation* du disque d'installation.

Sachez également que ces résultats sont établis uniquement d'après les données d'apprentissage. Pour évaluer comment le modèle peut se généraliser à d'autres données dans le monde réel, vous devez utiliser un noeud Partitionner destiné à contenir un sous-ensemble d'enregistrements à des fins de test et de validation. Pour plus d'informations, reportez-vous à la section Noeud Partitionner dans le chapitre 4 dans *Noeuds source, exécution et de sortie de IBM SPSS Modeler 15*.

Chapitre
18

Recyclage d'un modèle chaque mois (Réseau Bayésien)

Les réseaux Bayésiens permettent de créer un modèle de probabilité en combinant les preuves observées et enregistrées avec les connaissances réelles "de bon sens" pour établir la probabilité des occurences en utilisant des attributs apparemment sans lien.

Cet exemple utilise le flux *bayes_churn_retrain.str*, qui fait référence aux fichiers de données *telco_Jan.sav* et *telco_Feb.sav*. Ces fichiers sont accessibles dans le répertoire *Demos* de toute installation IBM® SPSS® Modeler. Vous pouvez y accéder à partir du groupe de programmes IBM® SPSS® Modeler du menu Démarrer de Windows. Le fichier *bayes_churn_retrain.str* se trouve dans le répertoire des *flux*.

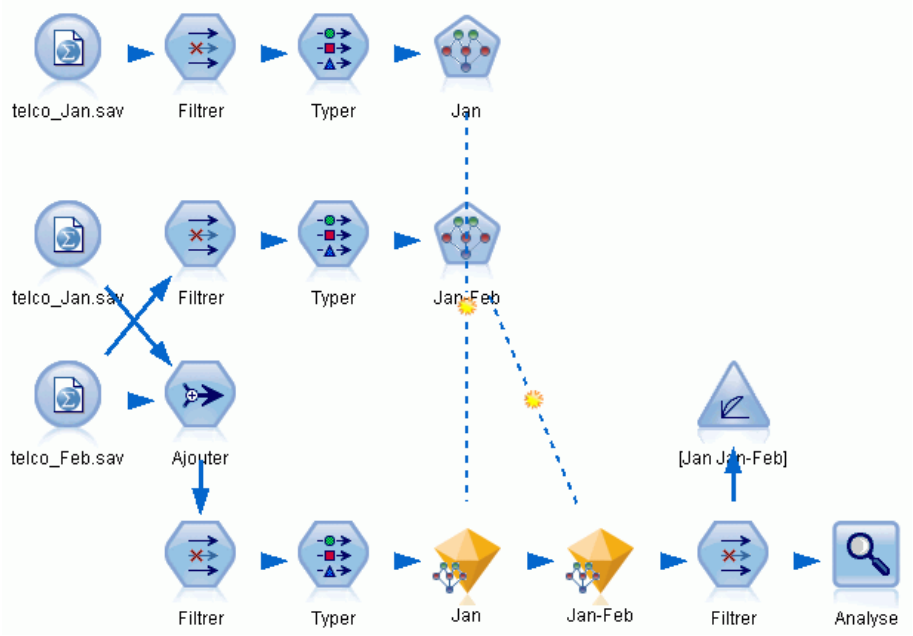
Par exemple, supposons qu'un fournisseur de télécommunications souhaite connaître le nombre de clients qui partent à la concurrence (attrition). Si les données client historiques peuvent être utilisées pour prédire quels clients sont les plus susceptibles d'attrition, ces clients peuvent être ciblés afin de recevoir des remises ou des offres pour les dissuader de passer à un autre fournisseur de services.

Cet exemple examine l'utilisation de données d'attrition mensuelles existantes pour prédire quels clients sont susceptibles d'attrition et l'ajout des données du mois suivant pour affiner et recycler ce modèle.

Création du flux

- Ajoutez un noeud source Statistics pointant vers *telco_Jan.sav* dans le dossier *Demos*.

Figure 18-1
Exemple de flux de réseau Bayésien

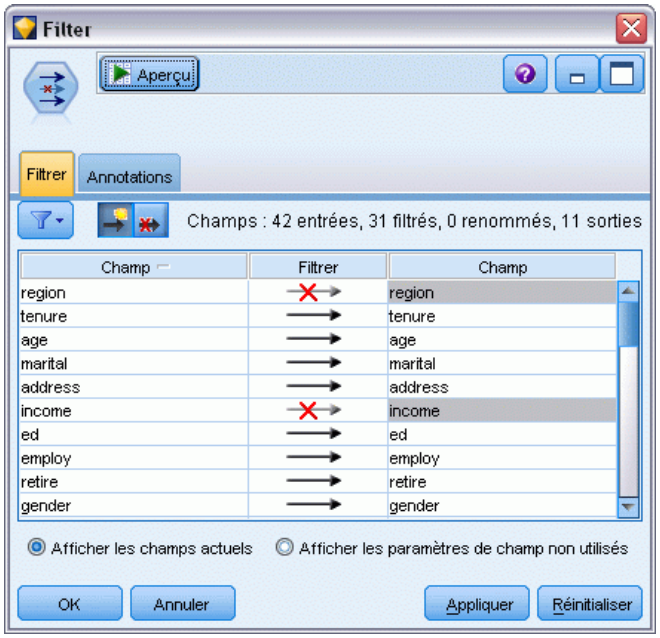


L'analyse précédente a démontré que plusieurs champs de données sont inutiles pour la prédiction de l'attrition. Ces champs peuvent être filtrés à partir de votre ensemble de données afin de réduire la durée du traitement lors de la création et du scoring de modèles.

- Ajoutez un noeud Filtrer au noeud Source.
- Excluez tous les champs à l'exception des champs *address*, *age*, *churn*, *custcat*, *ed*, *employ*, *gender*, *marital*, *reside*, *retire*, et *tenure*.

▶ Cliquez sur OK.

Figure 18-2
Filtrage des champs inutiles



▶ Ajoutez un noeud Typer au noeud Filtrer.

▶ Ouvrez le noeud Typer et cliquez sur le bouton Lire les valeurs pour remplir la colonne *Valeurs*.

- Pour que le noeud Evaluation puisse évaluer les valeurs True (vrai) et False (faux), définissez le niveau de mesure pour le champ *attrition* sur Booléen, et son rôle sur Cible. Cliquez sur OK.

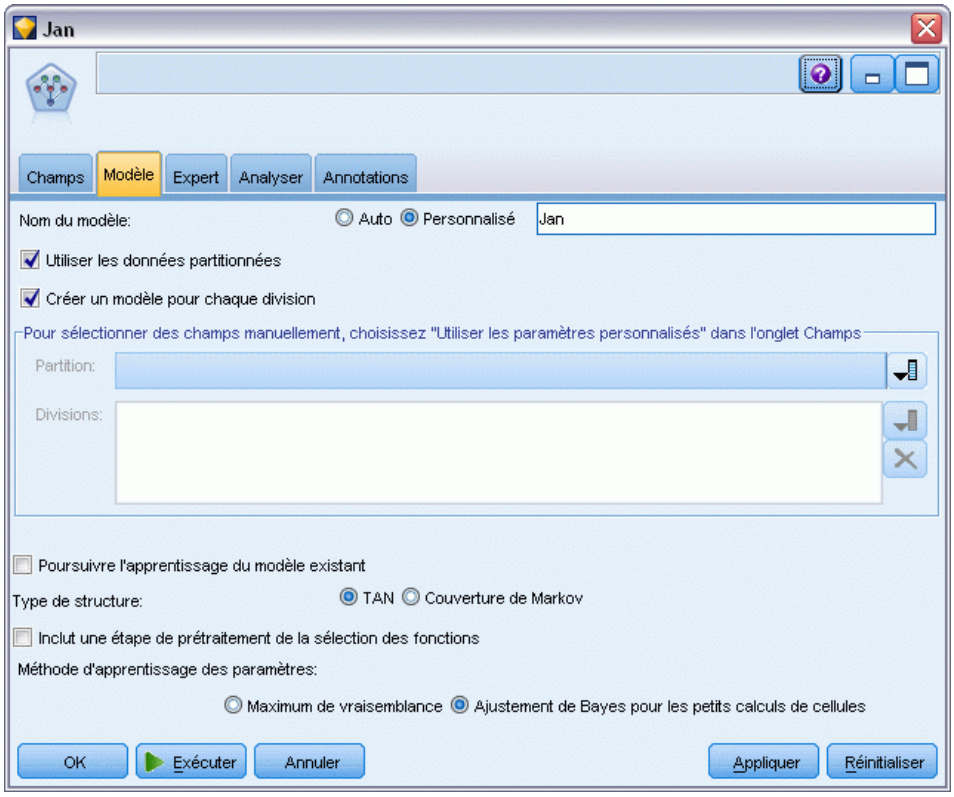
Figure 18-3
Sélection du champ cible

Vous pouvez créer plusieurs types de réseaux Bayésiens ; mais, dans cet exemple, vous créez un modèle Tree Augmented Naïve Bayes (TAN). Celui-ci crée un réseau étendu qui vous permet d'inclure tous les liens possibles entre les variables de données, créant ainsi un modèle initial fiable.

- Relier un noeud de réseau Bayésien au noeud Typer.
- Pour choisir le nom du modèle, dans l'onglet Modèle, sélectionnez Personnalisé puis saisissez Jan dans la zone de texte.
- Pour la méthode d'apprentissage des paramètres, sélectionnez Ajustement de Bayes pour les petits calculs de cellules.

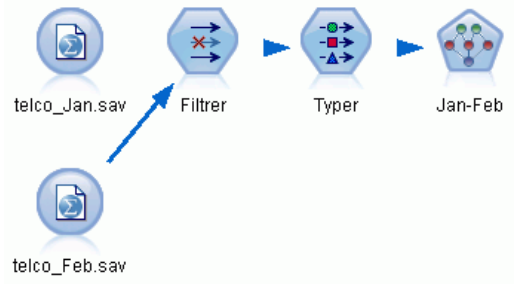
- Cliquez sur Exécuter. Le nugget de modèle est ajouté au flux et également à la palette Modèles en haut à droite.

Figure 18-4
Création d'un modèle Tree Augmented Naïve Bayes



- Ajoutez un noeud source Fichier de statistiques pointant vers *telco_Feb.sav* dans le dossier *Demos*.
- Attachez ce nouveau noeud source au noeud Filtrer (dans la boîte de dialogue d'avertissement, choisissez Remplacer pour remplacer la connexion au noeud source précédent).

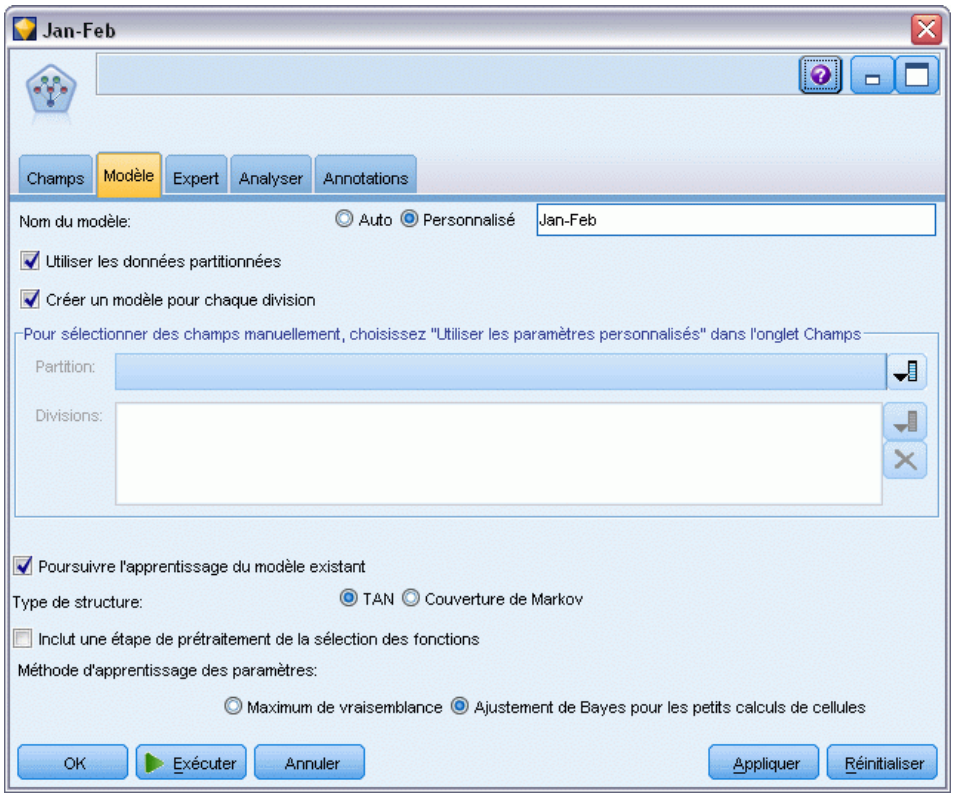
Figure 18-5
Ajout des données du deuxième mois



- Pour choisir le nom du modèle, dans l'onglet Modèle du noeud Réseau Bayésien, sélectionnez Personnalisé puis saisissez Jan-Feb dans la zone de texte.
- Sélectionnez Poursuivre l'apprentissage du modèle existant.

▶ Cliquez sur Exécuter. Le nugget de modèle remplace le nugget existant dans le flux mais il est également ajouté à la palette Modèles en haut à droite.

Figure 18-6
Recyclage du modèle

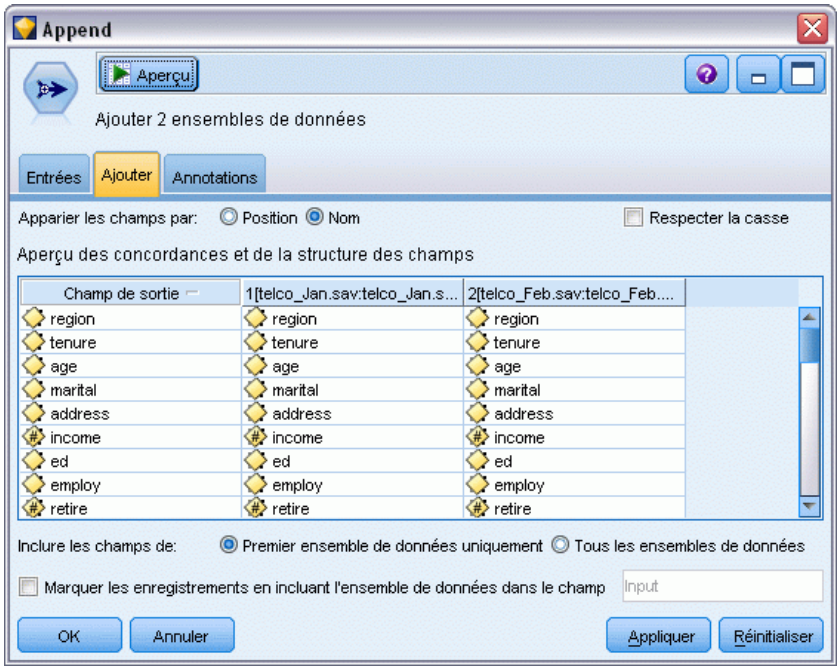


Evaluation du modèle

Pour comparer les modèles, vous devez combiner les deux ensembles de données.

▶ Ajoutez un noeud Ajouter et liez-y les noeuds source *telco_Jan.sav* et *telco_Feb.sav*.

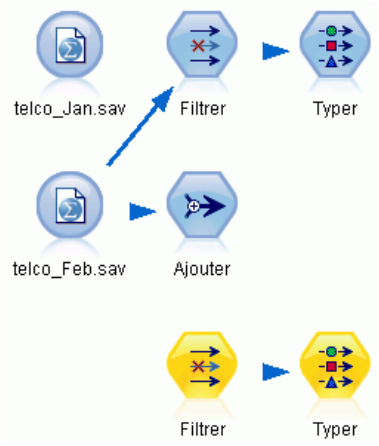
Figure 18-7
Ajoutez les deux sources de données



▶ Copiez les noeuds Filtrer et Typer précédents dans le flux et collez-les dans l'espace de travail de flux.

▶ Liez le noeud Ajouter au noeud Filtrer que vous venez de coller.

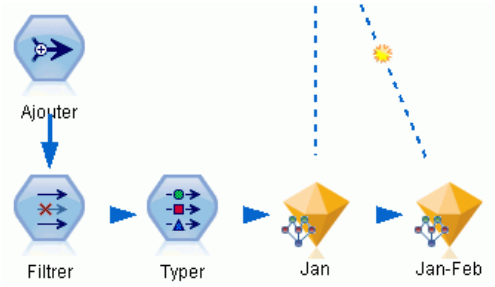
Figure 18-8
Collage des noeuds copiés dans le flux



Les nuggets des deux modèles de réseau Bayésien se trouvent dans la palette Modèles en haut à droite.

- Double-cliquez sur le nugget de modèle Jan pour l'ajouter au flux et reliez-le au nouveau noeud copié Typer.
- Liez le nugget de modèle Jan-Feb déjà dans le flux au nugget de modèle Jan.
- Ouvrez le nugget de modèle Jan.

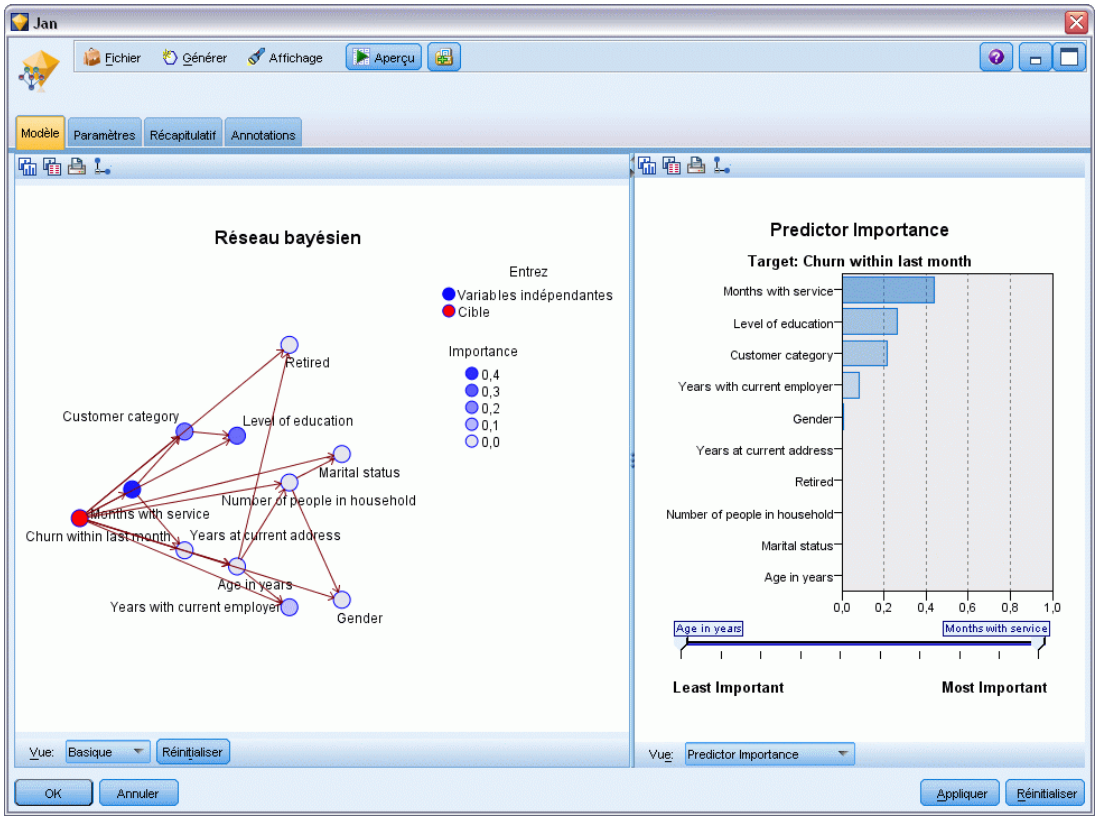
Figure 18-9
Ajout des nuggets au flux



L'onglet Modèle du nugget de modèle Réseau Bayésien est divisé en deux colonnes. La colonne de gauche contient un graphique de noeuds en réseau qui affiche la relation entre la cible et ses variables indépendantes les plus importantes, ainsi que la relation entre les variables indépendantes.

La colonne de droite affiche soit l'*Importance des variables indépendantes*, qui indique l'importance relative de chaque variable indépendante pour l'estimation du modèle ou les *Probabilités conditionnelles*, qui contiennent la valeur de la probabilité conditionnelle de chaque valeur de nœud et pour chaque combinaison de valeurs dans ses nœuds parent.

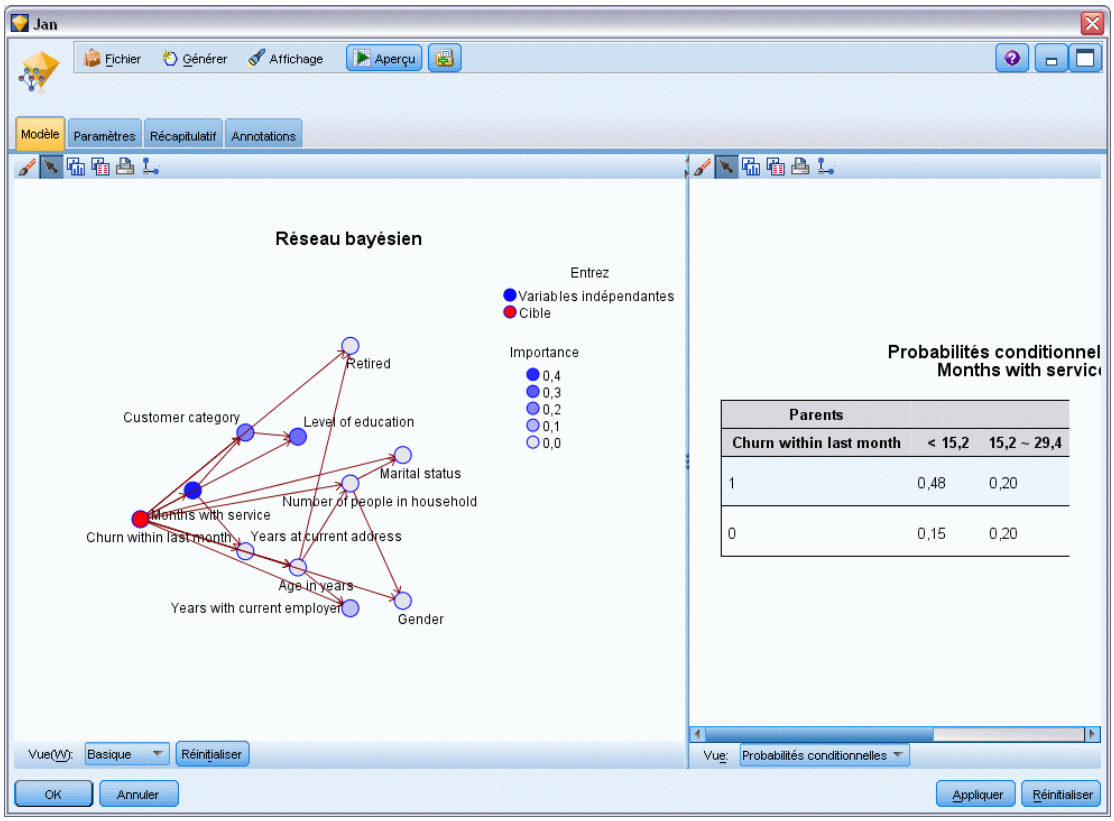
Figure 18-10
Modèle Réseau Bayésien affichant l'importance des variables indépendantes



Pour afficher les probabilités conditionnelles d'un noeud, cliquez sur ce noeud dans la colonne de gauche. La colonne de droite est mise à jour et affiche les détails appropriés.

Les probabilités conditionnelles sont affichées pour chaque intervalle de division des valeurs de données associé aux noeuds parent et aux noeuds frères.

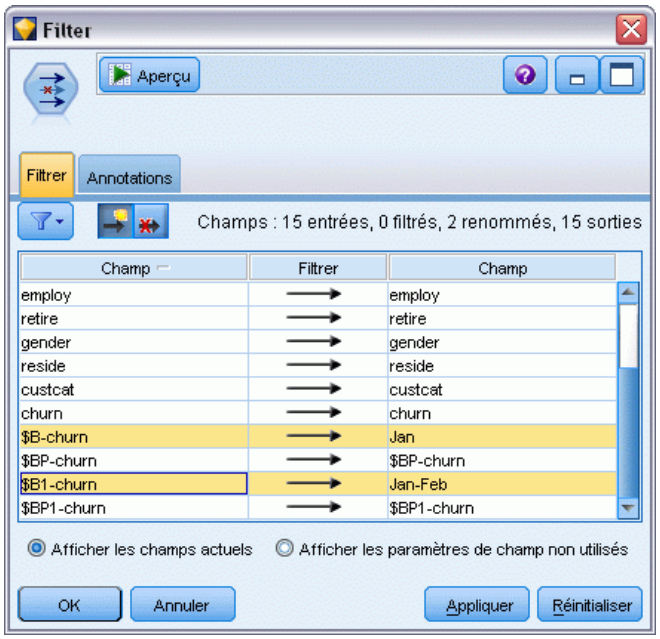
Figure 18-11
Modèle Réseau Bayésien affichant les probabilités conditionnelles



- Pour renommer les sorties de modèles et les rendre plus claires, liez un noeud Filtrer au nugget de modèle Jan-Feb.

▶ Dans la colonne de droite *Champ* , remplacez le nom $B-attrition par Jan et $B1-attrition par Jan-Fév.

Figure 18-12
Modifiez les noms des champs de modèle



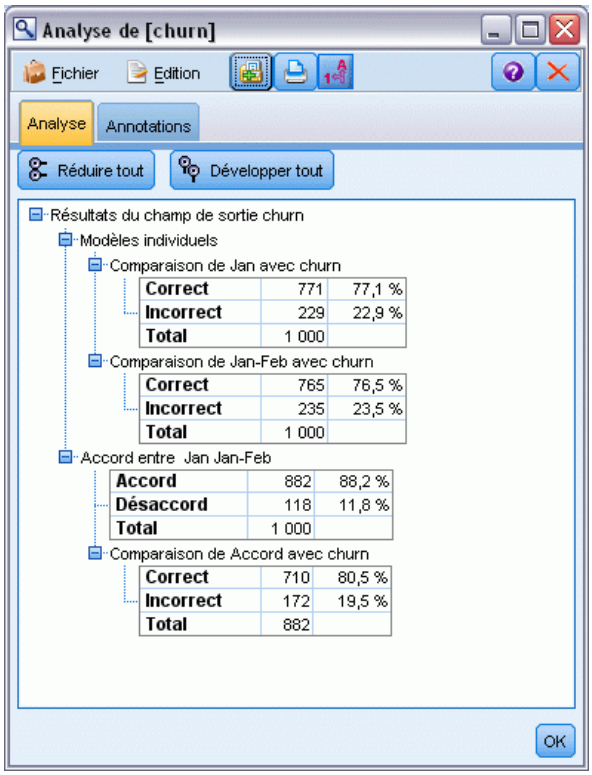
Pour vérifier la précision des prédictions d'attrition de chaque modèle, utilisez un noeud Analyse qui permet d'afficher cette précision en termes de pourcentage, à la fois pour les prédictions correctes et incorrectes.

▶ Reliez un noeud Analyse au noeud Filtrer.

▶ Ouvrez le noeud Analyse, puis cliquez sur Exécuter.

Cela montre que les deux modèles ont des degrés de précision semblables lors de la prédiction d'attrition.

Figure 18-13
Analyse de la précision des modèles



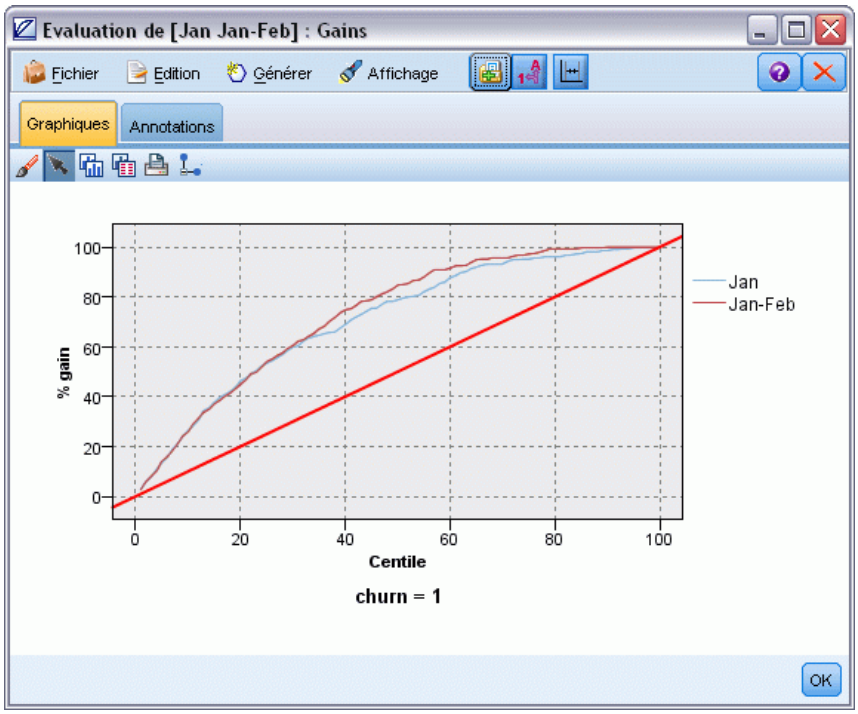
A la place du noeud Analyse, vous pouvez utiliser un graphique Evaluation pour comparer la précision des prédictions des modèles en créant un graphique de gains.

- Ajoutez un noeud de graphique Evaluation au noeud Filtrer.

et exécutez le noeud Graphiques en utilisant ses paramètres par défaut.

Comme le noeud Analyse, ce graphique affiche que chaque type de modèle produit des résultats similaires ; mais, le modèle recyclé qui utilise les données des deux mois est légèrement meilleur car ses prédictions ont un plus haut niveau de confiance.

Figure 18-14
Evaluation de la précision du modèle



Des explications sur les fondements mathématiques des méthodes de modélisation utilisées dans IBM® SPSS® Modeler sont présentées dans le *Guide des algorithmes de SPSS Modeler*, disponible dans le répertoire *\Documentation* du disque d'installation.

Sachez également que ces résultats sont établis uniquement d'après les données d'apprentissage. Pour évaluer comment le modèle peut se généraliser à d'autres données dans le monde réel, vous devez utiliser un noeud Partitionner destiné à contenir un sous-ensemble d'enregistrements à des fins de test et de validation. Pour plus d'informations, reportez-vous à la section Noeud Partitionner dans le chapitre 4 dans *Noeuds source, exécution et de sortie de IBM SPSS Modeler 15*.

Chapitre
19

Campagne publicitaire (R. neurones/Arbre C&RT)

Cet exemple s'appuie sur des données relatives à des gammes de produits destinés à la vente au détail et aux effets de la campagne publicitaire sur les ventes. (Ces données sont fictives.) Votre objectif, dans cet exemple, est de prévoir les effets des prochaines campagnes publicitaires. Comme dans l'exemple de surveillance d'état, le processus de Data mining se compose des étapes suivantes : exploration, préparation des données, apprentissage et tests.

Cet exemple utilise les flux nommés *goodsplot.str* et *goodslearn.str*, qui font référence aux fichiers de données nommés *GOODS1n* et *GOODS2n*. Ces fichiers sont disponibles dans le répertoire *Demos* de n'importe quelle installation de IBM® SPSS® Modeler. Ce dossier est accessible à partir du groupe de programmes IBM® SPSS® Modeler dans le menu Démarrer de Windows. Le flux *goodsplot.str* se situe dans le dossier des *flux* , tandis que le fichier *goodslearn.str* se situe dans le répertoire des *flux* .

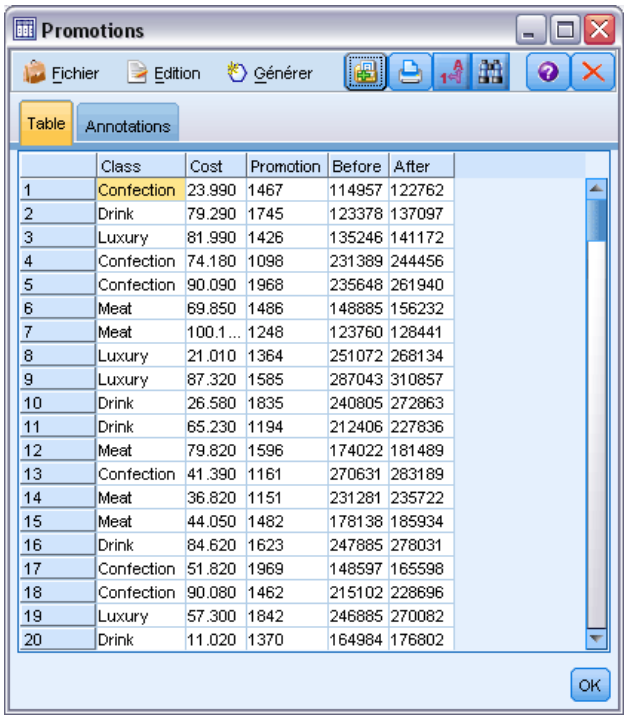
Examen des données

Chaque enregistrement comprend les éléments suivants :

- *Catégorie*. Type de produit.
- *Coût*. Prix unitaire.
- *Campagne publicitaire*. Somme consacrée à une campagne publicitaire particulière.
- *Avant*. Recettes avant la campagne publicitaire.
- *Après*. Recettes après la campagne publicitaire.

Le flux *goodsplot.str* contient un flux simple permettant d'afficher les données dans un tableau. Les deux champs relatifs aux recettes (*Avant* et *Après*) sont exprimés en termes absolus ; cependant, la valeur de l'augmentation des recettes après la campagne publicitaire (certainement due à celle-ci) vous sera probablement plus utile.

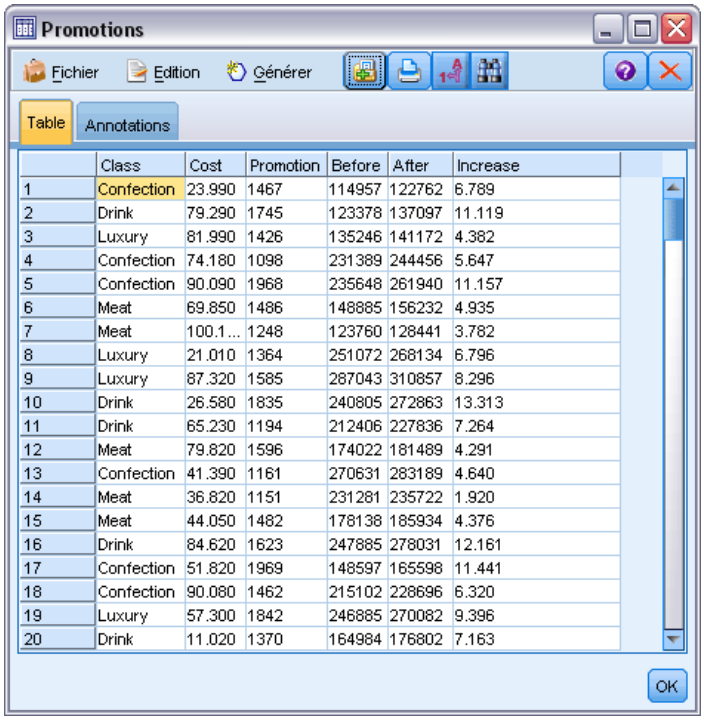
Figure 19-1
Effets de la campagne publicitaire sur les ventes du produit



	Class	Cost	Promotion	Before	After
1	Confection	23.990	1467	114957	122762
2	Drink	79.290	1745	123378	137097
3	Luxury	81.990	1426	135246	141172
4	Confection	74.180	1098	231389	244456
5	Confection	90.090	1968	235648	261940
6	Meat	69.850	1486	148885	156232
7	Meat	100.1...	1248	123760	128441
8	Luxury	21.010	1364	251072	268134
9	Luxury	87.320	1585	287043	310857
10	Drink	26.580	1835	240805	272863
11	Drink	65.230	1194	212406	227836
12	Meat	79.820	1596	174022	181489
13	Confection	41.390	1161	270631	283189
14	Meat	36.820	1151	231281	235722
15	Meat	44.050	1482	178138	185934
16	Drink	84.620	1623	247885	278031
17	Confection	51.820	1969	148597	165598
18	Confection	90.080	1462	215102	228696
19	Luxury	57.300	1842	246885	270082
20	Drink	11.020	1370	164984	176802

goodsplot.str contient également un noeud pour calculer cette valeur exprimée sous la forme d'un pourcentage de la recette avant la campagne publicitaire, dans un champ appelé *Augmentation*, et affiche un tableau indiquant ce champ.

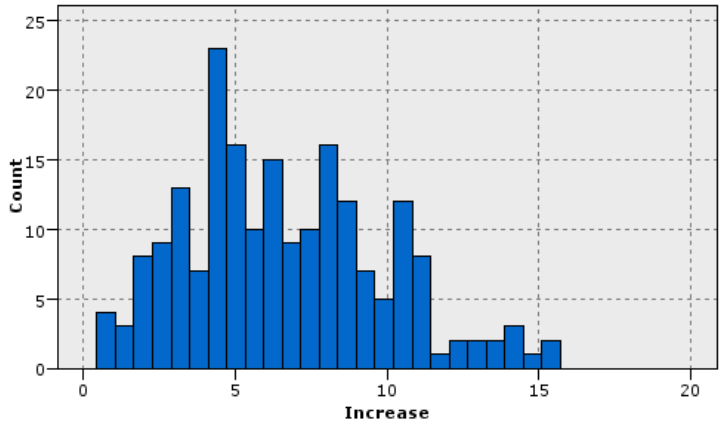
Figure 19-2
Augmentation des recettes après la campagne publicitaire



	Class	Cost	Promotion	Before	After	Increase
1	Confection	23.990	1467	114957	122762	6.789
2	Drink	79.290	1745	123378	137097	11.119
3	Luxury	81.990	1426	135246	141172	4.382
4	Confection	74.180	1098	231389	244456	5.647
5	Confection	90.090	1968	235648	261940	11.157
6	Meat	69.850	1486	148885	156232	4.935
7	Meat	100.1...	1248	123760	128441	3.782
8	Luxury	21.010	1364	251072	268134	6.796
9	Luxury	87.320	1585	287043	310857	8.296
10	Drink	26.580	1835	240805	272863	13.313
11	Drink	65.230	1194	212406	227836	7.264
12	Meat	79.820	1596	174022	181489	4.291
13	Confection	41.390	1161	270631	283189	4.640
14	Meat	36.820	1151	231281	235722	1.920
15	Meat	44.050	1482	178138	185934	4.376
16	Drink	84.620	1623	247885	278031	12.161
17	Confection	51.820	1969	148597	165598	11.441
18	Confection	90.080	1462	215102	228696	6.320
19	Luxury	57.300	1842	246885	270082	9.396
20	Drink	11.020	1370	164984	176802	7.163

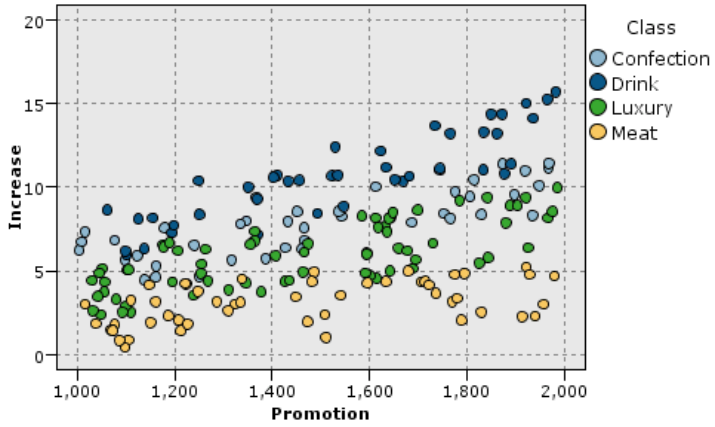
De plus, le flux affiche un histogramme de l'augmentation et un diagramme de dispersion de l'augmentation par rapport aux coûts de la campagne publicitaire, avec la catégorie de produits concernée.

Figure 19-3
Histogramme de l'augmentation des recettes



Le diagramme de dispersion fait apparaître que, pour chaque catégorie de produits, il existe une relation quasi-linéaire entre l'augmentation des recettes et les coûts engagés dans la campagne publicitaire. Il est donc probable qu'un arbre décision ou un réseau de neurones puisse prévoir, avec une précision relativement fiable, l'augmentation des recettes à l'aide des autres champs disponibles.

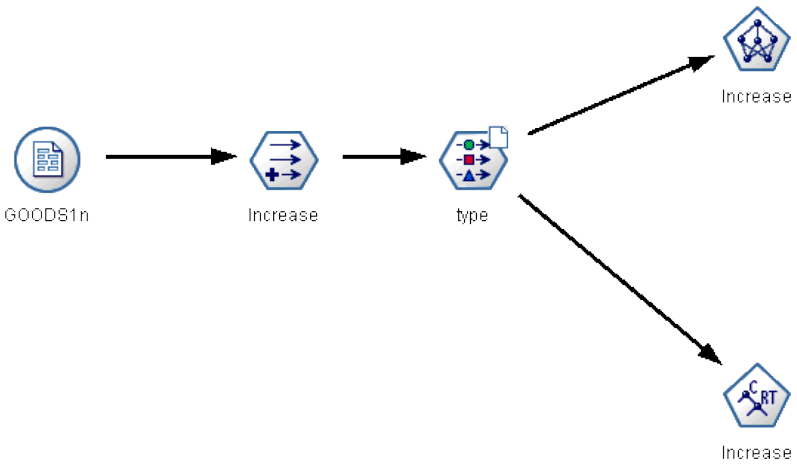
Figure 19-4
Rapport augmentation des recettes/dépenses publicitaires



Apprentissage et tests

Le flux *goodslearn.str* forme un réseau de neurones et un arbre décision pour effectuer cette prévision d'augmentation des recettes.

Figure 19-5
Modélisation du flux goodslearn.str (prodappren)



Une fois que vous avez exécuté les noeuds de modèle et généré les véritables modèles, vous pouvez tester les résultats du processus d'apprentissage. Pour ce faire, connectez l'arbre décision et le réseau en série entre le nœud Typer et un nouveau nœud Analyse, remplacez le fichier (de données) d'entrée par *PROD1n*, puis exécutez le nœud Analyse. A partir du résultat de ce nœud, notamment de la corrélation linéaire entre l'augmentation prévue et le résultat réel, vous remarquerez que les systèmes formés prévoient l'augmentation des recettes avec un niveau de fiabilité élevé.

Une étude plus poussée montrerait des cas où les systèmes formés commettent des erreurs assez importantes ; ces erreurs peuvent être identifiées par la représentation graphique de l'augmentation prévue par rapport à l'augmentation réelle. Dans ce cas, il vous suffirait de sélectionner les données déviantes à l'aide des diagrammes interactifs de IBM® SPSS® Modeler et, dans leurs propriétés, de rectifier la description des données et le processus d'apprentissage pour améliorer la précision des prévisions.

Chapitre
20
Surveillance d'état (R. neurones/C5.0)

Cet exemple se rapporte à la surveillance des informations d'état à partir d'un ordinateur, et à la difficulté à identifier et à prévoir les états de panne. Les données sont créées à partir d'une simulation fictive et se composent de plusieurs séries concaténées mesurées dans le temps. Chaque enregistrement rend compte du dernier état de la machine en indiquant les informations suivantes :

- *Heure*. Un entier.
- *Puissance*. Un entier.
- *Température*. Un entier.
- *Pression*. 0 si la situation est normale, 1 pour avertir d'un risque passager de pression.
- *Temps de bon fonctionnement*. Temps écoulé depuis la dernière intervention.
- *Etat*. Généralement 0. Il prend la valeur du code d'erreur (101, 202 ou 303) en cas d'erreur.
- *Résultat*. Le code d'erreur qui apparaît dans les séries temporelles ou 0 si aucune erreur ne se produit. (ces codes ne sont disponibles qu'a posteriori).

Cet exemple utilise les flux nommés *condplot.str* et *condlearn.str*, qui font référence aux fichiers de données *COND1n* et *COND2n*. Ces fichiers sont disponibles dans le répertoire *Demos* de n'importe quelle installation de IBM® SPSS® Modeler. Ce dossier est accessible à partir du groupe de programmes IBM® SPSS® Modeler dans le menu Démarrer de Windows. Les fichiers *condplot.str* et *condlearn.str* sont disponibles dans le répertoire des *flux* .

A chaque série temporelle correspond une série d'enregistrements commençant par une période d'activité normale suivie par la période menant à la panne, comme l'indique le tableau suivant :

Time	Puissance	Température	Pression	Temps de bon fonctionnement	Statut	Résultat
0	1059	259	0	404	0	0
1	1059	259	0	404	0	0
			...			
51	1059	259	0	404	0	0
52	1059	259	0	404	0	0
53	1007	259	0	404	0	303
54	998	259	0	404	0	303
			...			
89	839	259	0	404	0	303
90	834	259	0	404	303	303
0	965	251	0	209	0	0
1	965	251	0	209	0	0
			...			
51	965	251	0	209	0	0
52	965	251	0	209	0	0

Time	Puissance	Température	Pression	Temps de bon fonctionnement	Statut	Résultat
53	938	251	0	209	0	101
54	936	251	0	209	0	101
			...			
208	644	251	0	209	0	101
209	640	251	0	209	101	101

Le processus suivant est commun à la plupart des projets de Data mining :

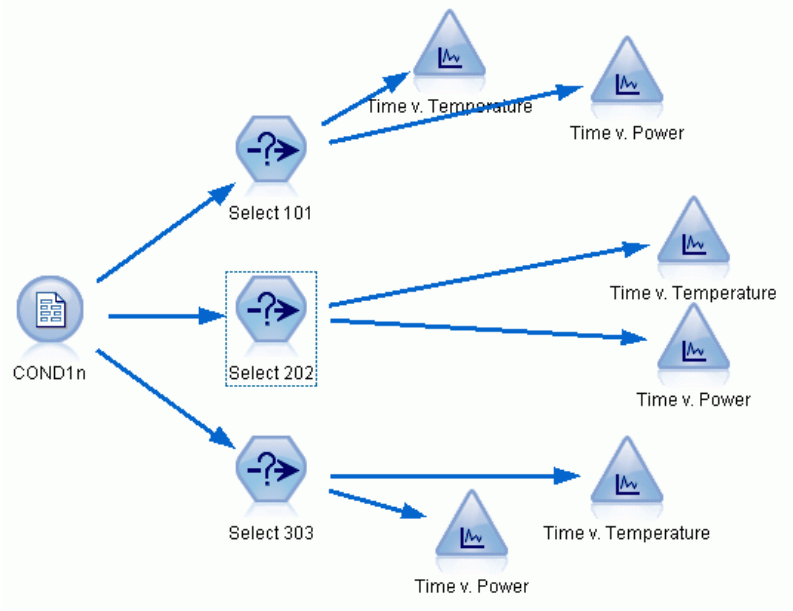
- Examinez les données permettant de déterminer les attributs utiles à la prévision ou à l'identification des états à connaître.
- Conservez ces attributs (s'ils existent) ou calculez-les, puis ajoutez-les aux données, si cela est nécessaire.
- Utilisez les données obtenues pour l'apprentissage des règles et des réseaux de neurones.
- Testez les systèmes formés à l'aide de données de test indépendantes.

Examen des données

Le fichier *condplot.str* illustre la première partie du processus. Il contient un flux qui reproduit plusieurs graphiques. Si les séries temporelles de température et de puissance contiennent des motifs visibles, vous pouvez distinguer les différentes conditions d'erreurs imminentes, voire prévoir le moment où l'erreur se produira. Pour la température et la puissance, le flux ci-dessous trace les séries temporelles associées aux trois codes d'erreur différents sur des graphiques

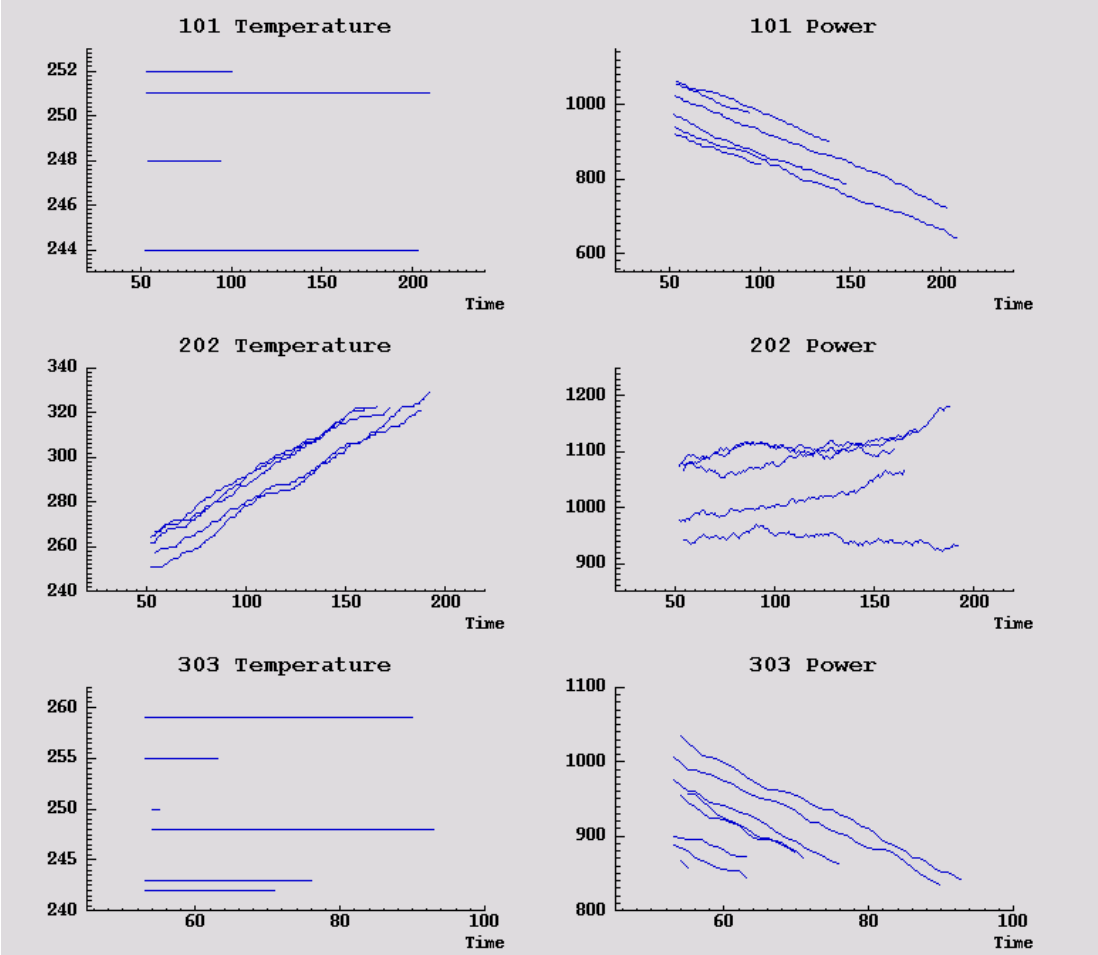
individuels, produisant six graphiques. Les noeuds Sélectionner séparent les données associées aux différents codes d'erreur.

Figure 20-1
Flux condplot (étatnuage)



Les résultats de ce flux sont indiqués dans cette figure.

Figure 20-2
Température et puissance dans le temps



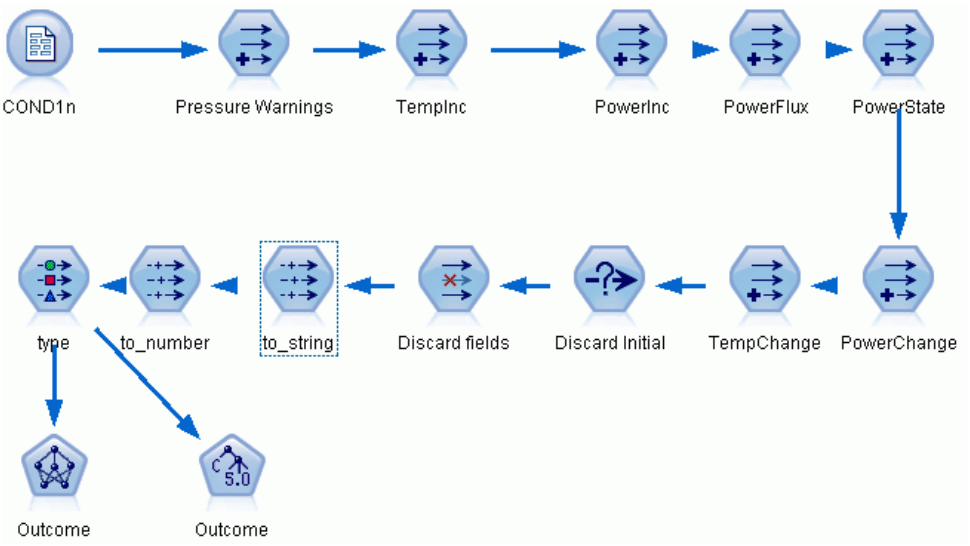
Les graphiques affichent clairement les motifs pouvant différencier jusqu'à 202 erreurs (de l'erreur 101 à l'erreur 303). Ces 202 erreurs rendent compte de l'augmentation de la température et de la fluctuation de la puissance dans le temps, ce qui n'est pas le cas des autres erreurs. Cependant, les motifs différenciant les erreurs 101 des erreurs 303 ne sont pas aussi explicites. Ces deux erreurs montrent une température régulière et une baisse de puissance, mais cette baisse est plus accentuée quand l'erreur 303 apparaît.

Ces graphiques permettent de constater que le changement et le degré de fluctuation de la température et de la puissance sont déterminants pour la prévision et l'identification des pannes. Ces attributs doivent donc être ajoutés aux données avant l'application des systèmes d'apprentissage.

Préparation des données

A partir des résultats du Data mining, le flux *condlearn.st* calcule les données appropriées et apprend à prévoir les pannes.

Figure 20-3
Flux condlearn (étatappren)



Le flux utilise plusieurs noeuds Calculer pour préparer les données en vue de la modélisation.

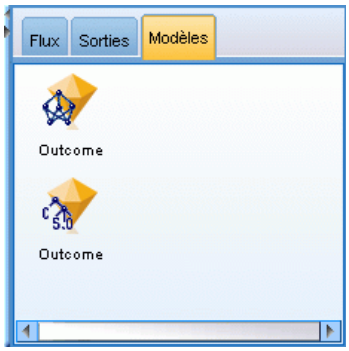
- **Noeud de type Délimité**. Lit le fichier de données *ETAT1n*.
- **Noeud Calculer Avertissements de pression**. Calcule le nombre d'avertissements de pression passagers. Réinitialisez cette valeur lorsque l'heure revient sur la valeur 0.
- **Noeud Calculer AugTemp**. Calcule le taux de changement passager de température à l'aide de @DIFF1.
- **Noeud Calculer AugPuiss**. Calcule le taux de changement passager de la puissance à l'aide de @DIFF1.
- **Noeud Calculer FluxPuiss**. Il s'agit d'un booléen avec la valeur True (vraie) si les variations de la puissance sont totalement différentes entre le dernier enregistrement et celui-ci, c'est-à-dire entre un pic de puissance et une baisse importante.
- **Noeud Calculer EtatPuiss**. L'état est *Stable* au début, puis devient *Fluctuant* lorsque deux flux de puissance successifs sont détectés. Il revient sur *Stable* uniquement lorsqu'il n'y a pas eu de flux de puissance pendant cinq intervalles de temps ou lorsque *Temps* est réinitialisé.
- **ModifPuiss**. Moyenne de *AugPuiss* sur les cinq derniers intervalles de temps.
- **ModifTemp**. Moyenne de *AugTemp* sur les cinq derniers intervalles de temps.
- **Noeud Abandonner (sélectionner)**. Supprime le premier enregistrement de chaque série temporelle pour éviter des écarts importants (inappropriés) de *Puissance* et de *Température* entre chaque enregistrement.

- **Noeud Abandonner les champs**. Réduit les champs des enregistrements à *Temps de bon fonctionnement*, *Etat*, *Résultat*, *Avertissements de pression*, *EtatPuiss*, *ModifPuiss* et *ModifTemp*.
- **Type**. Définit le rôle du *Résultat* en tant que Cible (le champ à prédire). En outre, définit le niveau de mesure de *Résultat* en tant que Nominal, *Avertissements de pression* en tant que Continu et *EtatPuiss* en tant que Booléen.

Apprentissage

L'exécution du flux dans *condlearn.str* permet l'apprentissage de la règle C5.0 et du réseau de neurones. Le temps d'apprentissage du réseau peut être long, mais vous pouvez interrompre le processus avant la fin pour enregistrer un réseau produisant des résultats satisfaisants. Une fois l'apprentissage terminé, l'onglet Modèles en haut à droite dans la fenêtre des gestionnaires clignote pour vous avertir que deux nuggets ont été créés : l'un représente le réseau de neurones et l'autre la règle.

Figure 20-4
Gestionnaire de modèles avec des nuggets de modèle



Les nuggets de modèle sont aussi ajoutés au flux existant, ce qui nous permet de tester le système ou d'exporter les résultats du modèle. Dans cet exemple, nous allons tester les résultats du modèle.

Testing

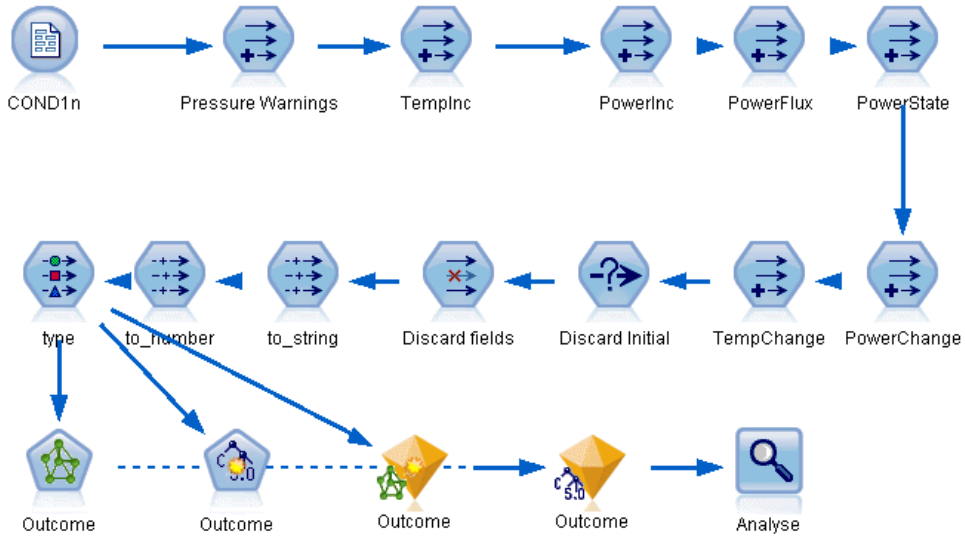
Les nuggets de modèle sont ajoutés au flux, tous deux étant connectés au noeud Typer.

▶ Repositionnez les nuggets comme indiqué, de sorte que le noeud Typer se connecte au nugget Réseau de neurones, qui se connecte au nugget C5.0.

▶ Reliez un noeud Analyse au nugget C5.0.

- Editez le nœud source d'origine pour lire le fichier *ETAT2n* (au lieu de *ETAT1n*), car *ETAT2n* contient des données de test non affichées.

Figure 20-5
Test du réseau formé



- Ouvrez le noeud Analyse, puis cliquez sur Exécuter.

  Ceci génère des résultats reflétant la précision du réseau formé et de la règle.

Chapitre
21

Classification des clients de services de télécommunications (analyse discriminante)

L'analyse discriminante est une technique statistique de classification des enregistrements sur la base des valeurs des champs d'entrée. Excepté le fait qu'elle utilise un champ cible catégoriel et non pas numérique, cette régression est semblable à la régression linéaire.

Par exemple, supposons qu'un fournisseur de télécommunications ait segmenté sa base de clientèle par modèles d'utilisation de service, classant ses clients en quatre groupes. Si les données démographiques peuvent être utilisées pour prévoir les groupes d'affectation, vous pouvez personnaliser les offres pour chaque client éventuel.

Cet exemple utilise le flux *telco_custcat_discriminant.str*, qui fait référence au fichier de données *telco.sav*. Ces fichiers sont disponibles dans le répertoire *Demos* de n'importe quelle installation de IBM® SPSS® Modeler. Ce dossier est accessible à partir du groupe de programmes IBM® SPSS® Modeler dans le menu Démarrer de Windows. Le fichier *telco_custcat_discriminant.str* se trouve dans le répertoire des *flux* .

Cet exemple est axé sur l'utilisation des données démographiques dans le but de prévoir des modèles d'utilisation. Le champ cible *custcat* possède quatre valeurs possibles qui correspondent aux quatre groupes de clients suivants :

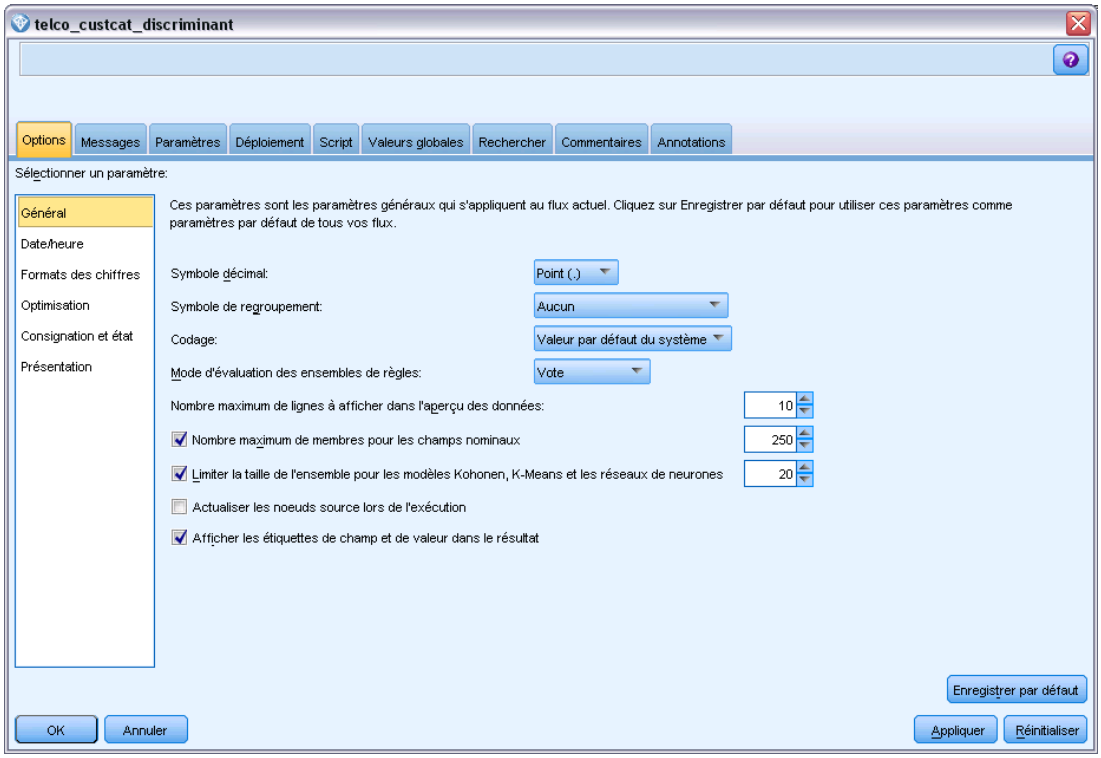
Valeur	Etiquette
1	Basic Service
2	E-Service
3	Plus Service
4	Total Service

Création du flux

- Définissez en premier lieu des propriétés de flux pour afficher les étiquettes de variable et de valeur dans la sortie. A partir du menu, sélectionnez :
  Fichier > Propriétés du flux... > Options > Général

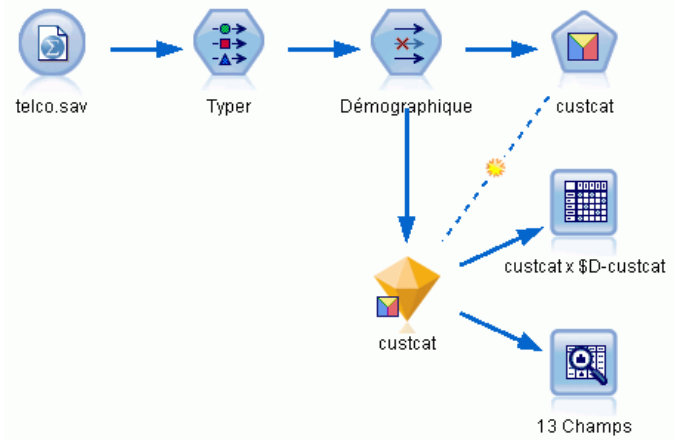
▶ Sélectionnez Afficher les étiquettes de champ et de valeur dans le résultat et cliquez sur OK.

Figure 21-1
Propriétés du flux



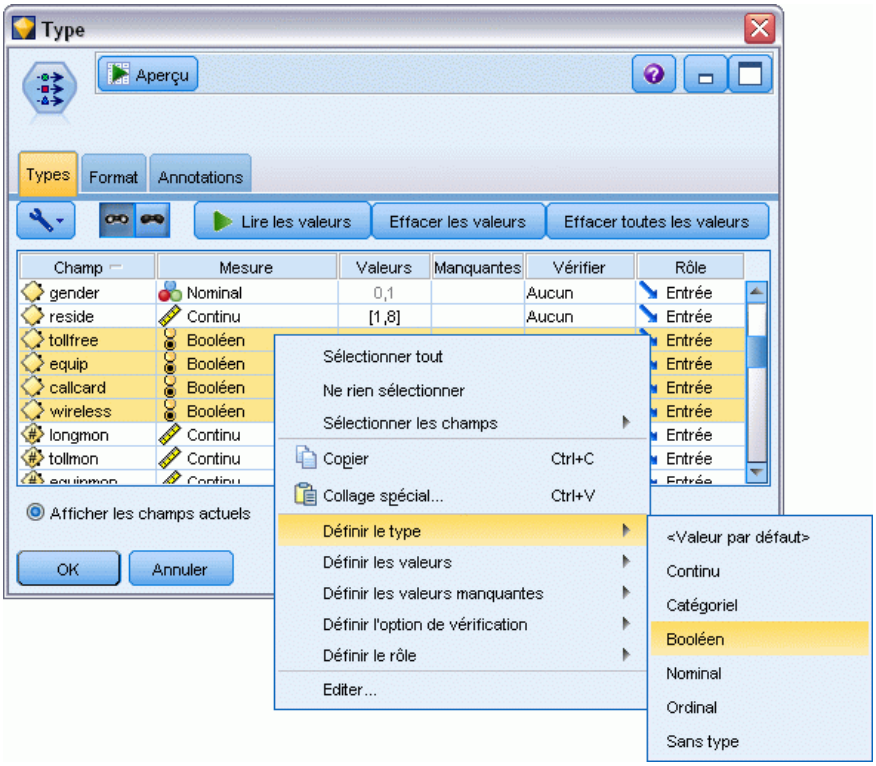
▶ Ajoutez un noeud source Fichier de statistiques pointant vers *telco.sav* dans le dossier *Demos*.

Figure 21-2
Exemple de flux permettant de classifier les clients par analyse discriminante



- Ajoutez un noeud Typer et cliquez sur Lire les valeurs, en vous assurant que tous les niveaux de mesure sont correctement paramétrés. Par exemple, la majorité des champs avec des valeurs 0 et 1 peuvent être considérés comme des champs booléens.

Figure 21-3
Configuration du niveau de mesure pour plusieurs champs



Astuce : Pour modifier les propriétés de plusieurs champs contenant des valeurs similaires (telles que 0/1), cliquez sur l'en-tête de colonne *Valeurs* afin de trier les champs en fonction de cette valeur. Maintenez la touche Maj enfoncée tout en utilisant la souris ou les touches fléchées pour sélectionner tous les champs à modifier. Cliquez ensuite sur la sélection avec le bouton droit de la souris pour modifier le niveau de mesure ou les autres attributs des champs sélectionnés.

Veuillez noter que puisqu'il est plus correct de considérer le *sexe* comme un champ avec un ensemble de deux valeurs plutôt que comme un champ booléen, laissez sa valeur de mesure sur Nominal.

▶ Définissez le rôle du champ *custcat* sur Cible. Le rôle de tous les autres champs doit être défini sur Entrée.

Figure 21-4
Définition du rôle de champ

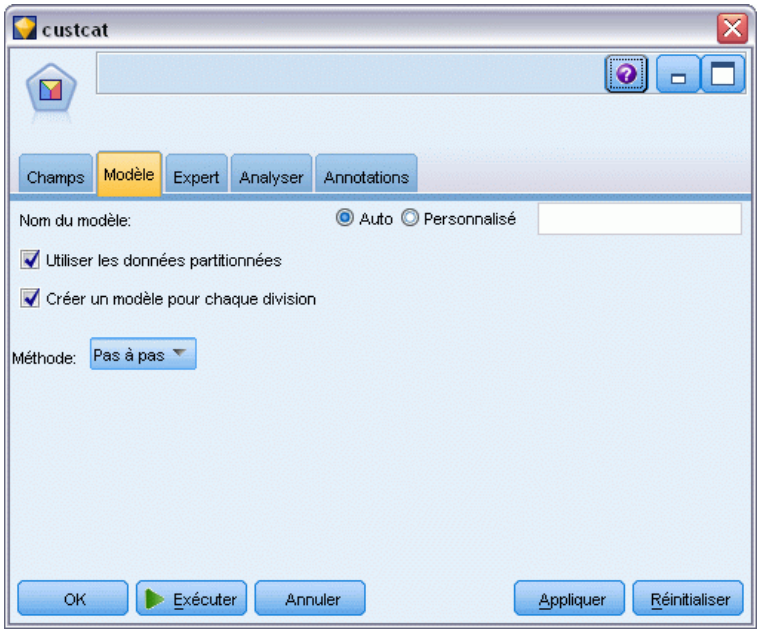
Cet exemple étant axé sur les données démographiques, utilisez un noeud Filtrer pour n'inclure que les champs pertinents (*region*, *age*, *marital*, *address*, *income*, *ed*, *employ*, *retire*, *gender*, *reside* et *custcat*). Les autres champs peuvent être exclus pour cette analyse.

Figure 21-5
Filtrage des champs démographiques

(Vous pouvez également paramétrer le rôle sur Aucun pour ces champs plutôt que de les exclure, ou sélectionner les champs que vous souhaitez utiliser dans le noeud de modélisation.)

▶ Dans le noeud discriminant, cliquez sur l'onglet Modèle et sélectionnez la méthode Pas à pas.

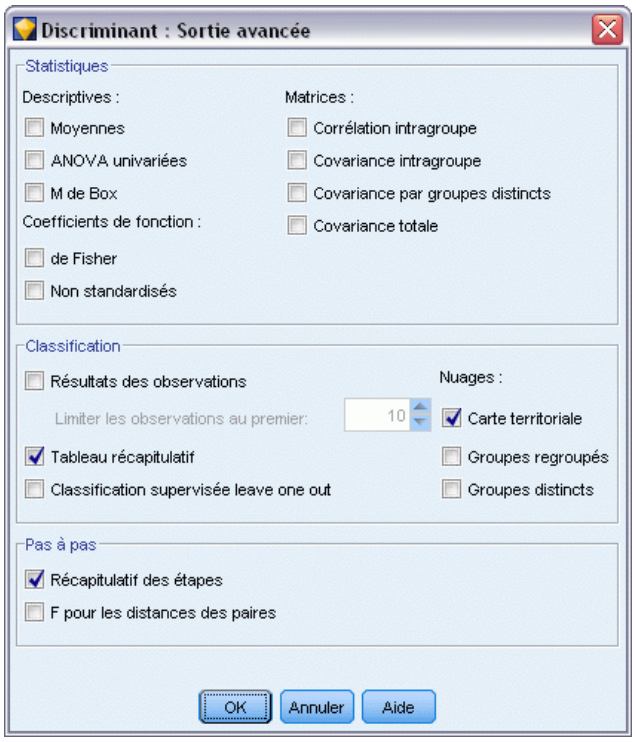
Figure 21-6
Choix des options de modèle



▶ Dans l'onglet Expert, paramétrez le mode sur Expert et cliquez sur Sortie.

▶ Sélectionnez Tableau récapitulatif, Carte territoriale et Récapitulatif des étapes dans la boîte de dialogue Sorties avancées puis cliquez sur OK.

Figure 21-7
Choix des options de sortie

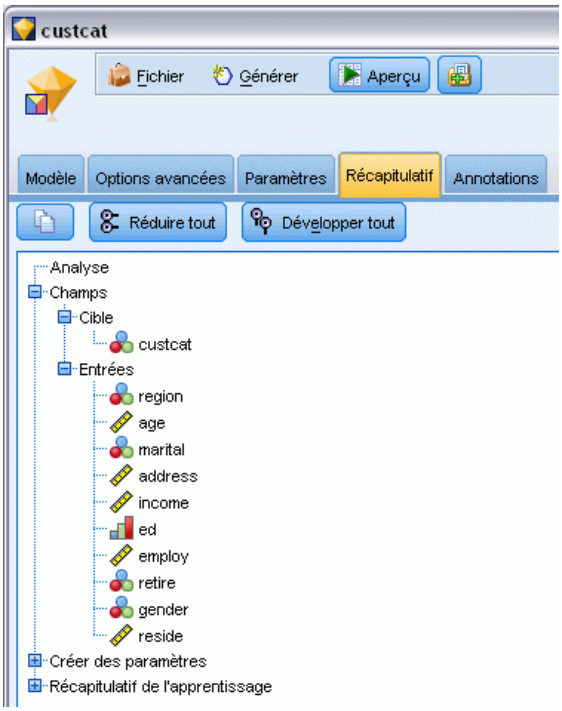


Examen du modèle

▶ Cliquez sur Exécuter pour créer le modèle qui est ajouté au flux et à la palette Modèles en haut à droite. Pour afficher ses détails, double-cliquez sur le nugget de modèle du flux.

L'onglet Récapitulatif affiche (entre autres) la cible et la liste complète des entrées (champs variables indépendantes) à examiner.

Figure 21-8
Récapitulatif du modèle avec champs cible et champs d'entrée



Pour des détails sur les résultats de l'analyse discriminante :

- Cliquez sur l'onglet Avancé.
- Cliquez sur le bouton « Lancer dans un navigateur externe » (juste en-dessous de l'onglet Modèle) pour afficher les résultats dans votre navigateur Web.

Analyse discriminante pas à pas

Figure 21-9
Variables absentes de l'analyse, étape 0

Pas		Tolérance	Tolérance minimale	F pour introduire	Lambda de Wilks
0	Age in years	1,000	1,000	7,521	,978
	Marital status	1,000	1,000	3,500	,990
	Years at current address	1,000	1,000	8,433	,975
	Household income in thousands	1,000	1,000	6,689	,980
	Level of education	1,000	1,000	61,454	,844
	Years with current employer	1,000	1,000	16,976	,951
	Retired	1,000	1,000	3,005	,991
	Gender	1,000	1,000	,373	,999
	Number of people in household	1,000	1,000	3,976	,988

Lorsque vous disposez de nombreuses valeurs prédites, la méthode pas à pas peut être utile car elle sélectionne automatiquement les « meilleures » variables à utiliser dans le modèle. La méthode pas à pas commence par un modèle qui n'inclut aucune des valeurs prédites. A chaque étape, la valeur prédite dotée de la valeur *F-to-enter* la plus importante, supérieure aux critères d'entrée (par défaut, 3,84), est ajoutée au modèle.

Figure 21-10
Variables absentes de l'analyse, étape 3

3	Age in years	,535	,535	,252	,795
	Marital status	,605	,593	1,507	,792
	Years at current address	,776	,771	3,514	,787
	Household income in thousands	,688	,657	,687	,794
	Retired	,917	,880	,353	,795
	Gender	,997	,931	,395	,795

Les variables qui ne sont toujours pas prises en compte dans l'analyse lors de la dernière étape ont toutes des valeurs *F-to-enter* inférieures à 3,84. Aucune variable supplémentaire n'est donc ajoutée.

Figure 21-11
Variables comprises dans l'analyse

Pas		Tolérance	F pour éliminer	Lambda de Wilks
1	Level of education	1,000	61,454	
2	Level of education	,953	59,108	,951
	Years with current employer	,953	14,933	,844
3	Level of education	,951	60,046	,940
	Years with current employer	,934	15,824	,834
	Number of people in household	,979	4,841	,807

Cette table affiche les statistiques des variables qui figurent dans l'analyse à chaque étape. La *tolérance* est la proportion de la variance d'une variable non justifiée par les autres variables indépendantes de l'équation. Une variable ayant une très faible tolérance n'apporte que peu d'informations à un modèle et peut générer des problèmes de calcul.

Les valeurs *F-to-remove* sont utiles pour décrire ce qui se passe si une variable est supprimée du modèle actuel (les autres variables étant conservées). La valeur *F-to-remove* de la variable entrante est identique à la valeur *F-to-enter* de l'étape précédente (affichée dans la table Variables absentes de l'analyse).

Avertissement relatif aux méthodes pas à pas

Les méthodes pas à pas sont pratiques, mais ont leurs limites. Sachez que, étant donné que les méthodes pas à pas sélectionnent des modèles uniquement sur la base du mérite statistique, elles risquent de choisir des valeurs prédites qui n'ont aucune **signification pratique**. Si vous connaissez bien les données et que vous avez des attentes particulières en ce qui concerne les valeurs prédites importantes, utilisez ces connaissances et évitez les méthodes pas à pas. Si, à l'inverse, vous avez de nombreuses valeurs prédites et que vous ne savez pas par où commencer, une analyse pas à pas et un ajustement du modèle sélectionné sont préférables à une absence complète de modèle.

Vérification de la qualité de l'ajustement

Figure 21-12
Valeurs propres

Fonction	Valeur propre	% de la variance	% cumulé	Corrélation canonique
1	,198(a)	80,2	80,2	,407
2	,048(a)	19,4	99,6	,214
3	,001(a)	,4	100,0	,031

Presque toute la variance expliquée par le modèle est liée aux deux premières fonctions discriminantes. Trois fonctions sont automatiquement ajustées. Cependant, du fait de sa valeur propre minime, vous pouvez ignorer la troisième en toute sécurité.

Figure 21-13
Lambda de Wilk

Test de la ou des fonctions	Lambda de Wilks	Khi-deux	ddl	Signification
de 1 à 3	,796	227,345	9	,000
de 2 à 3	,953	47,486	4	,000
3	,999	,929	1	,335

Le lambda de Wilk confirme que seules les deux premières fonctions sont utiles. Pour chaque ensemble de fonctions, cela permet de tester l'hypothèse selon laquelle les moyennes des fonctions répertoriées sont égales dans tous les groupes. Le test de la fonction 3 a une valeur de signification supérieure à 0,10. Par conséquent, cette fonction contribue peu au modèle.

Matrice de structure

Figure 21-14
Matrice de structure

	Fonction		
	1	**2**	**3**
Level of education	,966(*)	-,090	-,244
Years with current employer	-,182	,964(*)	-,193
Age in years(a)	-,162	,598(*)	-,285
Household income in thousands (a)	,109	,514(*)	-,190
Years at current address(a)	-,151	,394(*)	-,214
Retired(a)	-,108	,230(*)	-,137
Gender(a)	,008	,054(*)	,009
Number of people in household	,232	,097	,968(*)
Marital status(a)	,132	,134	,600(*)
Les corrélations intra-groupes combinés entre variables discriminantes et les variables des fonctions discriminantes canoniques standardisées sont ordonnées par tailles absolues des corrélations à l'intérieur de la fonction.			
*. Plus grande corrélation absolue entre chaque variable et une fonction discriminante quelconque.			
a. Cette variable n'est pas utilisée dans l'analyse.			

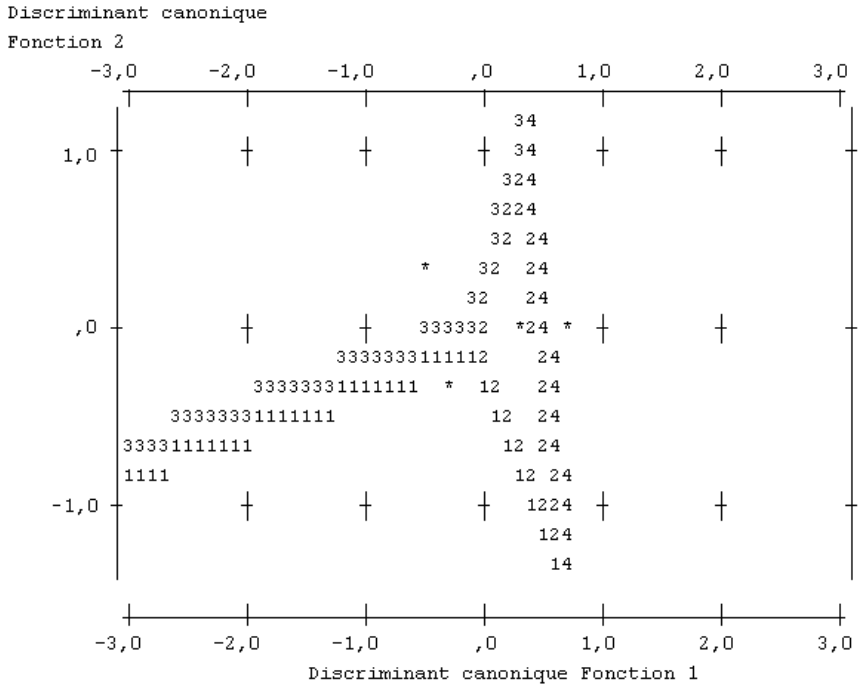
Lorsque plusieurs fonctions discriminantes existent, un astérisque (*) marque la corrélation absolue la plus importante de chaque variable avec l'une des fonctions canoniques. Dans chaque fonction, ces variables marquées sont ensuite triées en fonction de l'importance de la corrélation.

- La variable *Level of education* est la plus fortement corrélée avec la première fonction et est la seule variable la plus fortement corrélée avec cette fonction.

- Les variables *Years with current employer*, *Age in years*, *Household income in thousands*, *Years at current address*, *Retired* et *Gender* sont les plus fortement corrélées avec la deuxième fonction, bien que *Gender* et *Retired* soient plus faiblement corrélées que les autres. Les autres variables marquent cette fonction en tant que fonction de « stabilité ».
- Les variables *Number of people in household* et *Marital status* sont les plus fortement corrélées avec la troisième fonction discriminante. Cependant, cette fonction discriminante étant sans intérêt, ces valeurs prédites sont quasiment inutiles.

Carte territoriale

Figure 21-15
Carte territoriale



La carte territoriale peut vous aider à étudier les relations entre les groupes et les fonctions discriminantes. Associée aux résultats de la matrice de structure, elle donne une interprétation graphique des relations entre valeurs prédites et groupes. La première fonction, représentée sur l'axe horizontal, sépare le groupe 4 (clients *Total service*) des autres. Etant donné que la variable *Level of education* est fortement corrélée avec la première fonction de manière positive, les clients *Total service* sont, en règle générale, ceux qui ont le niveau d'éducation le plus élevé. La deuxième fonction sépare les groupes 1 et 3 (clients *Basic service* et *Plus service*). En règle générale, les clients *Plus service* ont travaillé plus longtemps et sont plus âgés que les clients *Basic service*. Les clients *E-service* ne se distinguent pas nettement des autres, bien que la carte laisse penser qu'ils ont tendance à avoir un niveau d'éducation important et une expérience professionnelle moyenne.

En général, la précision des centroïdes de groupe, marqués par des astérisques (*), par rapport aux lignes territoriales suggère que la séparation entre tous les groupes n'est pas très importante.

Seules les deux premières fonctions discriminantes sont tracées. Cependant, étant donné que la troisième fonction est relativement non significative, la carte territoriale fournit une vue complète du modèle discriminant.

Résultats de la classification supervisée

Figure 21-16
Résultats de la classification supervisée

		Customer category	Classe(s) d'affectation prévue(s)				Total
			Basic service	E-service	Plus service	Total service	
Original	Effectif	Basic service	125	11	61	69	266
		E-service	49	15	58	95	217
		Plus service	102	14	112	53	281
		Total service	40	16	37	143	236
	%	Basic service	47,0	4,1	22,9	25,9	100,0
		E-service	22,6	6,9	26,7	43,8	100,0
		Plus service	36,3	5,0	39,9	18,9	100,0
		Total service	16,9	6,8	15,7	60,6	100,0
a. 39,5% des observations originales classées correctement.							

Le lambda de Wilk vous permet de savoir que votre modèle permet d'obtenir des résultats pertinents, mais vous devez étudier les résultats de classification supervisée afin de déterminer à quel point ces résultats sont pertinents. D'après les données observées, le modèle « nul » (c'est-à-dire, un modèle sans valeurs prédites) classifie tous les clients dans le groupe modal, *Plus service*. Par conséquent, le modèle nul est correct 281/1 000 = 28,1 % du temps. Votre modèle obtient 11,4 % de plus, soit 39,5 % des clients. Votre modèle identifie particulièrement bien les clients *Total service*. Toutefois, il fonctionne très mal pour la classification des clients *E-service*. Vous devrez chercher une autre valeur prédite pour séparer ces clients.

Récapitulatif

Vous avez créé un modèle discriminant qui classe les clients dans l'un des quatre groupes d'« utilisation de service » prédéfinis, en fonction des informations démographiques collectées auprès de chacun des clients. Grâce à la matrice de structure et à la carte territoriale, vous avez identifié les variables les plus utiles à la segmentation de votre clientèle. Enfin, les résultats de classification supervisée montrent que le modèle n'est pas très performant en ce qui concerne la classification des clients *E-service*. Davantage de recherches sont nécessaires pour déterminer une autre variable de prévision classant mieux ces clients. Néanmoins, en fonction de ce que vous cherchez à prévoir, le modèle peut parfaitement correspondre à vos besoins. Par exemple, si l'identification des clients *E-service* ne vous intéresse pas, le modèle peut s'avérer assez précis pour vous. Cela peut être le cas lorsque le service en ligne est un produit d'appel qui

n'engrange que peu de bénéfices. Si, par exemple, votre retour sur investissement le plus élevé provient des clients *Plus service* ou *Total service*, il est possible que le modèle vous fournisse les informations nécessaires.

Sachez également que ces résultats sont établis uniquement d'après les données d'apprentissage. Pour évaluer comment le modèle peut se généraliser à d'autres données, vous pouvez utiliser un noeud Partitionner destiné à contenir un sous-ensemble d'enregistrements à des fins de test et de validation. Pour plus d'informations, reportez-vous à la section Noeud Partitionner dans le chapitre 4 dans *Noeuds source, exécution et de sortie de IBM SPSS Modeler 15*.

Vous trouverez des explications sur le fondement mathématique des méthodes de modélisation utilisées dans IBM® SPSS® Modeler dans le Guide des Algorithmes SPSS Modeler. Celui-ci est disponible dans le répertoire *Documentation* du disque d'installation.

Chapitre
22

Analyse de données de survie avec censure par intervalle (modèles linéaires généralisés)

Lors de l'analyse de données de survie avec censure par intervalle (c'est-à-dire lorsque l'heure exacte de l'événement d'intérêt n'est pas connue, la seule donnée connue étant qu'il a eu lieu au cours d'un intervalle donné), l'application du modèle de Cox aux risques des événements sur des intervalles aboutit à un modèle de régression log-log complémentaire.

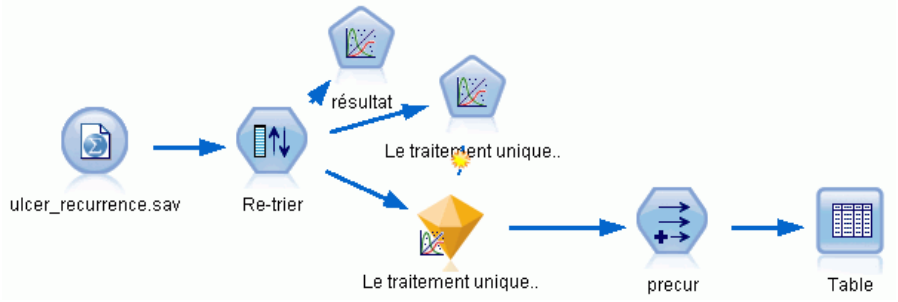
Des informations partielles, issues d'une étude visant à comparer l'efficacité de deux thérapies dans la prévention de la réapparition des ulcères, sont rassemblées dans le fichier *ulcer_recurrence.sav*. Cet ensemble de données a été présenté et analysé ailleurs. A l'aide des modèles linéaires généralisés, vous pouvez répliquer les résultats pour les modèles de régression log-log complémentaires.

Cet exemple utilise le flux nommé *ulcer_genlin.str*, qui fait référence au fichier de données *ulcer_recurrence.sav*. Le fichier de données se trouve dans le dossier *Demos* et le fichier de flux dans le sous-dossier *streams*. Pour plus d'informations, reportez-vous à la section Dossier Demos dans le chapitre 1 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.

Création du flux

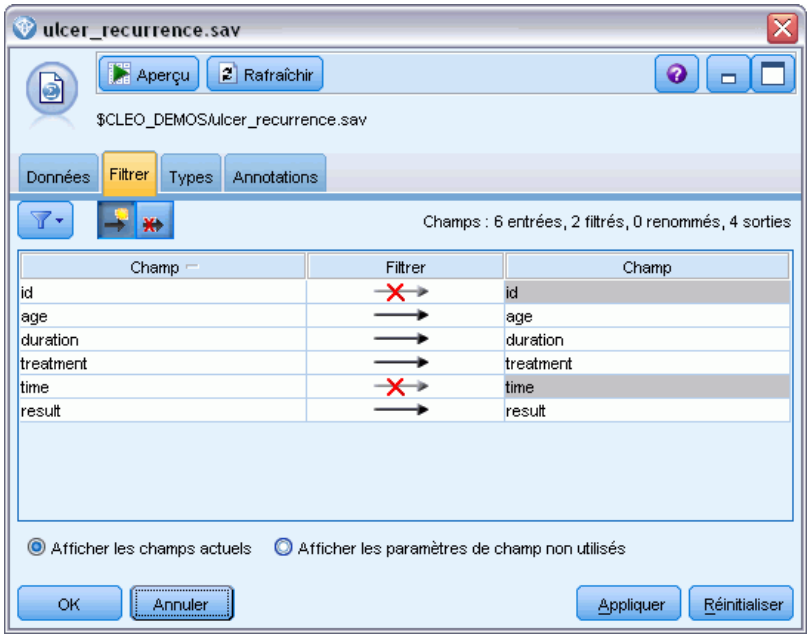
- Ajoutez un noeud Statistics pointant vers *ulcer_recurrence.sav* dans le dossier *Demos*.

Figure 22-1
Exemple de flux relatif à la prévision de la réapparition des ulcères



▶ Dans l'onglet Filtre du noeud source, excluez *id* et *time*.

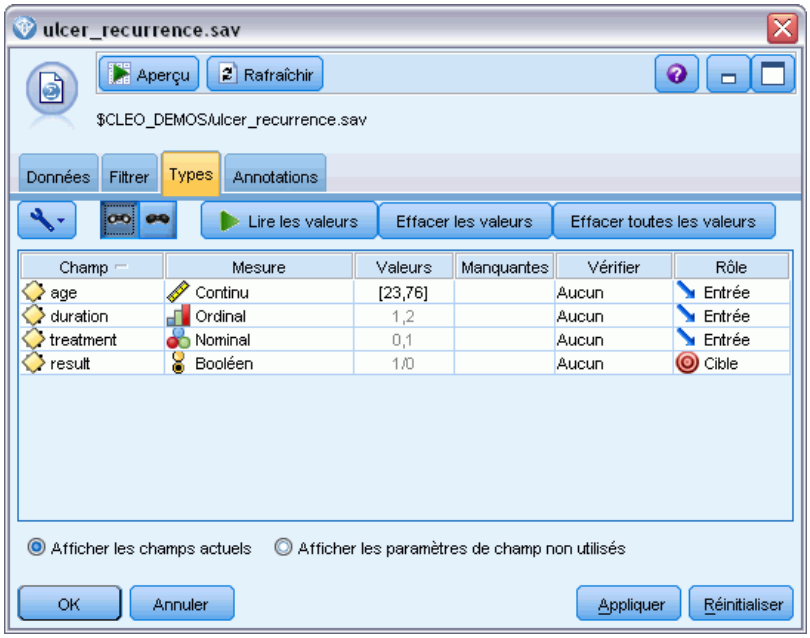
Figure 22-2
Filtrage des champs superflus



▶ Dans l'onglet Types du noeud source, définissez le rôle du champ *résultats* sur Cible et son niveau de mesure sur Booléen. Un résultat de 1 indique que l'ulcère est réapparu. Le rôle de tous les autres champs doit être défini sur Entrée.

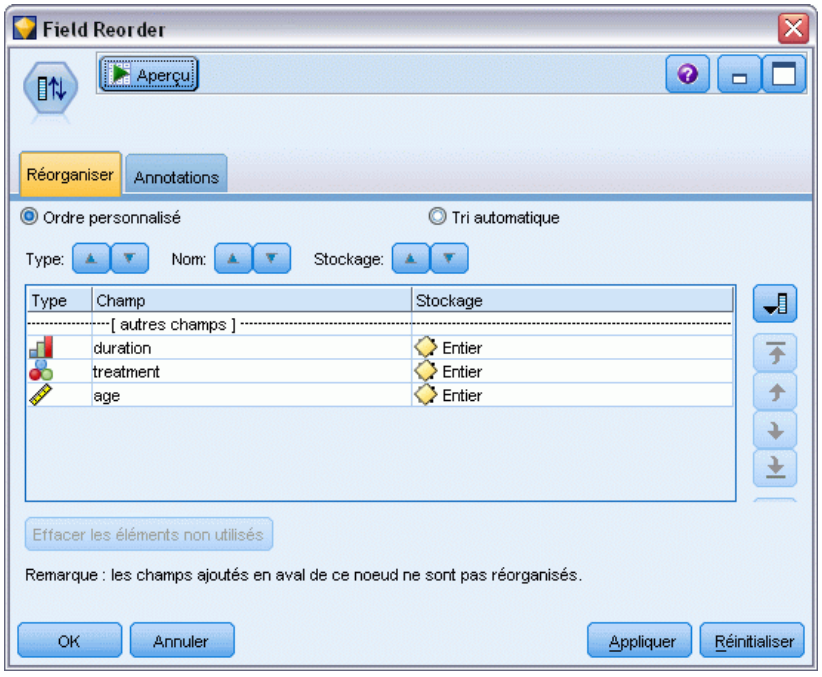
▶ Cliquez sur Lire les valeurs pour instancier les données.

Figure 22-3
Définition du rôle de champ



- Ajoutez un noeud Re-trier et spécifiez *duration*, *treatment* et *age* comme ordre des entrées. Il s'agit de l'ordre selon lequel les champs sont entrés dans le modèle ; il vous aide à répliquer les résultats de Collett.

Figure 22-4
Réorganisation des champs afin qu'ils soient entrés dans le modèle de la façon souhaitée

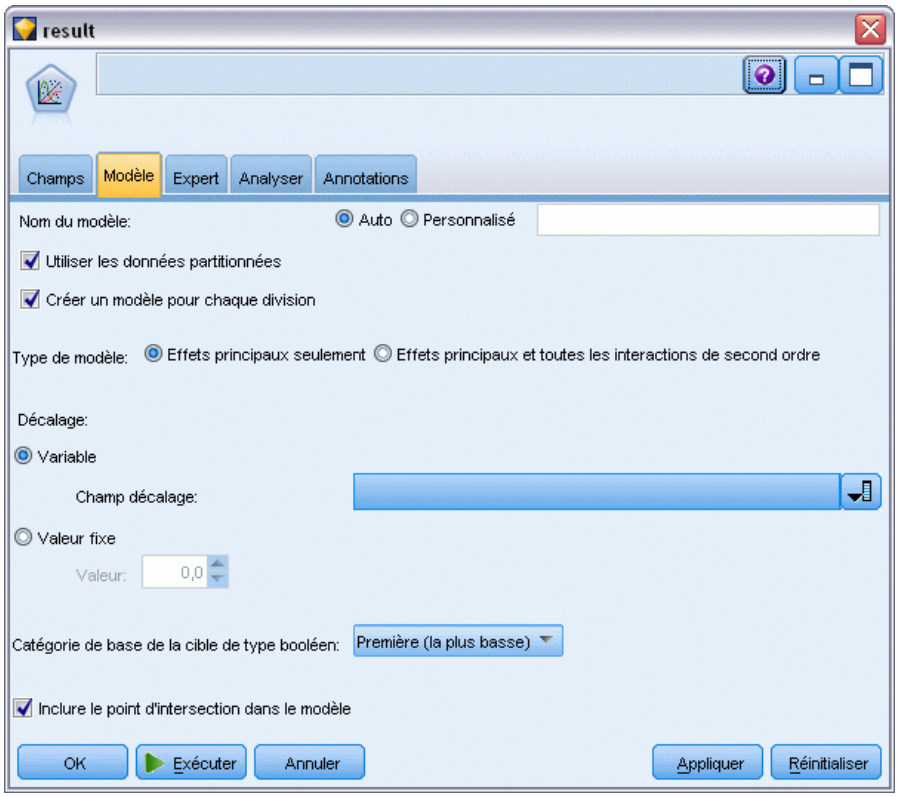


- Reliez un noeud Modèles linéaires généralisés au noeud source. Dans le noeud Modèles linéaires généralisés, cliquez sur l'onglet Modèle.

- Sélectionnez Premiers (valeur la plus faible) comme catégorie de référence pour la cible. Ce choix indique que la seconde catégorie est l'événement d'intérêt et que son effet sur le modèle se situe dans l'interprétation des estimations des paramètres. Une variable indépendante continue avec un coefficient positif indique une probabilité de réapparition accrue, avec des valeurs croissantes de la variable indépendante ; les catégories d'une variable indépendante nominale avec d'importants

coefficients indiquent une probabilité de réapparition accrue par rapport aux autres catégories de l'ensemble.

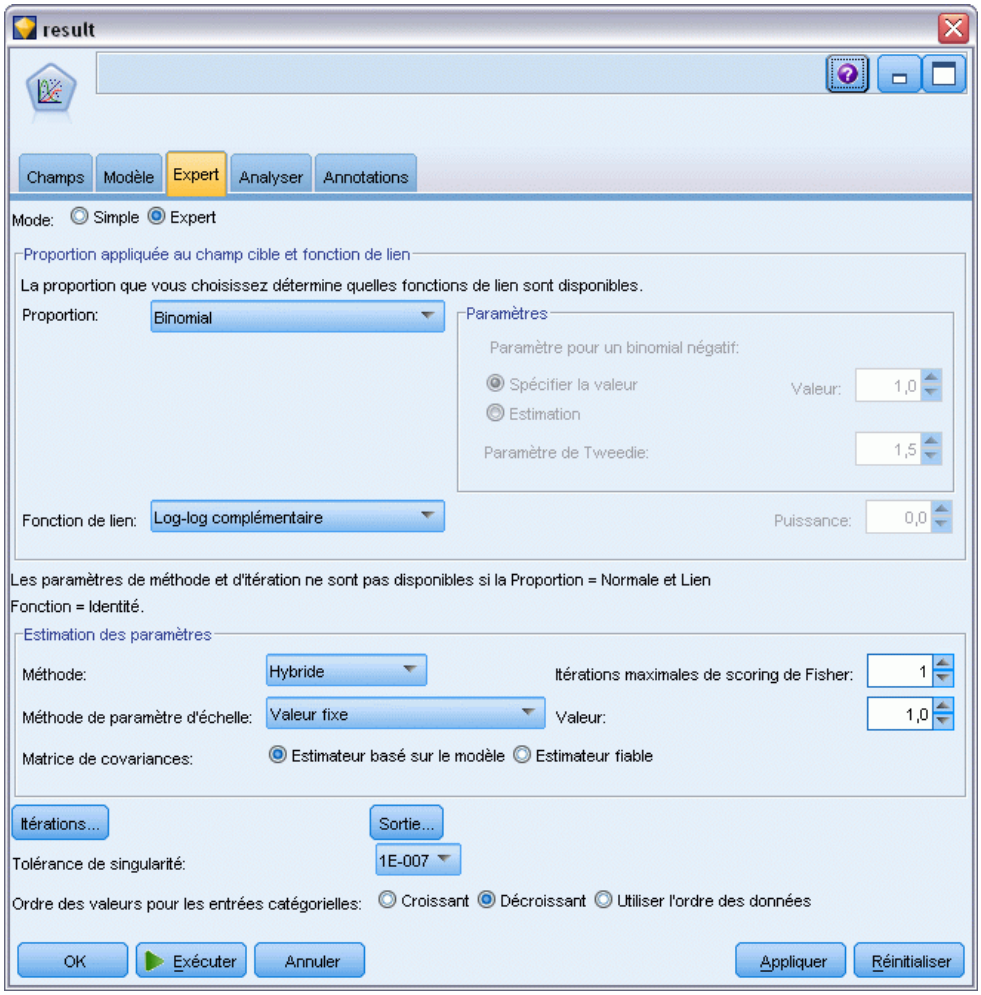
Figure 22-5
Choix des options de modèle



- Cliquez sur l'onglet Expert et sélectionnez Expert pour activer les options de modélisation expert.
- Sélectionnez Binomial pour la distribution et Log-log complémentaire pour la fonction de lien.
- Sélectionnez Valeur fixe comme méthode d'estimation du paramètre d'échelle et conservez la valeur par défaut 1,0.

- Sélectionnez l'ordre des catégories des facteurs Décroissant. Cela signifie que la première catégorie de chaque facteur constitue la catégorie de référence. L'effet de cette sélection sur le modèle porte sur l'interprétation des estimations de paramètre.

Figure 22-6
Choix des options expert



- Exécutez le flux pour créer le nugget de modèle, qui est ajouté à l'espace de travail du flux et à la palette Modèles dans l'angle supérieur droit. Pour consulter les détails du modèle, cliquez avec le bouton droit de la souris sur le nugget et sélectionnez Modifier ou Parcourir.

Tests des effets de modèle

Figure 22-7
Tests des effets pour le modèle Effets principaux

Source	Type III		
	Khi-deux de Wald	ddl	Sig.
(Ordonnée à l'origine)	,536	1	,464
duration	,003	1	,958
treatment	,382	1	,537
age	,358	1	,550
Variable dépendante : ResultModèle : (Ordonnée à l'origine), duration, treatment, age			

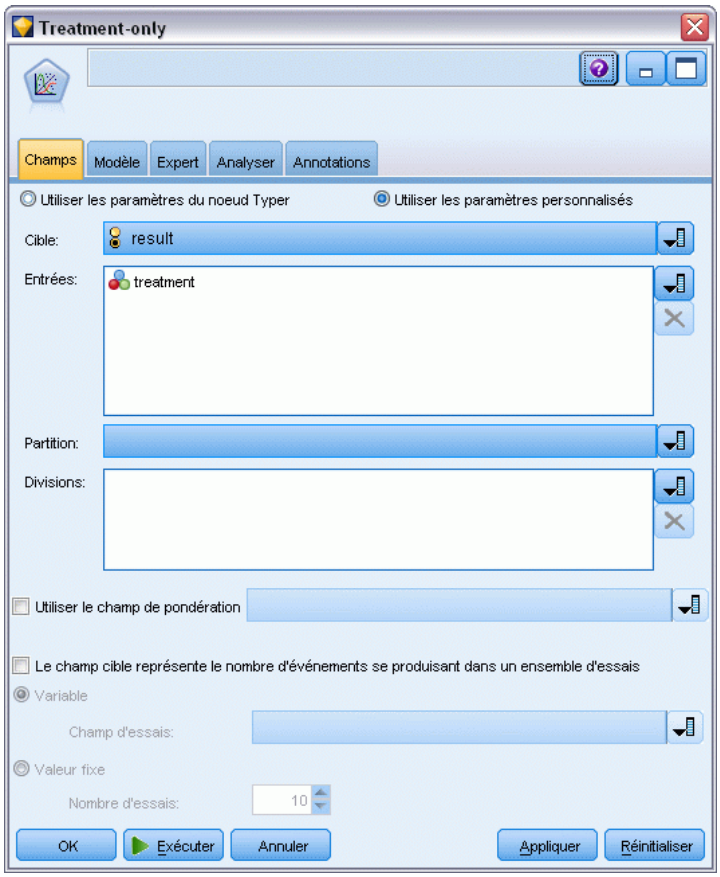
Aucun des effets du modèle n'est significatif d'un point de vue statistique ; toutefois, toute différence observable dans les effets du traitement a un intérêt du point de vue clinique. Nous allons donc ajuster un modèle réduit avec, pour seule caractéristique du modèle, le traitement.

Ajustement du modèle avec le traitement pour seule caractéristique

- Dans l'onglet Champs du noeud Modèles linéaires généralisés, cliquez sur Utiliser les paramètres personnalisés.
- Sélectionnez *result* comme cible.

▶ Sélectionnez *treatment* comme seule entrée.

Figure 22-8
Sélection des options de champ



▶ Exécutez le flux et ouvrez le nugget de modèle résultant.

Dans le nugget de modèle, sélectionnez l'onglet Avancé et accédez au bas de la liste.

Estimations des paramètres

Figure 22-9
Estimations des paramètres pour le modèle avec le traitement pour seule caractéristique

Paramètre	B	Erreur standard	Intervalle de confiance de Wald à 95 %		Test d'hypothèse		
			Inférieur	Supérieur	Khi-deux de Wald	ddl	Sig.
(Ordonnée à l'origine)	-1,442	,5012	-2,425	-,460	8,282	1	,004
[treatment=1]	,378	,6288	-,855	1,610	,361	1	,548
[treatment=0]	0(a)	.	.	.	.	.	.
(Échelle)	1(b)						

Variable dépendante : ResultModèle : (Ordonnée à l'origine), treatment, décalage = 0

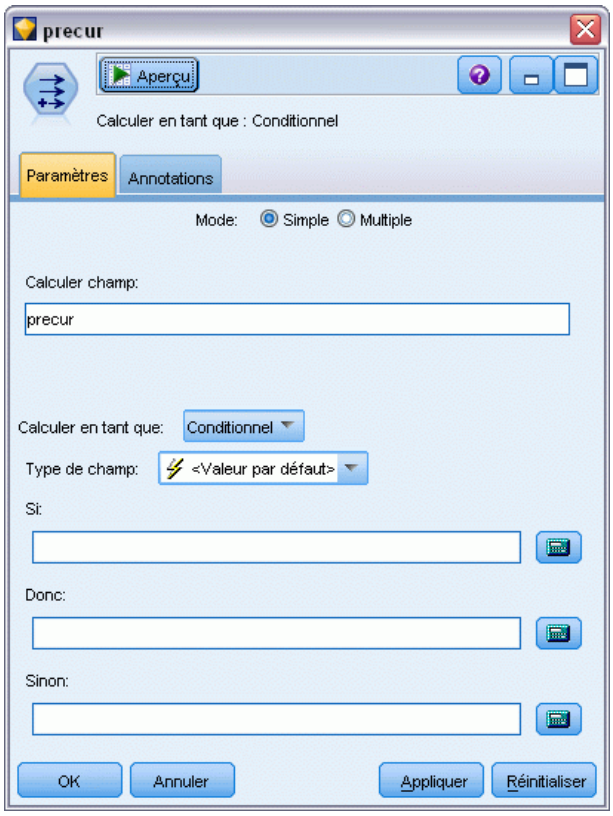
a. Défini sur zéro car ce paramètre est redondant.

b. Fixé à la valeur affichée.

L'effet du traitement (la différence, pour la variable indépendante linéaire, entre les deux niveaux de traitement ; c'est-à-dire le coefficient de *[treatment=1]*) n'est toujours pas significatif d'un point de vue statistique. Il suggère uniquement que le traitement *A [treatment=0]* semble meilleur que le traitement *B [treatment=1]* car l'estimation des paramètres pour le traitement *B* est supérieure à celle du traitement *A* et est donc associée à une probabilité de réapparition accrue dans les 12 premiers mois. La variable indépendante linéaire (constante + effet du traitement) est une estimation de $\log(-\log(1-P(recur_{12,t}))$, où $P(recur_{12,t})$ est la probabilité de réapparition à 12 mois pour le traitement t(=*A* ou *B*). Ces probabilités prédites sont générées pour chaque observation de l'ensemble de données.

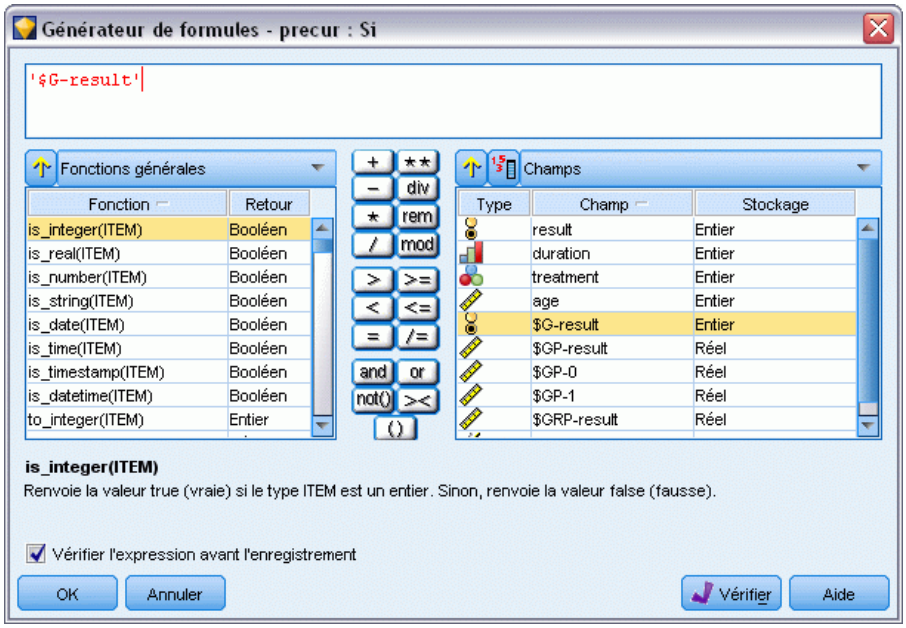
Réapparition prédite et probabilités de survie

Figure 22-10
Options des paramètres du noeud Calculer



- Pour chaque patient, le modèle détermine le score du résultat prédit et de la probabilité de ce résultat prédit. De façon à visualiser les probabilités de réapparition prédites, copiez le modèle généré dans la palette et reliez un noeud Calculer.
- Dans l'onglet Paramètres, saisissez le champ de calcul precur.
- Choisissez de le calculer comme champ Conditionnel.
- Cliquez sur le bouton représentant une calculatrice pour ouvrir le Générateur de formules pour la condition If (Si).

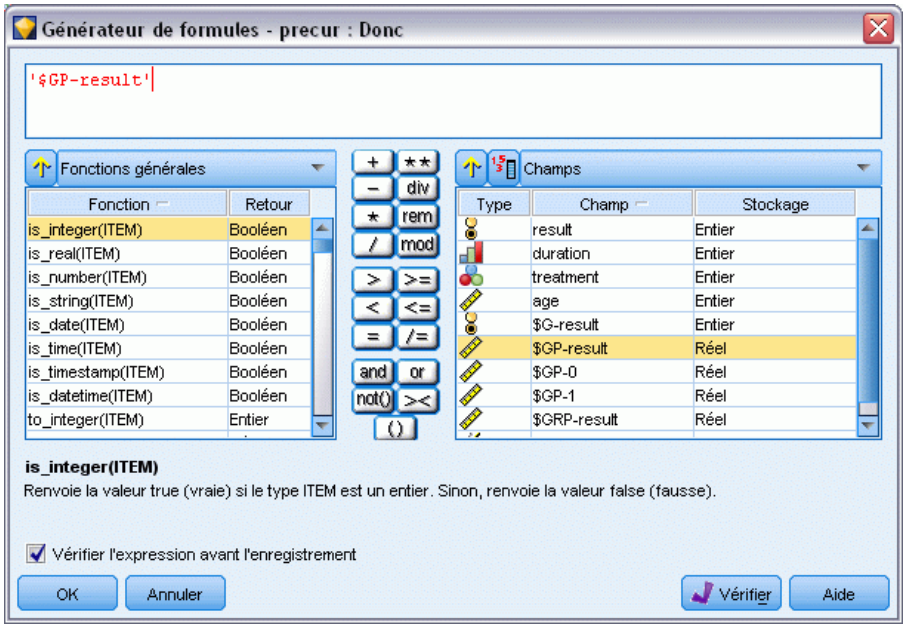
Figure 22-11
Noeud Calculer : Générateur de formules pour la condition Si



- Insérez le champ *$G-result* dans la formule.
- Cliquez sur OK.

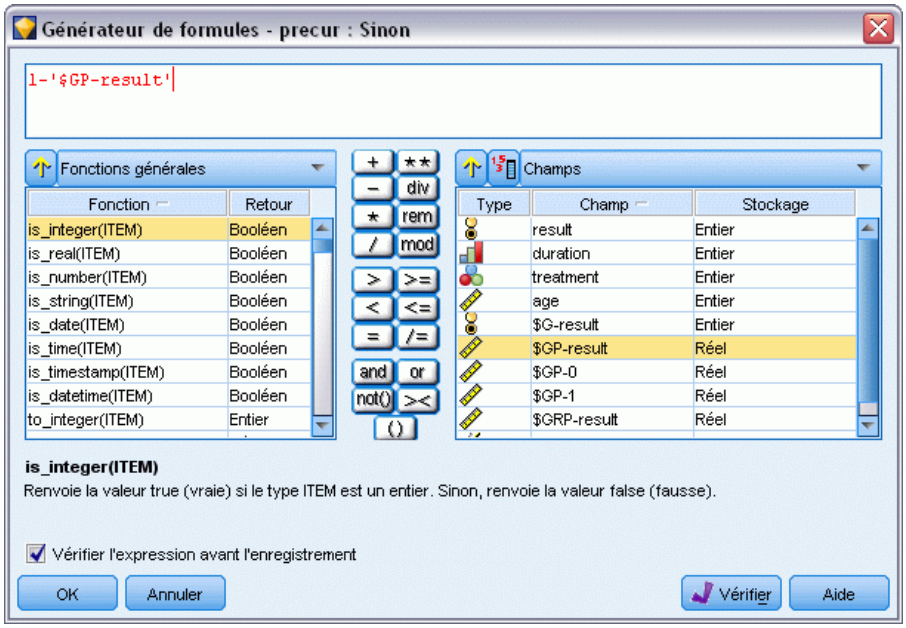
Le champ de calcul *precur* prendra la valeur de la formule Then (Donc) lorsque *$G-result* est égal à 1 et la valeur de la formule Else (Sinon) lorsqu'il est égal à 0.

Figure 22-12
Noeud Calculer : Générateur de formules pour la formule Then



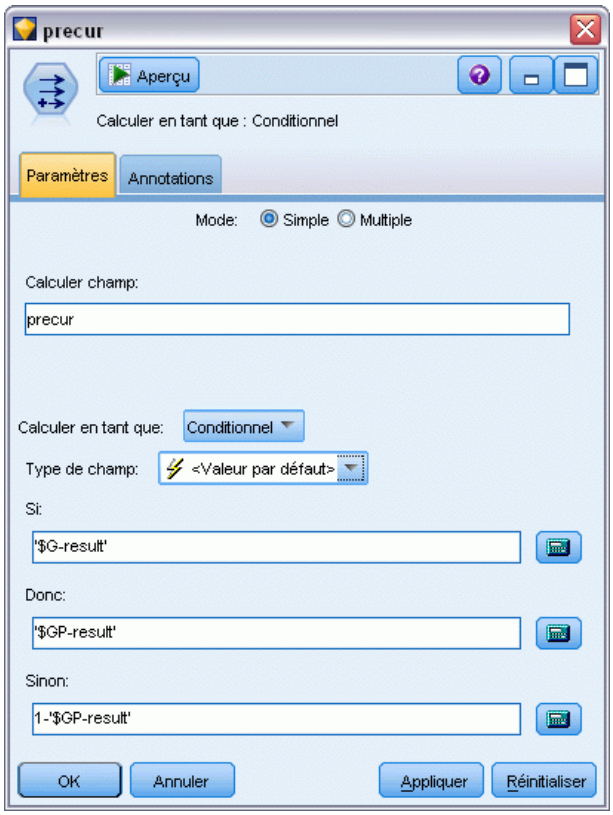
- Cliquez sur le bouton représentant une calculatrice pour ouvrir le Générateur de formules pour la formule Then.
- Insérez le champ *$GP-result* dans la formule.
- Cliquez sur OK.

Figure 22-13
Noeud Calculer : Générateur de formules pour la formule Else



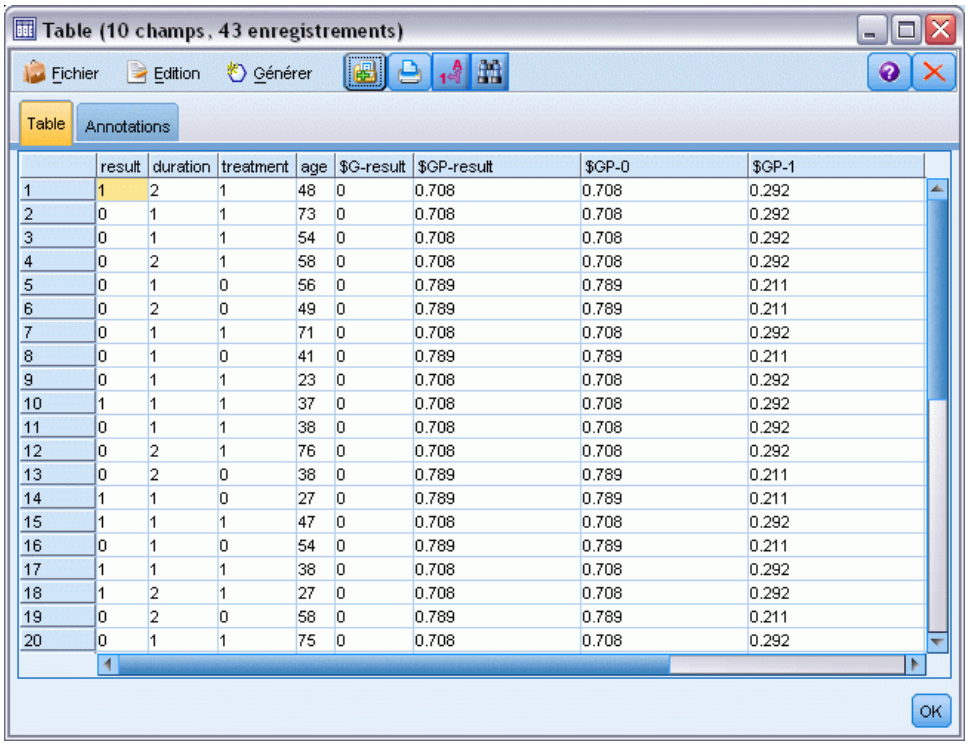
- Cliquez sur le bouton représentant une calculatrice pour ouvrir le Générateur de formules pour la formule Else.
- Saisissez 1- dans la formule puis insérez-y le champ *$GP-result*.
- Cliquez sur OK.

Figure 22-14
Options des paramètres du noeud Calculer



► Liez un noeud Table au noeud Calculer et exécutez-le.

Figure 22-15
Probabilités prévues



	result	duration	treatment	age	$G-result	$GP-result	$GP-0	$GP-1
1	1	2	1	48	0	0.708	0.708	0.292
2	0	1	1	73	0	0.708	0.708	0.292
3	0	1	1	54	0	0.708	0.708	0.292
4	0	2	1	58	0	0.708	0.708	0.292
5	0	1	0	56	0	0.789	0.789	0.211
6	0	2	0	49	0	0.789	0.789	0.211
7	0	1	1	71	0	0.708	0.708	0.292
8	0	1	0	41	0	0.789	0.789	0.211
9	0	1	1	23	0	0.708	0.708	0.292
10	1	1	1	37	0	0.708	0.708	0.292
11	0	1	1	38	0	0.708	0.708	0.292
12	0	2	1	76	0	0.708	0.708	0.292
13	0	2	0	38	0	0.789	0.789	0.211
14	1	1	0	27	0	0.789	0.789	0.211
15	1	1	1	47	0	0.708	0.708	0.292
16	0	1	0	54	0	0.789	0.789	0.211
17	1	1	1	38	0	0.708	0.708	0.292
18	1	2	1	27	0	0.708	0.708	0.292
19	0	2	0	58	0	0.789	0.789	0.211
20	0	1	1	75	0	0.708	0.708	0.292

La probabilité estimée que les patients voient réapparaître leur ulcère dans les 12 premiers mois est de 0,211 pour les patients recevant le traitement *A* et de 0,292 pour ceux qui reçoivent le traitement *B*. Notez que $1-P(recur_{12,\ t})$ est la probabilité de survie à 12 mois, laquelle peut s'avérer plus intéressante pour les analystes de la survie.

Modélisation de la probabilité de réapparition par période

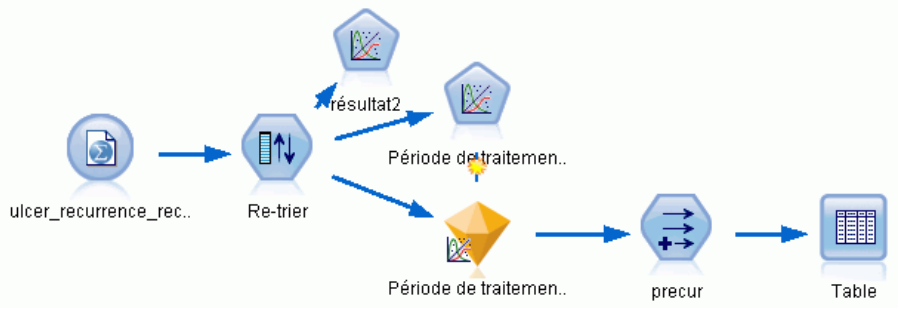
Ce modèle présente un inconvénient : il ignore les informations recueillies lors du premier examen. En effet, chez de nombreux patients, l'ulcère n'est pas réapparu durant les six premiers mois. Pour obtenir un « meilleur » modèle, il serait nécessaire de modéliser une réponse binaire qui enregistre si l'événement est survenu ou n'est pas survenu durant chaque intervalle. L'ajustement de ce modèle nécessite une reconstruction de l'ensemble de données d'origine, disponible dans le fichier *ulcer_recurrence_recoded.sav*. Pour plus d'informations, reportez-vous à la section Dossier Demos dans le chapitre 1 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*. Ce fichier contient deux variables supplémentaires :

- *Period*, qui enregistre si l'observation correspond à la période du premier ou du second examen
- *Result by period*, qui enregistre si une réapparition est survenue ou non pour le patient donné durant la période donnée.

Chaque observation d'origine (le patient) fournit une observation pour chaque intervalle où il demeure dans l'ensemble de risques. Par exemple, le patient 1 fournit deux observations : une observation pour la période du premier examen durant laquelle aucune réapparition n'est survenue et une observation pour la période du second examen durant laquelle une réapparition a été enregistrée. Le patient 10, en revanche, ne fournit qu'une seule observation car une réapparition a été enregistrée dans la première période. Les patients 16, 28 et 34 ont abandonné l'étude après six mois et ne fournissent donc qu'une seule observation au nouvel ensemble de données.

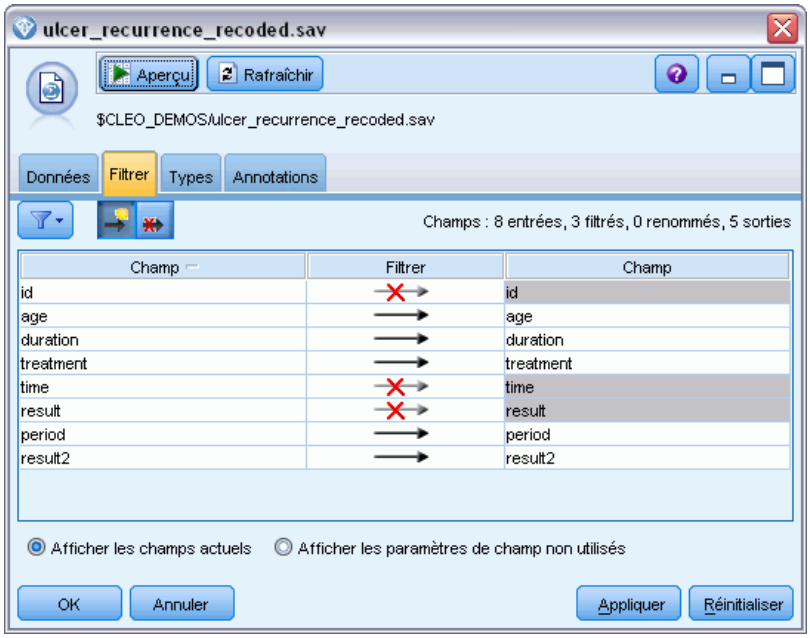
▶ Ajoutez un noeud Statistics pointant vers *ulcer_recurrence_recoded.sav* dans le dossier *Demos*.

Figure 22-16
Exemple de flux relatif à la prévision de la réapparition des ulcères



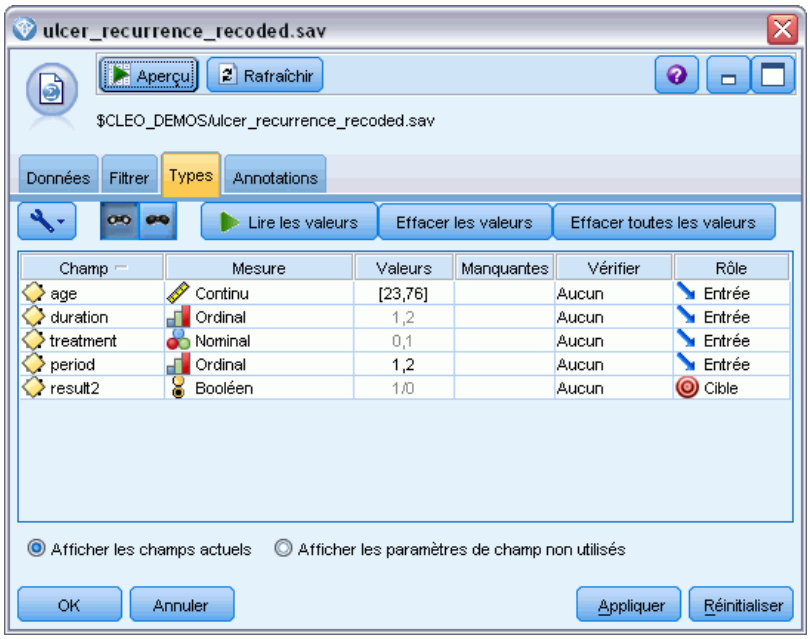
▶ Dans l'onglet Filtrer du noeud source, excluez *id*, *time* et *result*.

Figure 22-17
Filtrage des champs superflus



- Dans l'onglet Types du noeud source, définissez le rôle du champ *result2* sur Cible et son niveau de mesure sur Booléen. Le rôle de tous les autres champs doit être défini sur Entrée.

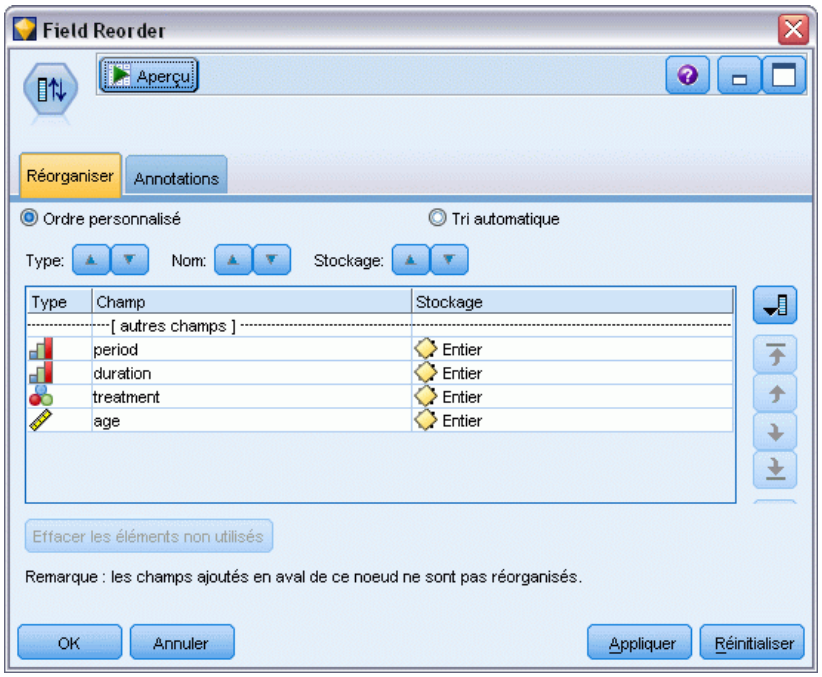
Figure 22-18
Définition du rôle de champ



- Ajoutez un noeud Re-trier et spécifiez *period*, *duration*, *treatment* et *age* comme ordre des entrées. Le fait d'avoir *period* comme première entrée (et de ne pas inclure la caractéristique de constante

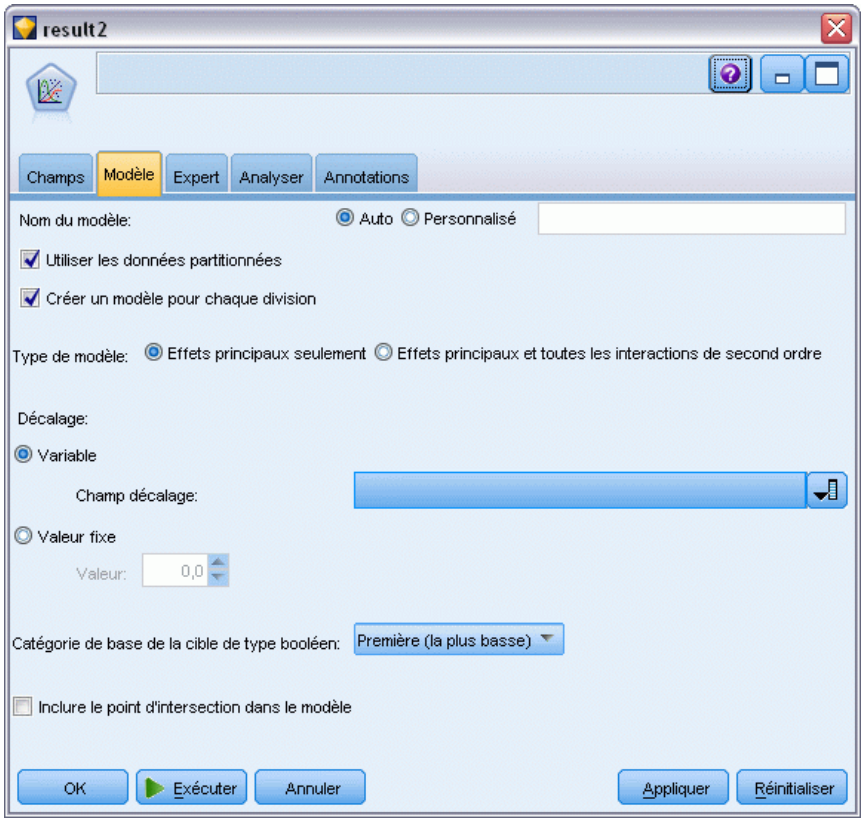
dans le modèle) permet d'ajuster un ensemble complet de variables factices pour capturer les effets de la période.

Figure 22-19
Réorganisation des champs afin qu'ils soient entrés dans le modèle de la façon souhaitée



▶ Dans le noeud Modèles linéaires généralisés, cliquez sur l'onglet Modèle.

Figure 22-20
Choix des options de modèle

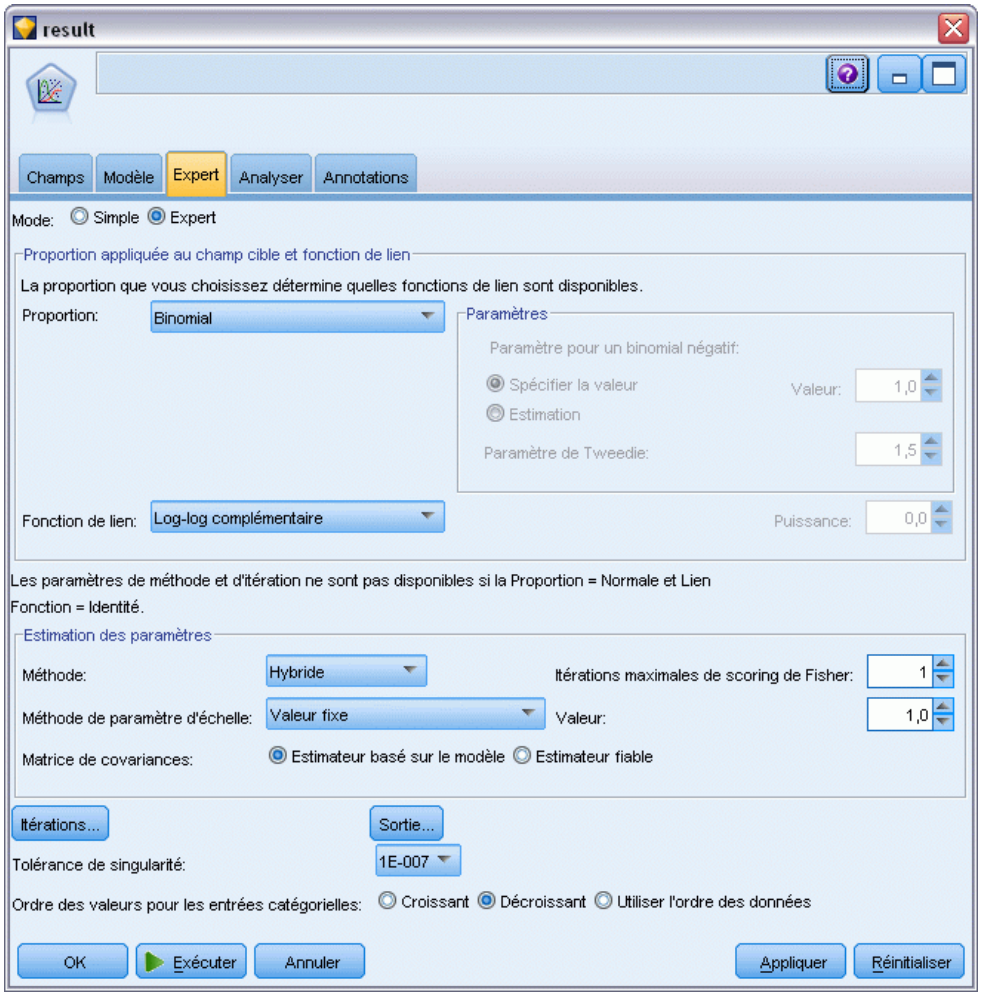


▶ Sélectionnez Premiers (valeur la plus faible) comme catégorie de référence pour la cible. Ce choix indique que la seconde catégorie est l'événement d'intérêt et que son effet sur le modèle se situe dans l'interprétation des estimations des paramètres.

▶ Désélectionnez Inclure une constante au modèle.

▶ Cliquez sur l'onglet Expert et sélectionnez Expert pour activer les options de modélisation expert.

Figure 22-21
Choix des options expert



▶ Sélectionnez Binomial pour la distribution et Log-log complémentaire pour la fonction de lien.

▶ Sélectionnez Valeur fixe comme méthode d'estimation du paramètre d'échelle et conservez la valeur par défaut 1,0.

▶ Sélectionnez l'ordre des catégories des facteurs Décroissant. Cela signifie que la première catégorie de chaque facteur constitue la catégorie de référence. L'effet de cette sélection sur le modèle porte sur l'interprétation des estimations de paramètre.

▶ Exécutez le flux pour créer le nugget de modèle, qui est ajouté à l'espace de travail du flux et à la palette Modèles dans l'angle supérieur droit. Pour consulter les détails du modèle, cliquez avec le bouton droit de la souris sur le nugget et sélectionnez Modifier ou Parcourir.

Tests des effets de modèle

Figure 22-22
Tests des effets pour le modèle Effets principaux

Source	Type III		
	Khi-deux de Wald	ddl	Sig.
period	,464	1	,496
duration	,000	1	,988
treatment	,117	1	,732
age	,314	1	,575
Variable dépendante : Result by periodModèle : period, duration, treatment, age			

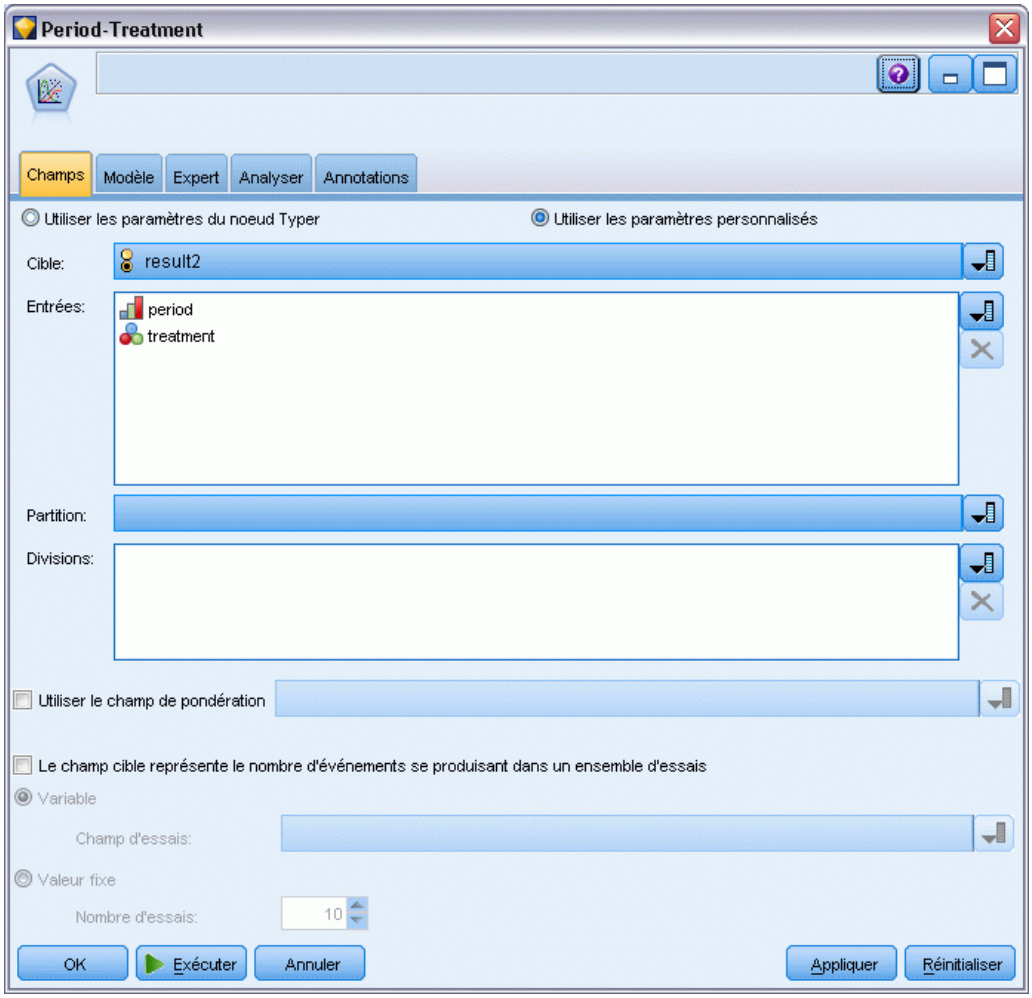
Aucun des effets du modèle n'est significatif d'un point de vue statistique ; toutefois, toute différence observable dans les effets de la période et du traitement a un intérêt du point de vue clinique ; Nous allons donc ajuster un modèle réduit, avec ces seules caractéristiques de modèle.

Ajustement du modèle réduit

- Dans l'onglet Champs du noeud Modèles linéaires généralisés, cliquez sur Utiliser les paramètres personnalisés.
- Sélectionnez *result2* comme cible.

▶ Sélectionnez *period* et *treatment* comme entrées.

Figure 22-23
Sélection des options de champ



▶ Exécutez le noeud et parcourez le modèle généré. Ensuite, copiez le modèle généré dans la palette, reliez un noeud Table et exécutez-le.

Estimations des paramètres

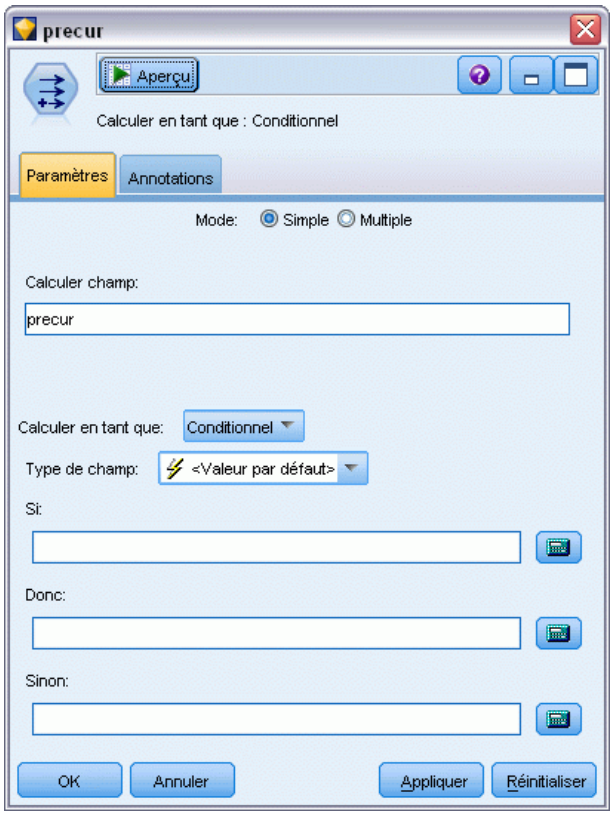
Figure 22-24
Estimations des paramètres pour le modèle avec le traitement pour seule caractéristique

Paramètre	B	Erreur standard	Intervalle de confiance de Wald à 95 %		Test d'hypothèse		
			Inférieur	Supérieur	Khi-deux de Wald	ddl	Sig.
[period=2]	-1,794	,5792	-2,929	-,659	9,597	1	,002
[period=1]	-2,206	,5912	-3,365	-1,047	13,926	1	,000
[treatment=1]	,195	,6279	-1,035	1,426	,097	1	,756
[treatment=0]	0(a)	.	.	.	.	.	.
(Échelle)	1(b)						
Variable dépendante : Result by periodModèle : period, treatment							
a. Défini sur zéro car ce paramètre est redondant.							
b. Fixé à la valeur affichée.							

L'effet du traitement n'est toujours pas significatif d'un point de vue statistique. Il suggère uniquement que le traitement *A* semble meilleur que le traitement *B* car l'estimation des paramètres pour le traitement *B* est associée à une probabilité de réapparition accrue durant les 12 premiers mois. Les valeurs de période sont, d'un point de vue statistique, significativement différentes de 0, mais ceci est dû au fait qu'un terme de constante n'est pas ajusté. L'effet de la période (la différence entre les valeurs de la variable indépendante linéaire de *[period=1]* et *[period=2]*) n'est pas significatif, d'un point de vue statistique, comme l'indiquent les tests des effets du modèle. La variable indépendante linéaire (effet de la période + effet du traitement) est une estimation de $\log(-\log(1-P(recur_{p,\,t}))$, où $P(recur_{p,\,t})$ est la probabilité de réapparition au cours de la période p(=1 ou 2, représentant six mois ou 12 mois), compte tenu du traitement t(=*A* ou *B*). Ces probabilités prédites sont générées pour chaque observation de l'ensemble de données.

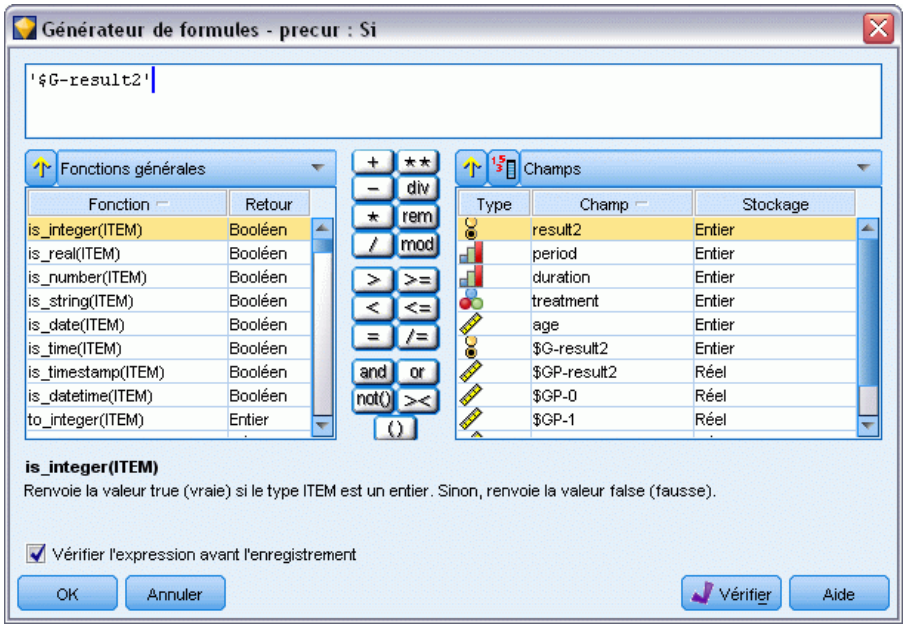
Réapparition prédite et probabilités de survie

Figure 22-25
Options des paramètres du noeud Calculer



- Pour chaque patient, le modèle détermine le score du résultat prédit et de la probabilité de ce résultat prédit. De façon à visualiser les probabilités de réapparition prédites, copiez le modèle généré dans la palette et reliez un noeud Calculer.
- Dans l'onglet Paramètres, saisissez le champ de calcul precur.
- Choisissez de le calculer comme champ Conditionnel.
- Cliquez sur le bouton représentant une calculatrice pour ouvrir le Générateur de formules pour la condition If (Si).

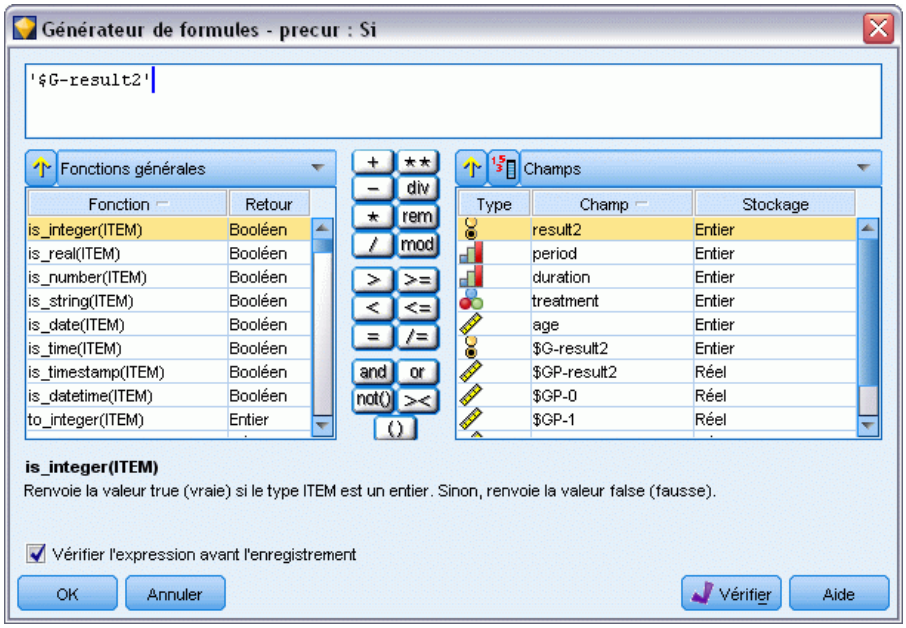
Figure 22-26
Noeud Calculer : Générateur de formules pour la condition Si



- Insérez le champ *$G-result2* dans la formule.
- Cliquez sur OK.

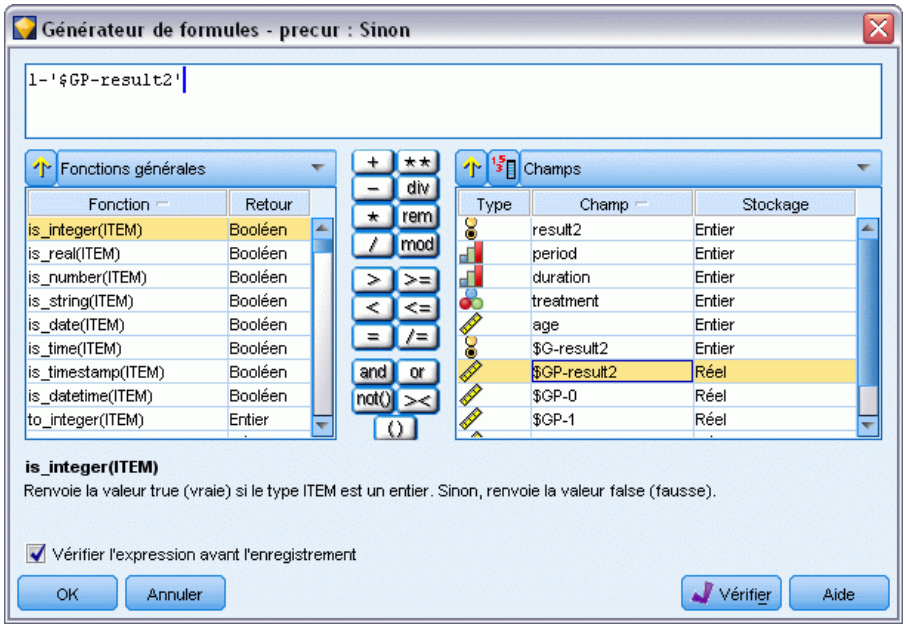
Le champ de calcul *precur* prendra la valeur de la formule Then lorsque *$G-result2* est égal à 1 et la valeur de la formule Else lorsqu'il est égal à 0.

Figure 22-27
Noeud Calculer : Générateur de formules pour la formule Then



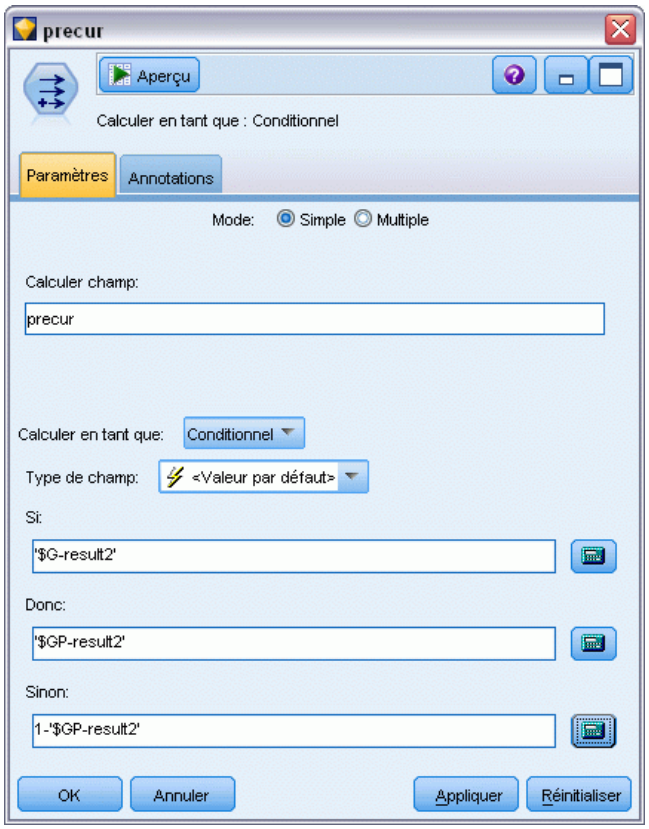
- Cliquez sur le bouton représentant une calculatrice pour ouvrir le Générateur de formules pour la formule Then.
- Insérez le champ *$GP-result2* dans la formule.
- Cliquez sur OK.

Figure 22-28
Noeud Calculer : Générateur de formules pour la formule Else



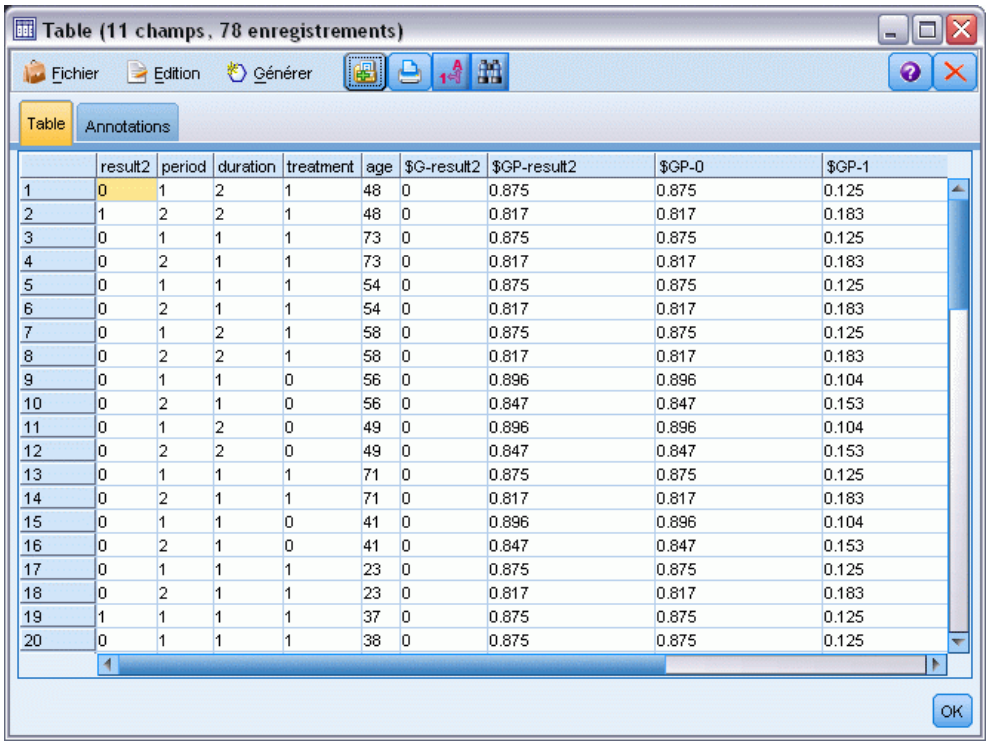
- Cliquez sur le bouton représentant une calculatrice pour ouvrir le Générateur de formules pour la formule Else.
- Saisissez 1- dans la formule puis insérez-y le champ *$GP-result2*.
- Cliquez sur OK.

Figure 22-29
Options des paramètres du noeud Calculer



▶ Liez un noeud Table au noeud Calculer et exécutez-le.

Figure 22-30
Probabilités prévues



	result2	period	duration	treatment	age	$G-result2	$GP-result2	$GP-0	$GP-1
1	0	1	2	1	48	0	0.875	0.875	0.125
2	1	2	2	1	48	0	0.817	0.817	0.183
3	0	1	1	1	73	0	0.875	0.875	0.125
4	0	2	1	1	73	0	0.817	0.817	0.183
5	0	1	1	1	54	0	0.875	0.875	0.125
6	0	2	1	1	54	0	0.817	0.817	0.183
7	0	1	2	1	58	0	0.875	0.875	0.125
8	0	2	2	1	58	0	0.817	0.817	0.183
9	0	1	1	0	56	0	0.896	0.896	0.104
10	0	2	1	0	56	0	0.847	0.847	0.153
11	0	1	2	0	49	0	0.896	0.896	0.104
12	0	2	2	0	49	0	0.847	0.847	0.153
13	0	1	1	1	71	0	0.875	0.875	0.125
14	0	2	1	1	71	0	0.817	0.817	0.183
15	0	1	1	0	41	0	0.896	0.896	0.104
16	0	2	1	0	41	0	0.847	0.847	0.153
17	0	1	1	1	23	0	0.875	0.875	0.125
18	0	2	1	1	23	0	0.817	0.817	0.183
19	1	1	1	1	37	0	0.875	0.875	0.125
20	0	1	1	1	38	0	0.875	0.875	0.125

Il est possible de résumer les probabilités de réapparition estimées de la façon suivante :

Traitement	**6 mois**	**12 mois**
S	0.104	0.153
B	0.125	0.183

A partir de ces données, la probabilité de survie sur 12 mois peut être estimée sous la forme $1-(P(recur_{1,t}) + P(recur_{2,t})\times(1-P(recur_{1,t})))$; par conséquent, pour chaque traitement :

A : 1 − (0.104 + 0.153*0.896) = 0.759

B : 1 − (0.125 + 0.183*0.875) = 0.715

ce qui montre de nouveau une préférence, significative d'un point de vue autre que statistique, pour *A* comme étant le meilleur traitement.

Récapitulatif

A l'aide des modèles linéaires généralisés, vous avez ajusté une série de modèles de régression log-log complémentaires pour des données de survie avec censure par intervalle. Même si le choix du traitement *A* est privilégié, l'obtention d'un résultat significatif du point de vue statistique

peut nécessiter une étude plus importante. Toutefois, d'autres pistes sont à explorer avec les données existantes.

- Il peut être utile de réajuster le modèle avec des effets d'interaction, notamment entre *Period* et *Treatment group*.

Vous trouverez des explications sur le fondement mathématique des méthodes de modélisation utilisées dans IBM® SPSS® Modeler dans le *Guide des Algorithmes SPSS Modeler*.

Chapitre
23

Utilisation de la régression de Poisson pour analyser les taux de dommage aux navires (modèles linéaires généralisés)

Un modèle linéaire généralisé peut être utilisé pour ajuster une régression de Poisson pour l'analyse des données d'effectif. Par exemple, un ensemble de données présenté et analysé ailleurs () relate les dommages que les vagues causent aux cargos. Le nombre d'incidents peut être modélisé comme se produisant selon l'effectif défini par un test Poisson en fonction des valeurs des variables indépendantes, et le modèle résultant peut permettre de déterminer les types de navire les plus exposés aux dommages.

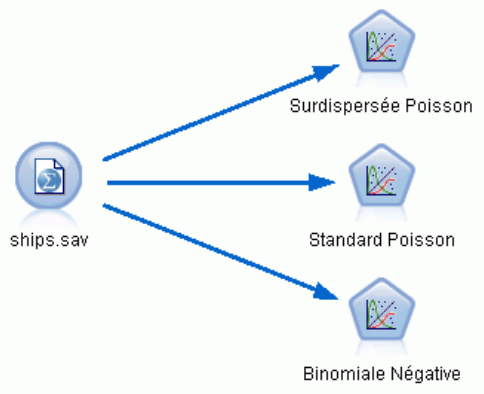
Cet exemple utilise le flux *ships_genlin.str*, qui fait référence au fichier de données *ships.sav*. Le fichier de données se trouve dans le dossier *Demos* et le fichier de flux dans le sous-dossier *streams*. Pour plus d'informations, reportez-vous à la section Dossier Demos dans le chapitre 1 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.

La modélisation des calculs des cellules brutes peut se révéler insatisfaisante dans cette situation car le *total des mois de service* varie selon le type de navire. Les variables qui mesurent le degré d'« exposition » au risque sont traitées dans le modèle linéaire généralisé en tant que variables de décalage. D'autre part, une régression de Poisson considère que le log de la variable dépendante est linéaire dans les variables indépendantes. Par conséquent, pour utiliser les modèles linéaires généralisés pour ajuster une régression de Poisson aux taux d'accidents, vous devez utiliser le *logarithme du total des mois de service*.

Ajustement d'une régression de Poisson « surdispersée »

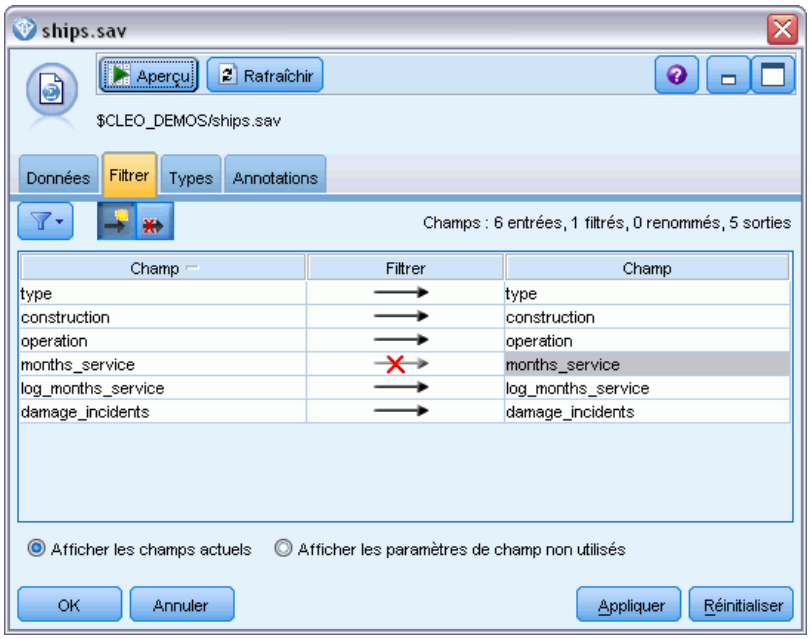
- Ajoutez un noeud source Statistics pointant vers *ships.sav* dans le dossier *Demos*.

Figure 23-1
Exemple de flux utilisé pour l'analyse des taux de dommage



- Dans l'onglet Filtrer du noeud source, excluez le champ *months_service*. Les valeurs transformées en log de cette variable sont contenues dans *log_months_service*, qui sera utilisé dans l'analyse.

Figure 23-2
Filtrage d'un champ inutile

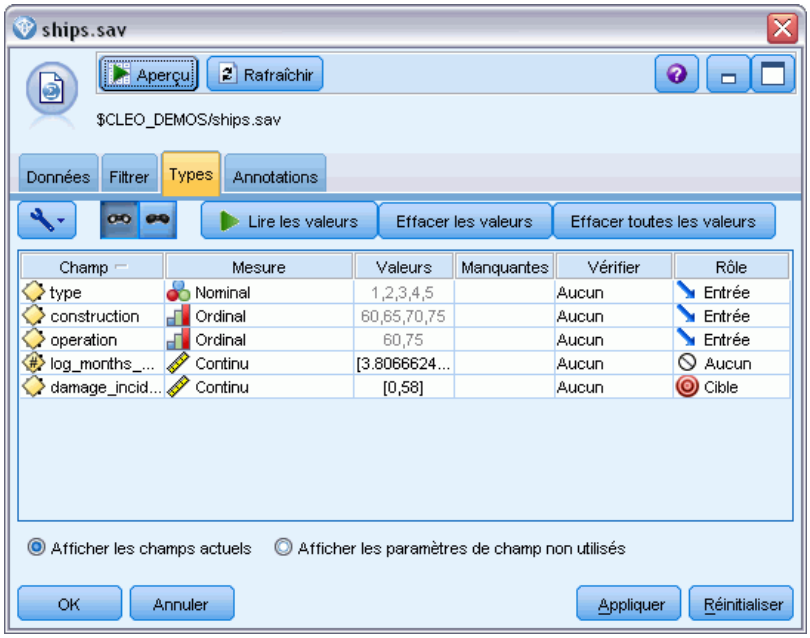


(Vous pouvez également régler le rôle de ce champ sur Aucun dans l'onglet Types au lieu de l'exclure ou sélectionner les champs que vous souhaitez utiliser dans le noeud de modélisation.)

- Dans l'onglet Types du noeud source, définissez le rôle du champ *damage_incidents* sur Cible. Le rôle de tous les autres champs doit être défini sur Entrée.

▶ Cliquez sur Lire les valeurs pour instancier les données.

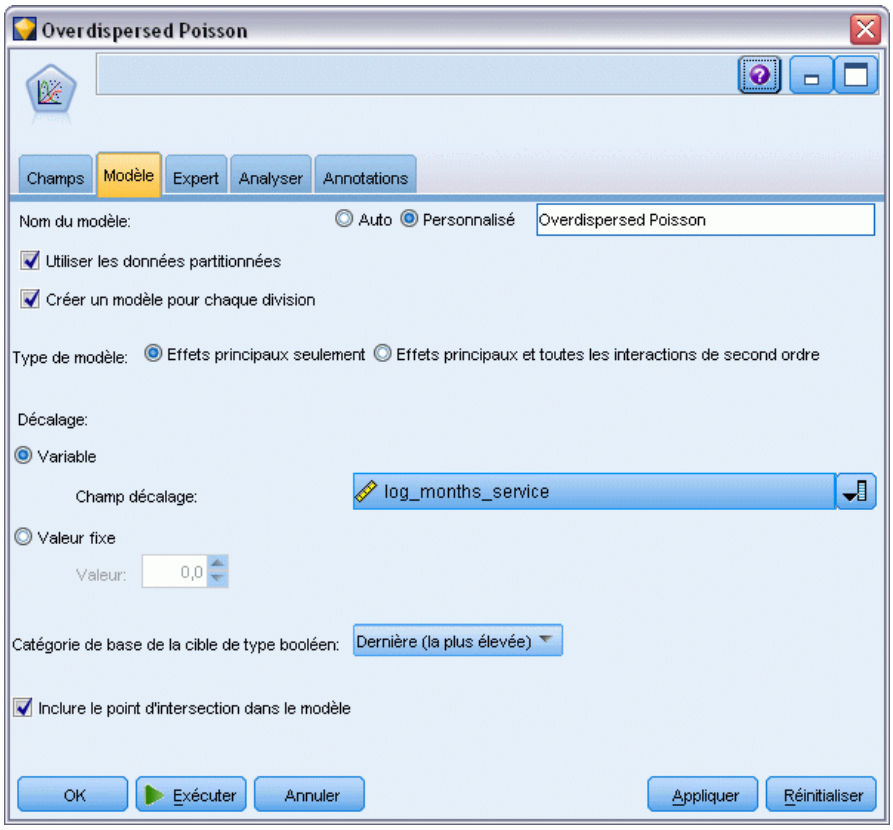
Figure 23-3
Définition du rôle de champ



▶ Reliez un noeud Modèles linéaires généralisés au noeud source. Dans le noeud Modèles linéaires généralisés, cliquez sur l'onglet Modèle.

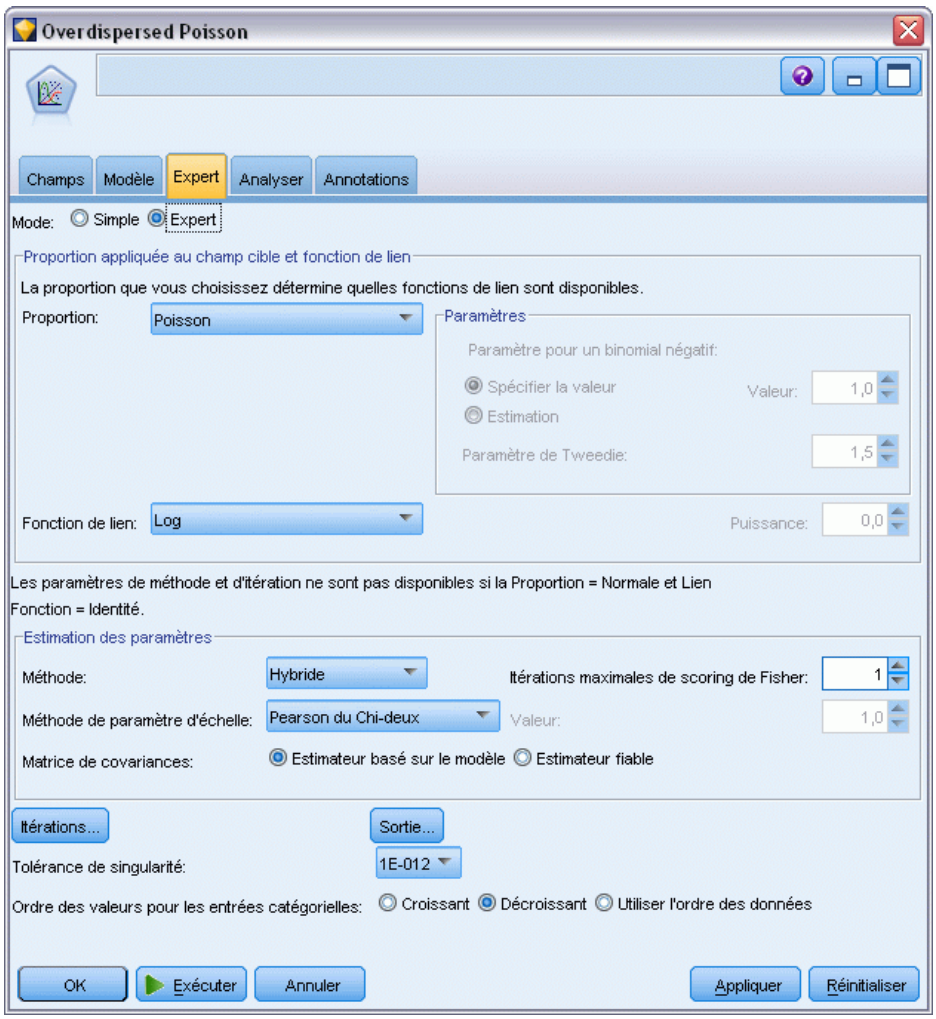
▶ Sélectionnez *log_months_service* comme variable de décalage.

Figure 23-4
Choix des options de modèle



- Cliquez sur l'onglet Expert et sélectionnez Expert pour activer les options de modélisation expert.

Figure 23-5
Choix des options expert



- Sélectionnez Poisson comme distribution pour la réponse et Log comme fonction de lien.
- Sélectionnez Pearson du Chi-deux comme méthode d'estimation du paramètre d'échelle. Le paramètre d'échelle est généralement considéré comme étant égal à 1 dans une régression de Poisson. Cependant, McCullagh et Nelder utilisent l'estimation Pearson du Chi-deux pour obtenir des estimations de variance et des niveaux de signification plus prudents.
- Sélectionnez l'ordre des catégories des facteurs Décroissant. Cela signifie que la première catégorie de chaque facteur constitue la catégorie de référence. L'effet de cette sélection sur le modèle porte sur l'interprétation des estimations de paramètre.
- Cliquez sur Exécuter pour créer le nugget de modèle qui est ajouté à l'espace de travail du flux et à la palette Modèles en haut à droite. Pour consulter les détails du modèle, cliquez avec le bouton droit de la souris sur le nugget et sélectionnez Modifier ou Parcourir, puis sur l'onglet Avancé.

Statistiques de qualité de l'ajustement

Figure 23-6
Qualités d'ajustement des statistiques

	Valeur	ddl	Valeur/ddl
Déviance	38,695	25	1,548
Déviance mise à l'échelle	22,883	25	
Khi-deux de Pearson	42,275	25	1,691
Khi-deux de Pearson mis à l'échelle	25,000	25	
Log-vraisemblance(b,c)	-68,281		
Log-vraisemblance ajustée(d)	-40,379		
Critère d'information d'Akaike (AIC)	154,562		
AIC corrigé d'échantillon fini (AICC)	162,062		
Critère d'information Bayésien (BIC)	168,299		
AIC cohérent (CAIC)	177,299		
Variable dépendante : Number of damage incidentsModèle : (Ordonnée à l'origine), type, construction, operation, décalage = log_months_service			
a. Les critères d'information sont de type "valeur faible préférée".			
b. La fonction de log-vraisemblance complète est affichée et utilisée dans le calcul des critères d'information.			
c. La fonction de log-vraisemblance est basée sur un paramètre d'échelle fixé à 1.			
d. La fonction de log-vraisemblance est basée sur un paramètre d'échelle estimé et est utilisée dans le test composite d'adaptation du modèle.			

Le tableau des qualités d'ajustement des statistiques contient des mesures permettant de comparer les modèles en concurrence. Par ailleurs, la *valeur/df* des statistiques de la déviance et du Pearson du Chi-deux donne des estimations pour le paramètre d'échelle. Ces valeurs doivent être proches de 1,0 pour une régression de Poisson. Des valeurs supérieures à 1,0 indiquent que l'ajustement du modèle surdispersé peut être judicieux.

Test composite

Figure 23-7
Test composite

Khi-deux du rapport de vraisemblance	ddl	Sig.
107,633	8	,000
Variable dépendante : Number of damage incidentsModèle : (Ordonnée à l'origine), type, construction, operation, décalage = log_months_service		
a. Compare le modèle ajusté au modèle avec constante seulement.		

Le test composite est un test Chi-deux de rapport de vraisemblance du modèle actuel par rapport au modèle nul (dans ce cas, la constante). Une valeur de signification inférieure à 0,05 indique que le modèle actuel permet d'obtenir de meilleurs résultats que le modèle nul.

Tests des effets de modèle

Figure 23-8
Tests des effets de modèle

Source	Type III		
	Khi-deux de Wald	ddl	Sig.
(Ordonnée à l'origine)	2138,657	1	,000
type	15,415	4	,004
construction	17,242	3	,001
operation	6,249	1	,012
Variable dépendante : Number of damage incidentsModèle : (Ordonnée à l'origine), type, construction, operation, décalage = log_months_service			

Chaque terme du modèle est testé afin de déterminer s'il présente un effet. Les termes dont les valeurs de signification sont inférieures à 0,05 ont un effet visible. Chacun des termes des effets principaux contribue au modèle.

Estimations des paramètres

Figure 23-9
Estimations des paramètres

Paramètre	B	Erreur standard	Intervalle de confiance de Wald à 95 %		Test d'hypothèse		
			Inférieur	Supérieur	Khi-deux de Wald	ddl	Sig.
(Ordonnée à l'origine)	-6,406	,2828	-6,960	-5,852	513,238	1	,000
[type=5]	,326	,3067	-,276	,927	1,127	1	,288
[type=4]	-,076	,3779	-,817	,665	,040	1	,841
[type=3]	-,687	,4279	-1,526	,151	2,581	1	,108
[type=2]	-,543	,2309	-,996	-,091	5,536	1	,019
[type=1]	0(a)	.	.	.	.	.	.
[construction=75]	,453	,3032	-,141	1,048	2,236	1	,135
[construction=70]	,818	,2208	,386	1,251	13,743	1	,000
[construction=65]	,697	,1946	,316	1,079	12,835	1	,000
[construction=60]	0(a)	.	.	.	.	.	.
[operation=75]	,384	,1538	,083	,686	6,249	1	,012
[operation=60]	0(a)	.	.	.	.	.	.
(Échelle)	1,691(b)						

Variable dépendante : Number of damage incidentsModèle : (Ordonnée à l'origine), type, construction, operation, décalage = log_months_service

a. Défini sur zéro car ce paramètre est redondant.

b. Calcul en fonction du Khi-deux de Pearson.

Le tableau des estimations de paramètres récapitule l'effet de chaque variable indépendante. Bien que l'interprétation des coefficients dans ce modèle soit difficile de par la nature de la fonction de lien, les signes des coefficients des covariables et les valeurs relatives des coefficients des niveaux de facteur peuvent fournir des informations importantes sur les effets des variables indépendantes dans le modèle.

- Pour les covariables, des coefficients positifs (négatifs) indiquent des relations positives (inverses) entre les variables indépendantes et les résultats. Une valeur croissante d'une covariable avec un coefficient positif correspond à un taux croissant d'incidents provoquant des dommages.
- Dans le cas des facteurs, un niveau de facteur présentant un coefficient supérieur indique un impact plus important des dommages. Le signe d'un coefficient d'un niveau de facteur dépend de l'effet de ce niveau de facteur par rapport à la modalité de référence.

Vous pouvez tirer les conclusions suivantes en fonction des estimations des paramètres :

- Le type de navire *B [type=2]* présente un taux de dommage considérablement inférieur du point de vue statistique (valeur *p* de 0,019) (coefficient estimé de –0,543) que le type *A [type=1]*, la catégorie de référence. Le type *C [type=3]* présente en réalité un paramètre estimé inférieur à celui de *B*, mais la variabilité de l'estimation de *C* trouble l'effet.

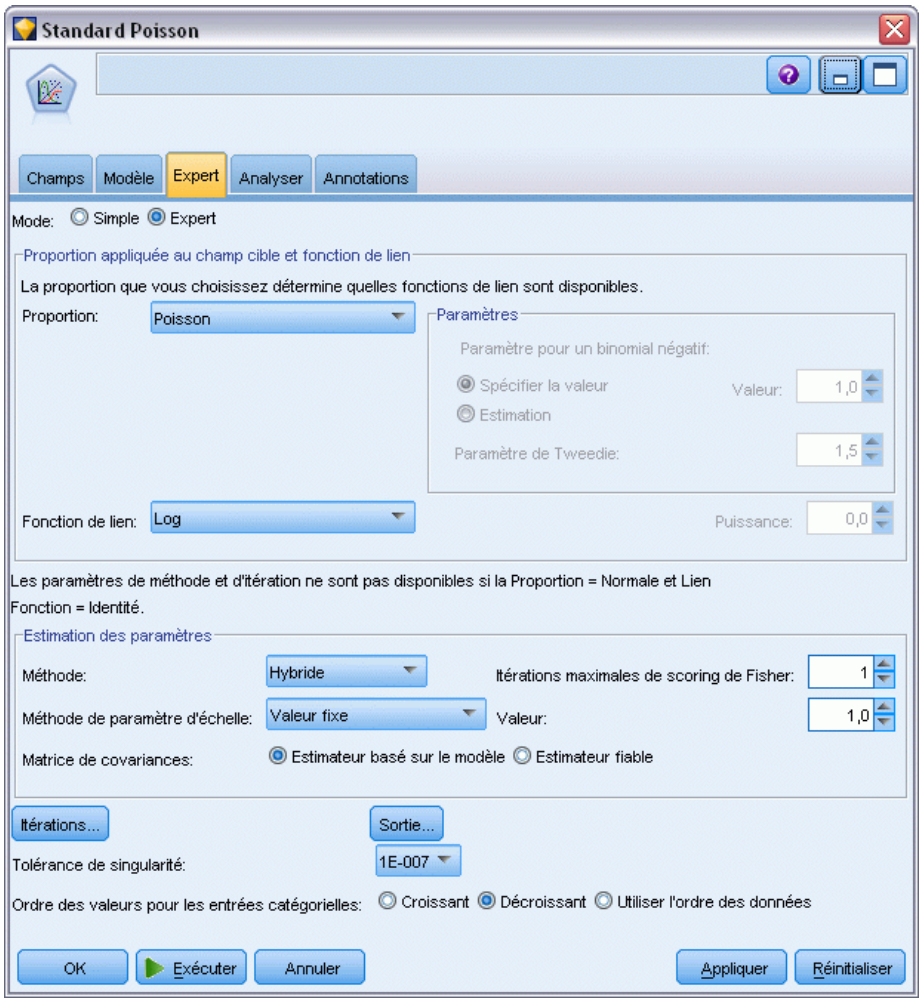
Reportez-vous aux moyennes marginales estimées pour toutes les relations entre les niveaux du facteur.

- Les navires construits entre 1965 et 1969 *[construction=65]* et, 1970 et 1974 *[construction=70]* présentent des taux de dommage considérablement supérieurs du point de vue statistique (valeurs p <0,001) (coefficients estimés de 0,697 et de 0,818, respectivement) que ceux construits entre 1960 et 1964 *[construction=60]*, la catégorie de référence. Reportez-vous aux moyennes marginales estimées pour toutes les relations entre les niveaux du facteur.
- Les navires en service entre 1975 et 1979 *[operation=75]* présentent des taux de dommage considérablement supérieurs du point de vue statistique (valeur p de 0,012) (coefficient estimé de 0,384) que ceux en service entre 1960 et 1974 *[operation=60]*.

Ajustement des modèles alternatifs

Le problème concernant la régression de Poisson « surdispersée » est qu'il n'existe pas de manière formelle de la tester par rapport à la régression de Poisson « standard ». Toutefois, un test formel conseillé afin de déterminer l'existence d'une surdispersion consiste à effectuer un test de rapport de vraisemblance entre une régression de Poisson « standard » et une régression binomiale négative, l'ensemble des autres paramètres étant égaux. En cas d'absence de surdispersion dans la régression de Poisson, la statistique −2×(log de vraisemblance du modèle de Poisson − log de vraisemblance du modèle binomial négatif) doit présenter une proportion de mélange : la moitié de sa masse de probabilité sur 0 et le reste dans une distribution Chi-deux avec 1 degré de liberté.

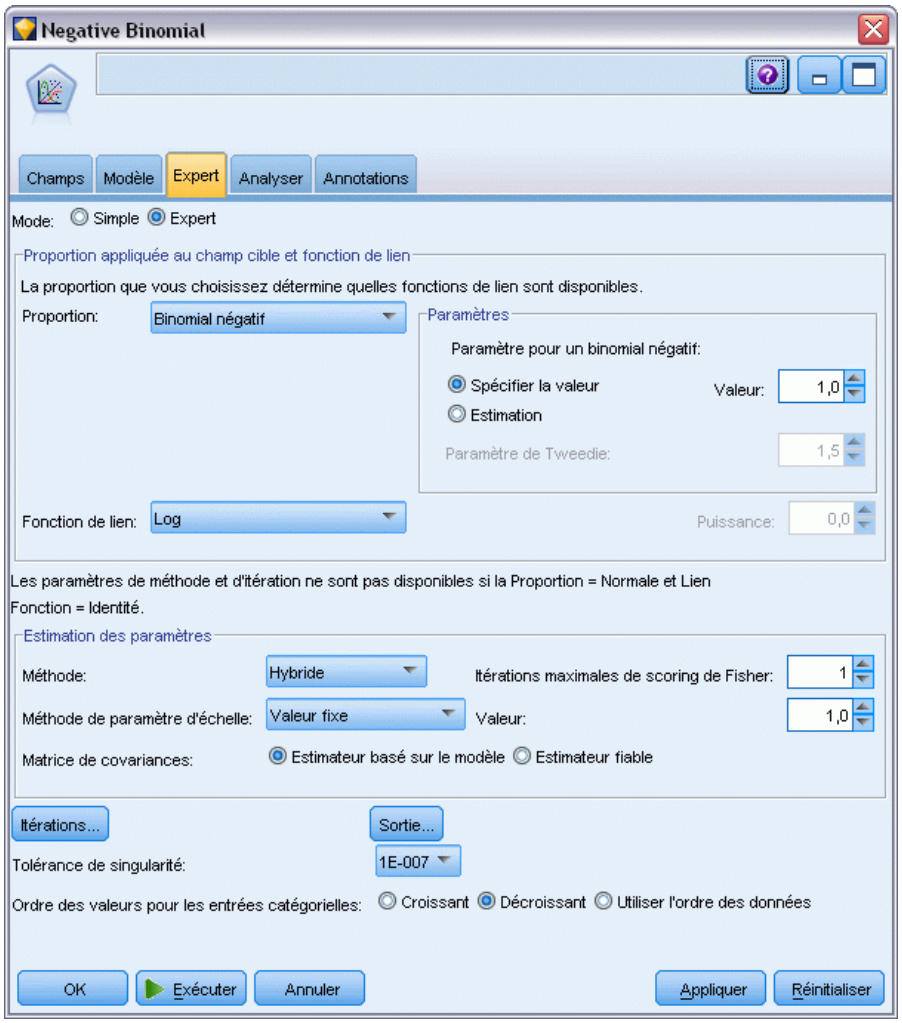
Figure 23-10
Onglet Expert



Pour ajuster la régression de Poisson “standard”, copiez et collez le noeud Modèles linéaires généralisés, liez-le au noeud source, ouvrez le nouveau noeud, puis cliquez sur l’onglet Expert.

- Sélectionnez Valeur fixe comme méthode d’estimation du paramètre d’échelle. Par défaut, cette valeur est 1.

Figure 23-11
Onglet Expert



- Pour ajuster la régression binomiale négative, copiez et collez le noeud Modèles linéaires généralisés, liez-le au noeud source, ouvrez le nouveau noeud, puis cliquez sur l'onglet Expert.
- Sélectionnez la loi binomiale négative. Conservez la valeur par défaut de 1 pour le paramètre secondaire.
- Exécutez le flux et accédez à l'onglet Avancé dans les nuggets de modèle nouvellement créés.

Statistiques de qualité de l'ajustement

Figure 23-12
Qualités d'ajustement des statistiques pour la régression de Poisson standard

	Valeur	ddl	Valeur/ddl
Déviance	38,695	25	1,548
Déviance mise à l'échelle	22,883	25	
Khi-deux de Pearson	42,275	25	1,691
Khi-deux de Pearson mis à l'échelle	25,000	25	
Log-vraisemblance(b,c)	-68,281		
Log-vraisemblance ajustée(d)	-40,379		
Critère d'information d'Akaike (AIC)	154,562		
AIC corrigé d'échantillon fini (AICC)	162,062		
Critère d'information Bayésien (BIC)	168,299		
AIC cohérent (CAIC)	177,299		
Variable dépendante : Number of damage incidentsModèle : (Ordonnée à l'origine), type, construction, operation, décalage = log_months_service			
a. Les critères d'information sont de type "valeur faible préférée".			
b. La fonction de log-vraisemblance complète est affichée et utilisée dans le calcul des critères d'information.			

Le log de vraisemblance indiqué pour la régression de Poisson standard est –68,281. Comparez ce chiffre à celui du modèle binomial négatif.

Figure 23-13
Qualités d'ajustement des statistiques pour la régression binomiale négative

	Valeur	ddl	Valeur/ddl
Déviance	11,145	25	,446
Déviance mise à l'échelle	11,145	25	
Khi-deux de Pearson	8,815	25	,353
Khi-deux de Pearson mis à l'échelle	8,815	25	
Log-vraisemblance(b)	-83,725		
Critère d'information d'Akaike (AIC)	185,450		
AIC corrigé d'échantillon fini (AICC)	192,950		
Critère d'information Bayésien (BIC)	199,187		
AIC cohérent (CAIC)	208,187		
Variable dépendante : Number of damage incidentsModèle : (Ordonnée à l'origine), type, construction, operation, décalage = log_months_service			
a. Les critères d'information sont de type "valeur faible préférée".			
b. La fonction de log-vraisemblance complète est affichée et utilisée dans le calcul des critères d'information.			

Le log de vraisemblance indiqué pour la régression binomiale négative est –83,725. Il est en réalité *inférieur* au log de vraisemblance de la régression de Poisson, ce qui indique (sans le test de rapport de vraisemblance) que cette régression binomiale négative n'offre pas d'amélioration par rapport à la régression de Poisson.

Toutefois, la valeur de 1 choisie pour le paramètre secondaire de la loi binomiale négative peut ne pas être optimale pour cet ensemble de données. Une autre manière de tester la surdispersion consiste à ajuster un modèle binomial négatif avec un paramètre secondaire égal à 0 et à demander le test du multiplicateur de Lagrange dans la boîte de dialogue Sortie de l'onglet Expert. Si le test n'est pas concluant, la surdispersion ne doit pas poser de problème à cet ensemble de données.

Récapitulatif

A l'aide des modèles linéaires généralisés, vous avez ajusté trois modèles différents pour les données d'effectif. Il s'est avéré que la régression binomiale négative n'offre aucune amélioration par rapport à la régression de Poisson. La régression de Poisson surdispersée semble offrir une alternative raisonnable au modèle de Poisson standard, mais il n'existe pas de test formel permettant de choisir l'un plutôt que l'autre.

Vous trouverez des explications sur le fondement mathématique des méthodes de modélisation utilisées dans IBM® SPSS® Modeler dans le *Guide des Algorithmes SPSS Modeler*.

Chapitre
24

Ajustement d'une régression gamma à des déclarations de sinistre automobile (modèles linéaires généralisés)

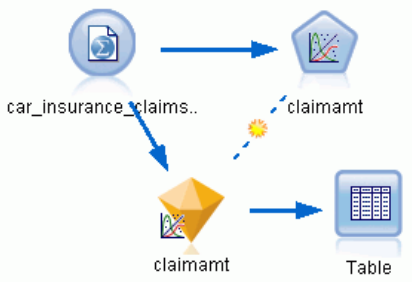
Un modèle linéaire généralisé permet d'ajuster une régression gamma pour l'analyse de données d'intervalle positif. Par exemple, un ensemble de données présenté et analysé ailleurs () porte sur les déclarations de sinistre automobile. La somme moyenne des déclarations peut être modélisée comme suivant une distribution gamma, en utilisant une fonction de lien inverse pour associer la moyenne de la variable dépendante à une combinaison linéaire de variables indépendantes. Afin de représenter le nombre variable de déclarations servant à calculer les sommes moyennes des déclarations, vous pouvez indiquer la pondération de mise à l'échelle *Number of claims*.

Cet exemple utilise le flux *car-insurance_genlin.str*, qui fait référence au fichier de données *car_insurance_claims.sav*. Le fichier de données se trouve dans le dossier *Demos* et le fichier de flux dans le sous-dossier *streams*. Pour plus d'informations, reportez-vous à la section Dossier Demos dans le chapitre 1 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.

Création du flux

- Ajoutez un noeud Statistics pointant vers *car_insurance_claims.sav* dans le dossier *Demos*.

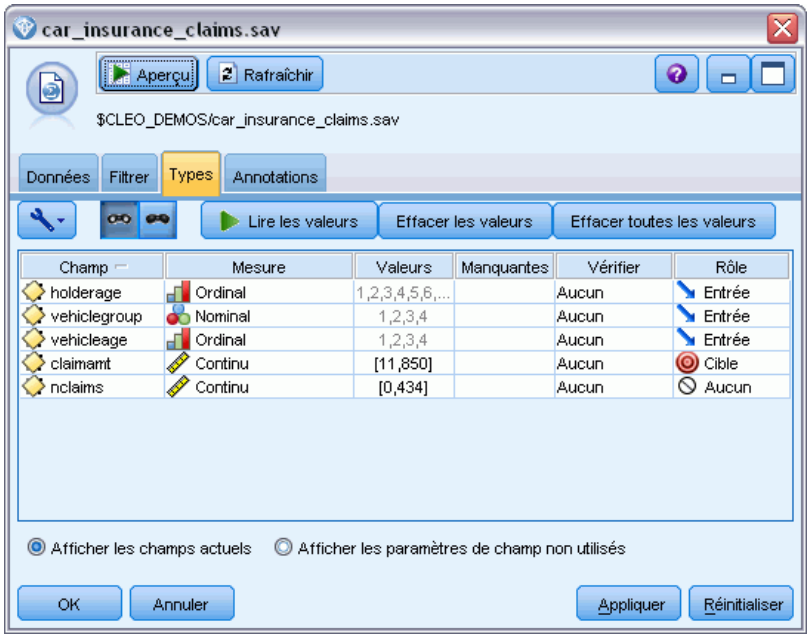
Figure 24-1
Exemple de flux relatif à la prévision des déclarations de sinistre automobile



- Dans l'onglet Types du noeud source, définissez le rôle du champ *claimamt* sur Cible. Le rôle de tous les autres champs doit être défini sur Entrée.

▶ Cliquez sur Lire les valeurs pour instancier les données.

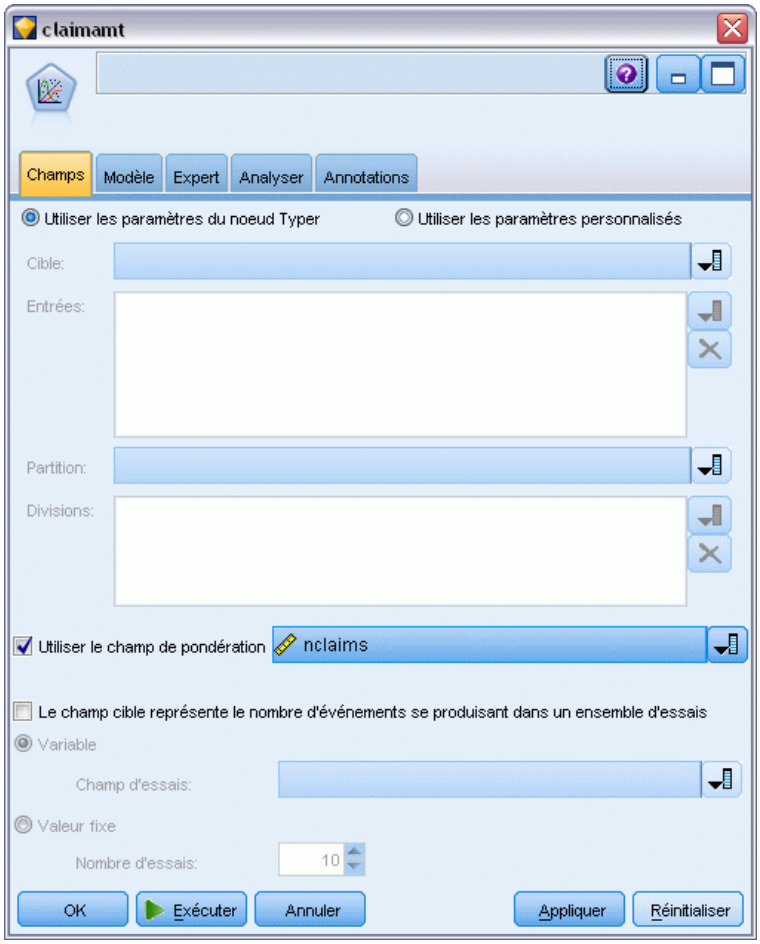
Figure 24-2
Définition du rôle de champ



▶ Reliez un noeud Modèles linéaires généralisés au noeud source. Dans le noeud Modèles linéaires généralisés, cliquez sur l'onglet Champs.

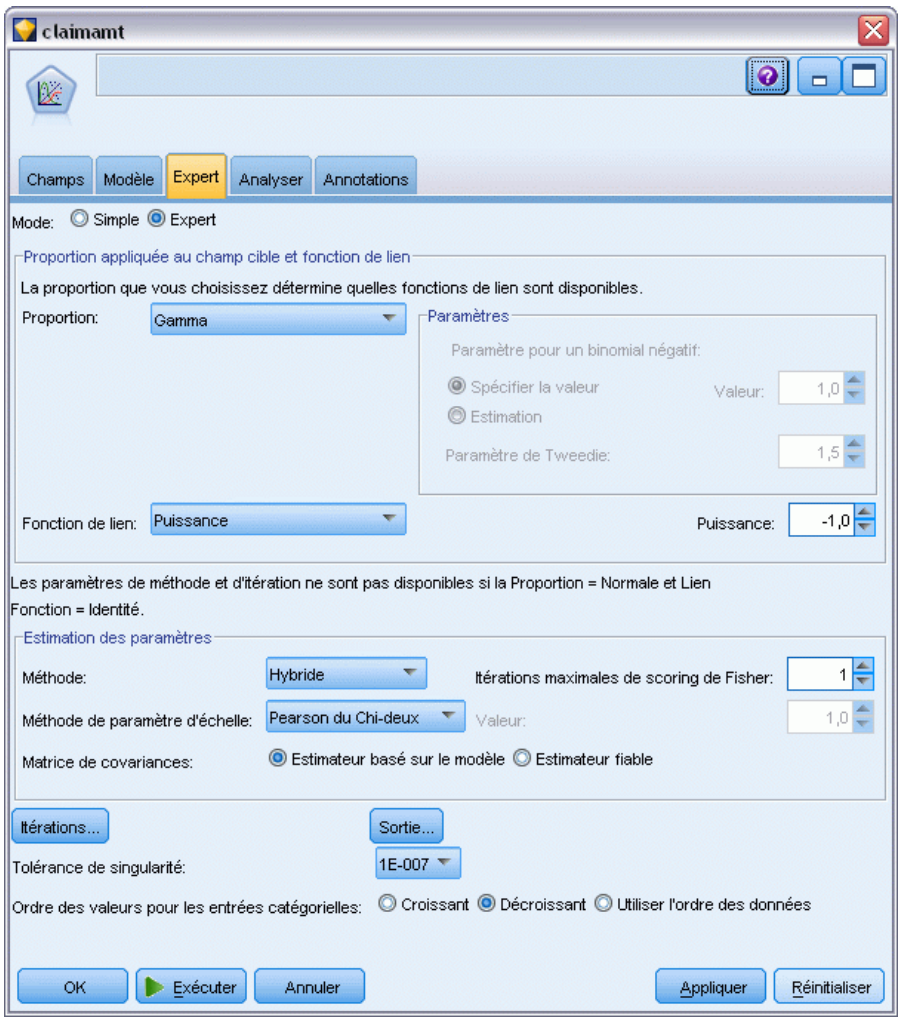
▶ Sélectionnez le champ de pondération *nclaims*.

Figure 24-3
Sélection des options de champ



▶ Cliquez sur l'onglet Expert et sélectionnez Expert pour activer les options de modélisation expert.

Figure 24-4
Choix des options expert



▶ Sélectionnez la distribution de réponse Gamma.

▶ Sélectionnez la fonction de lien Puissance, puis saisissez l'exposant de la fonction de puissance -1,0. Il s'agit d'un lien inverse.

▶ Sélectionnez la méthode d'estimation du paramètre d'échelle Pearson du Chi-deux. Il s'agit de la méthode utilisée par McCullagh et Nelder. Nous allons donc la suivre de manière à reproduire leurs résultats.

▶ Sélectionnez l'ordre des catégories des facteurs Décroissant. Cela signifie que la première catégorie de chaque facteur constitue la catégorie de référence. L'effet de cette sélection sur le modèle porte sur l'interprétation des estimations de paramètre.

▶ Cliquez sur Exécuter pour créer le nugget de modèle qui est ajouté à l'espace de travail du flux et à la palette Modèles en haut à droite. Pour afficher les détails du modèle, cliquez avec le bouton

droit de la souris sur le nugget de modèle et choisissez Modifier ou Parcourir, puis sélectionnez l'onglet Avancé.

Estimations des paramètres

Figure 24-5
Estimations des paramètres

Paramètre	B	Erreur standard	Intervalle de confiance de Wald à 95 %		Test d'hypothèse		
			Inférieur	Supérieur	Khi-deux de Wald	ddl	Sig.
(Ordonnée à l'origine)	,003	,0004	,003	,004	66,593	1	,000
[holderage=8]	,001	,0004	,000	,002	4,898	1	,027
[holderage=7]	,001	,0004	,000	,002	5,046	1	,025
[holderage=6]	,001	,0004	,000	,002	5,740	1	,017
[holderage=5]	,001	,0004	,001	,002	10,682	1	,001
[holderage=4]	,000	,0004	,000	,001	1,268	1	,260
[holderage=3]	,000	,0004	,000	,001	,720	1	,396
[holderage=2]	,000	,0004	-,001	,001	,054	1	,816
[holderage=1]	0(a)	.	.	.	.	.	.
[vehiclegroup=4]	-,001	,0002	-,002	-,001	61,883	1	,000
[vehiclegroup=3]	-,001	,0002	-,001	,000	13,039	1	,000
[vehiclegroup=2]	3,77E-005	,0002	,000	,000	,050	1	,823
[vehiclegroup=1]	0(a)	.	.	.	.	.	.
[vehicleage=4]	,004	,0004	,003	,005	88,175	1	,000
[vehicleage=3]	,002	,0002	,001	,002	53,013	1	,000
[vehicleage=2]	,000	,0001	,000	,001	13,191	1	,000
[vehicleage=1]	0(a)	.	.	.	.	.	.
(Échelle)	1,209(b)						

Variable dépendante : Average cost of claimsModèle : (Ordonnée à l'origine), holderage, vehiclegroup, vehicleage

a. Défini sur zéro car ce paramètre est redondant.

b. Calcul en fonction du Khi-deux de Pearson.

Le test composite et les tests d'effets de modèle (non affichés) indiquent que le modèle permet d'obtenir de meilleurs résultats que le modèle nul et que chaque caractéristique effet principal contribue à ce modèle. Le tableau des estimations de paramètre contient les mêmes valeurs que celles obtenues par McCullagh et Nelder pour les niveaux de facteur et le paramètre d'échelle.

Récapitulatif

Grâce à Modèles linéaires généralisés, vous venez d'ajuster une régression gamma aux données concernant les déclarations. Même si une fonction de lien canonique pour la distribution gamma a été utilisée dans ce modèle, un lien log produit également des résultats raisonnables. En général, il est difficile, voire impossible, de comparer directement des modèles avec différentes fonctions de lien. Toutefois, le lien log constitue un cas particulier du lien de puissance, où l'exposant est égal à 0. Par conséquent, vous pouvez comparer les déviances d'un modèle avec un lien log et un modèle avec un lien de puissance pour déterminer celui qui offre le meilleur ajustement (reportez-vous, par exemple, à la section 11.3 concernant McCullagh et Nelder).

Vous trouverez des explications sur le fondement mathématique des méthodes de modélisation utilisées dans IBM® SPSS® Modeler dans le *Guide des Algorithmes SPSS Modeler*.

Chapitre
25

Classification des échantillons de cellules (SVM)

Support Vector Machine (SVM) est une technique de classification et de régression particulièrement adaptée aux larges ensembles de données. Un large ensemble de données est un ensemble contenant un nombre important de variables indépendantes, comme c'est le cas dans le domaine de la bio-informatique (l'application des technologies de l'information aux données biochimiques et biologiques).

Un chercheur en médecine a obtenu un ensemble de données contenant les caractéristiques d'un certain nombre d'échantillons de cellules humaines supposées favoriser le développement du cancer. L'analyse des données originales indiquait que de nombreuses caractéristiques différaient considérablement entre les échantillons bénins et malins. Ce chercheur en médecine souhaite développer un modèle SVM qui peut utiliser les valeurs des caractéristiques de ces cellules dans des échantillons d'autres patients pour savoir au plus tôt si leurs échantillons peuvent être bénins ou malins.

Cet exemple utilise le flux nommé *svm_cancer.str*, disponible dans le dossier *Demos* du sous-dossier des *flux*. Le fichier de données est *cell_samples.data*. Pour plus d'informations, reportez-vous à la section Dossier Demos dans le chapitre 1 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.

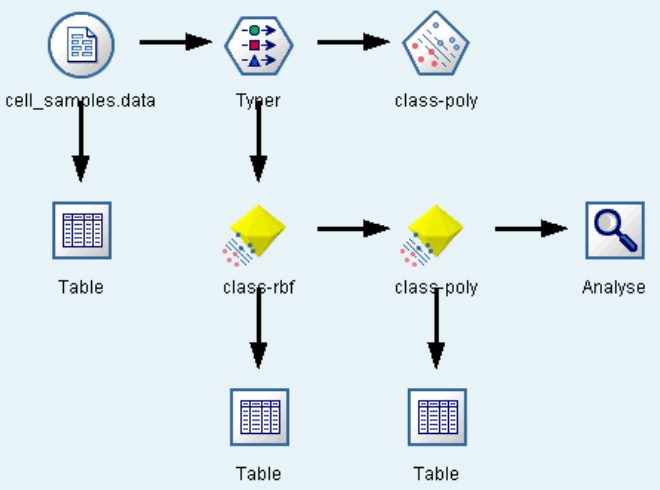
Cet exemple utilise un ensemble de données disponible au public dans le référentiel d'apprentissage automatique UCI(Asuncion et Newman, 2007). Cet ensemble de données est constitué de plusieurs centaines d'enregistrements d'échantillons de cellules humaines, chacun d'entre eux contenant les valeurs d'un ensemble de caractéristiques des cellules. Les champs de chaque enregistrement sont :

Nom de champ	Description
ID	Identifiant du patient
Clump	Epaisseur de l'agglutination
UnifSize	Uniformité de la taille des cellules
UnifShape	Uniformité de la forme des cellules
MargAdh	Adhésion marginale
SingEpiSize	Taille des cellules épithéliales
BareNuc	Noyau nu
BlandChrom	Chromatine terne
NormNucl	Nucléole normal
Mit	Mitoses
Class	Bénine ou maligne

Dans cet exemple, nous utilisons un ensemble de données contenant un nombre relativement petit de variables indépendantes dans chaque enregistrement.

Création du flux

Figure 25-1
Exemple de flux présentant la modélisation SVM



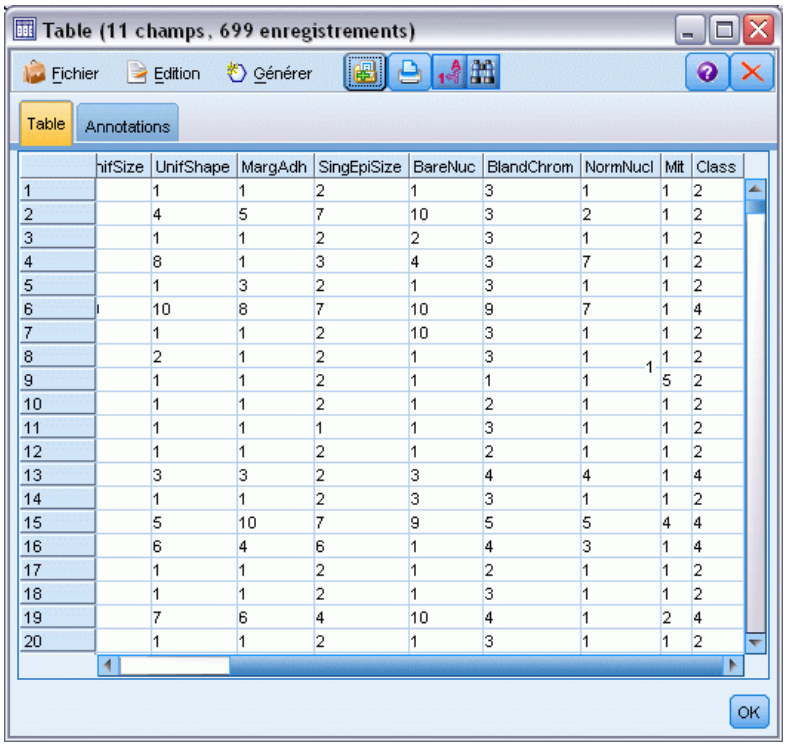
- Créez un nouveau flux et ajoutez un noeud source Délimité pointant vers *cell_samples.data* dans le dossier *Demos* de votre installation IBM® SPSS® Modeler.

  Examinons les données du fichier source.

- Ajoutez un noeud Table au flux.

- Liez le noeud Table au noeud Délimité et exécutez le flux.

Figure 25-2
Données source de SVM



	hifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1		1	1	2	1	3	1	1	2
2		4	5	7	10	3	2	1	2
3		1	1	2	2	3	1	1	2
4		8	1	3	4	3	7	1	2
5		1	3	2	1	3	1	1	2
6		10	8	7	10	9	7	1	4
7		1	1	2	10	3	1	1	2
8		2	1	2	1	3	1	1	2
9		1	1	2	1	1	1	5	2
10		1	1	2	1	2	1	1	2
11		1	1	1	1	3	1	1	2
12		1	1	2	1	2	1	1	2
13		3	3	2	3	4	4	1	4
14		1	1	2	3	3	1	1	2
15		5	10	7	9	5	5	4	4
16		6	4	6	1	4	3	1	4
17		1	1	2	1	2	1	1	2
18		1	1	2	1	3	1	1	2
19		7	6	4	10	4	1	2	4
20		1	1	2	1	3	1	1	2

Le champ *ID* contient les identifiants du patient. Les caractéristiques des échantillons de cellules de chaque patient se trouvent dans les champs *Clump* à *Mit*. Les valeurs vont de 1 à 10, 1 étant le plus proche de bénin.

Le champ *Class* contient le diagnostique, confirmé par plusieurs procédures médicales, établissant si les échantillons sont bénins (valeur = 2) ou malins (valeur = 4).

Figure 25-3
Paramètres du noeud Typer

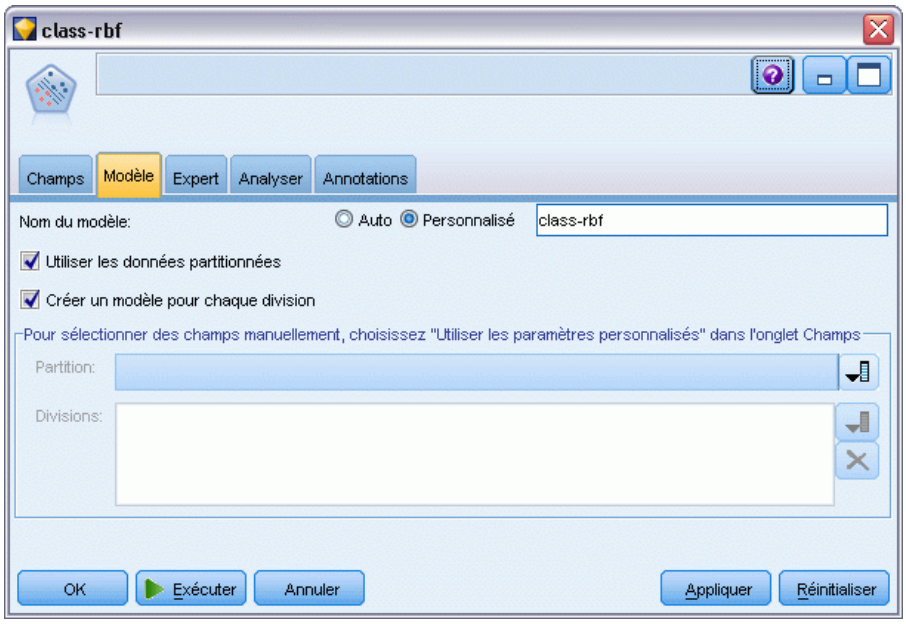
- Ajoutez un noeud Typer et liez-le au noeud Délimité.
- Ouvrez le noeud Typer.

  Nous voudrions que le modèle prédise la valeur de *Class* (c'est-à-dire, bénine (=2) ou maligne (=4)). Ce champ ne pouvant avoir qu'une des deux valeurs possibles, il est nécessaire de modifier son niveau de mesure pour refléter ceci.

- Dans la colonne Mesure du champ *Class* (le dernier de la liste), cliquez sur la valeur Continu et modifiez-la en Booléen.
- Cliquez sur Lire les valeurs.
- Dans la colonne Rôle, définissez le rôle du champ *ID* (l'identifiant du patient) sur Aucun, cette valeur n'étant pas utilisée comme variable indépendante ou comme cible du modèle.
- Définissez le rôle de la cible, *Class*, sur Cible et laissez le rôle de tous les autres champs (variables indépendantes) sur Entrée.
- Cliquez sur OK.

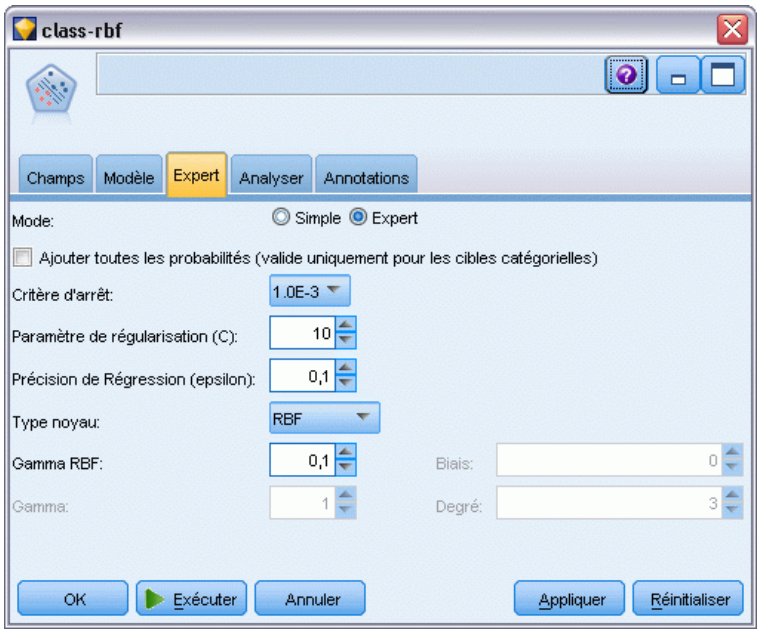
  Le noeud SVM propose plusieurs fonctions du noyau permettant son exécution. Comme il n'est pas évident de savoir quelle fonction est la plus appropriée à un ensemble de données spécifique, nous choisirons plusieurs fonctions afin de comparer leurs résultats. Commençons par la fonction par défaut, RBF (Fonction radiale de base).

Figure 25-4
Paramètres de l'onglet Modèle



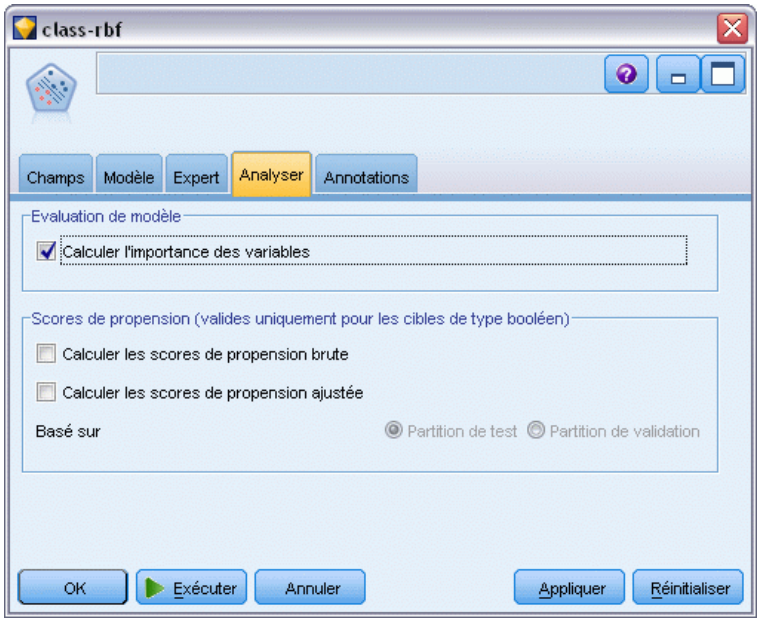
- Dans la palette Modélisation, reliez un noeud SVM au noeud Typer.
- Ouvrez le noeud SVM. Dans l'onglet Modèle, cliquez sur l'option Personnalisé pour le nom du modèle et entrez *class-rbf* dans le champ de texte adjacent.

Figure 25-5
Onglet Expert - Paramètres par défaut



▶ Dans l'onglet Expert, définissez le Mode sur Expert pour faciliter la lecture mais ne modifiez aucune des options par défaut. Veuillez noter que type noyau est défini sur RBF par défaut. Toutes les options sont grisées en mode Simple.

Figure 25-6
Paramètres de l'onglet Analyser

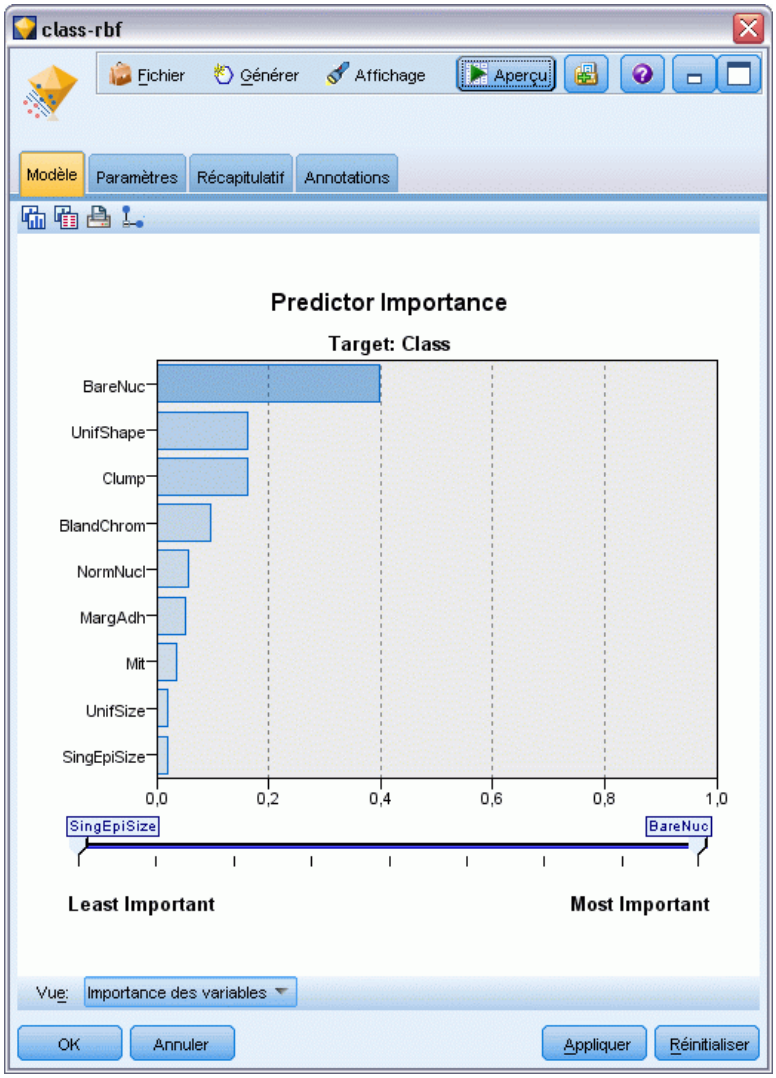


▶ Dans l'onglet Analyser, sélectionnez la case Calculer l'importance de la variable.

- Cliquez sur Exécuter. Le nugget de modèle est placé dans le flux et dans la palette Modèles en haut à droite de l'écran.
- Double-cliquez sur le nugget de modèle dans le flux.

Examen des données

Figure 25-7
Graphique de l'importance des variables indépendantes

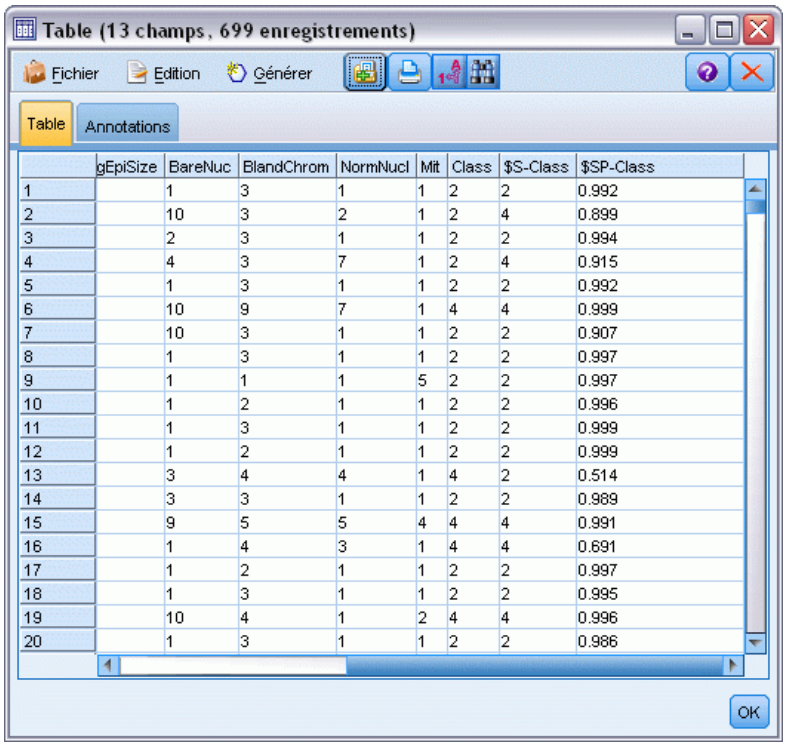


Dans l'onglet Modèle, le graphique d'importance des variables indépendantes présente l'effet relatif des différents champs sur la prévision. Ceci nous indique que *BareNuc* a bien l'effet le plus important alors que *UnifShape* et *Clump* sont également relativement importants.

- Cliquez sur OK.

- Liez un noeud Table au nugget de modèle *class-rbf*.
- Ouvrez le noeud Table, puis cliquez sur Exécuter.

Figure 25-8
Champs ajoutés pour la valeur de prévision et de confiance



	gEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class	$S-Class	$SP-Class
1		1	3	1	1	2	2	0.992
2		10	3	2	1	2	4	0.899
3		2	3	1	1	2	2	0.994
4		4	3	7	1	2	4	0.915
5		1	3	1	1	2	2	0.992
6		10	9	7	1	4	4	0.999
7		10	3	1	1	2	2	0.907
8		1	3	1	1	2	2	0.997
9		1	1	1	5	2	2	0.997
10		1	2	1	1	2	2	0.996
11		1	3	1	1	2	2	0.999
12		1	2	1	1	2	2	0.999
13		3	4	4	1	4	2	0.514
14		3	3	1	1	2	2	0.989
15		9	5	5	4	4	4	0.991
16		1	4	3	1	4	4	0.691
17		1	2	1	1	2	2	0.997
18		1	3	1	1	2	2	0.995
19		10	4	1	2	4	4	0.996
20		1	3	1	1	2	2	0.986

- Le modèle a créé deux champs supplémentaires. Faites défiler les sorties de la table vers la droite pour les voir :

Nouveau nom de champ	Description
$S-Class	Valeur de la *classe* prédite par le modèle.
$SP-Class	Score de propension de cette prévision (la probabilité qu'a cette prévision d'être vraie, une valeur de 0,0 à 1,0).

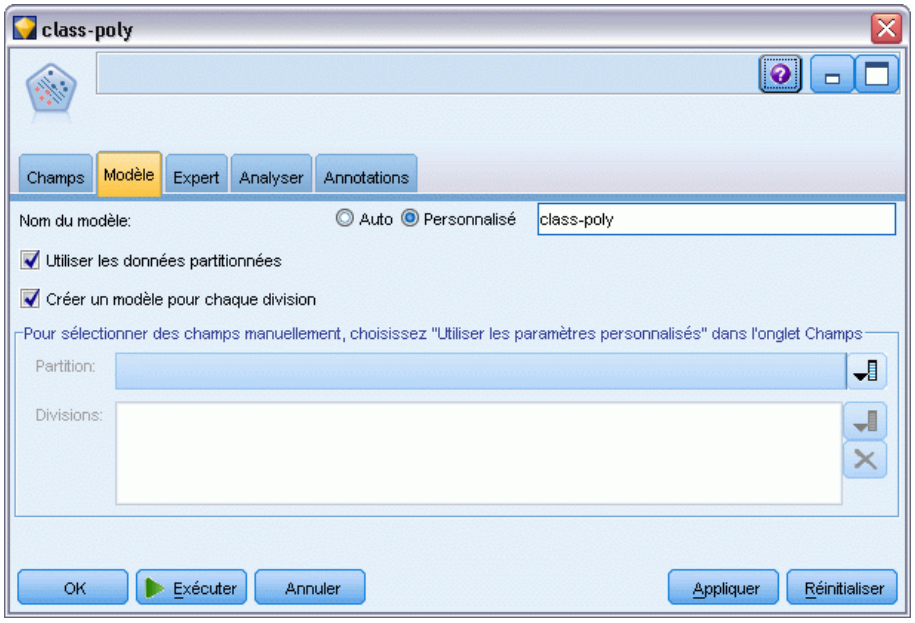
D'un coup d'oeil à la table, nous pouvons voir que les scores de propension (dans la colonne *$SP-Class*) de la majorité des enregistrements sont assez élevés.

Mais il existe des exceptions importantes ; par exemple, l'enregistrement pour le patient 1041801 à la ligne 13 dont la valeur de 0,514 est anormalement basse. De plus, si l'on compare *Class* à *$S-Class*, il est évident que ce modèle a effectué plusieurs prévisions incorrectes, même lorsque le score de propension était relativement élevé (par exemple, lignes 2 et 4).

Voyons si le résultat peut être meilleur en choisissant un autre type de fonction.

Essai d'une autre fonction

Figure 25-9
Définition d'un nouveau nom pour le modèle



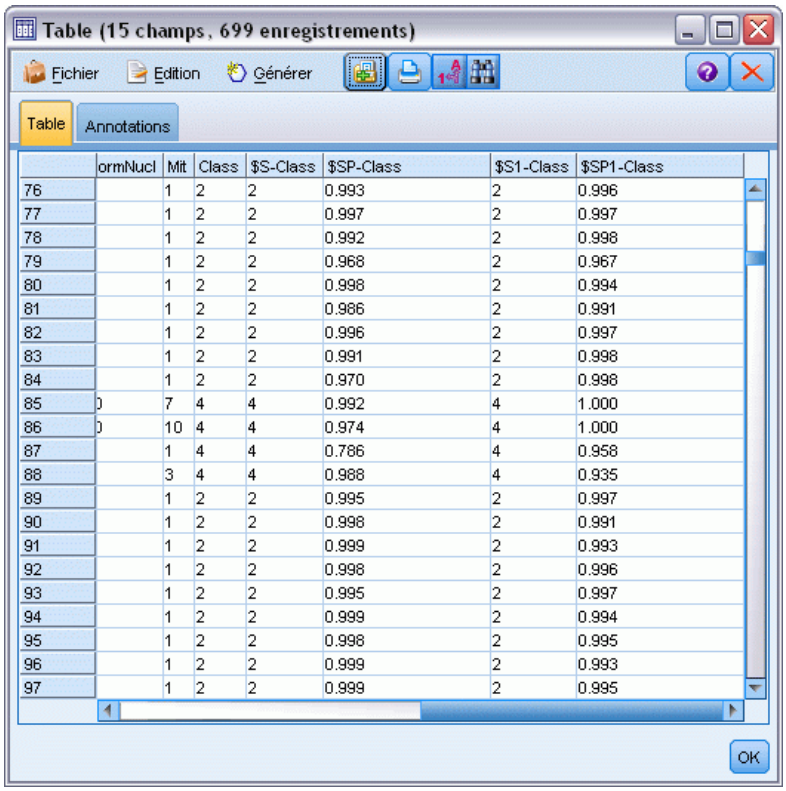
- Fermez la fenêtre de sortie Table.
- Reliez un deuxième noeud de modélisation SVM au noeud Typer.
- Ouvrez le nouveau noeud SVM.
- Dans l'onglet Modèle, choisissez Personnalisé et saisissez *class-poly* comme nom de modèle.

Figure 25-10
Paramètres de l'onglet Expert pour polynomial

- Dans l'onglet Expert, définissez le Mode sur Expert.
- Définissez le type du noyau sur Polynomial et cliquez sur Exécuter. Le nugget de modèle *class-poly* est ajouté au flux et dans la palette Modèles en haut à droite de l'écran.
- Connectez le nugget de modèle *class-rbf* au nugget de modèle *class-poly* (choisissez Remplacer dans la boîte de dialogue d'avertissement).
- Liez un noeud Table au nugget *class-poly*.
- Ouvrez le noeud Table, puis cliquez sur Exécuter.

Comparaison des résultats

Figure 25-11
Champs ajoutés pour la fonction Polynomiale



	ormNucl	Mit	Class	$S-Class	$SP-Class	$S1-Class	$SP1-Class
76		1	2	2	0.993	2	0.996
77		1	2	2	0.997	2	0.997
78		1	2	2	0.992	2	0.998
79		1	2	2	0.968	2	0.967
80		1	2	2	0.998	2	0.994
81		1	2	2	0.986	2	0.991
82		1	2	2	0.996	2	0.997
83		1	2	2	0.991	2	0.998
84		1	2	2	0.970	2	0.998
85	0	7	4	4	0.992	4	1.000
86	0	10	4	4	0.974	4	1.000
87		1	4	4	0.786	4	0.958
88		3	4	4	0.988	4	0.935
89		1	2	2	0.995	2	0.997
90		1	2	2	0.998	2	0.991
91		1	2	2	0.999	2	0.993
92		1	2	2	0.998	2	0.996
93		1	2	2	0.995	2	0.997
94		1	2	2	0.999	2	0.994
95		1	2	2	0.998	2	0.995
96		1	2	2	0.999	2	0.993
97		1	2	2	0.999	2	0.995

- Faites défiler les sorties de la table vers la droite pour voir les champs ajoutés.

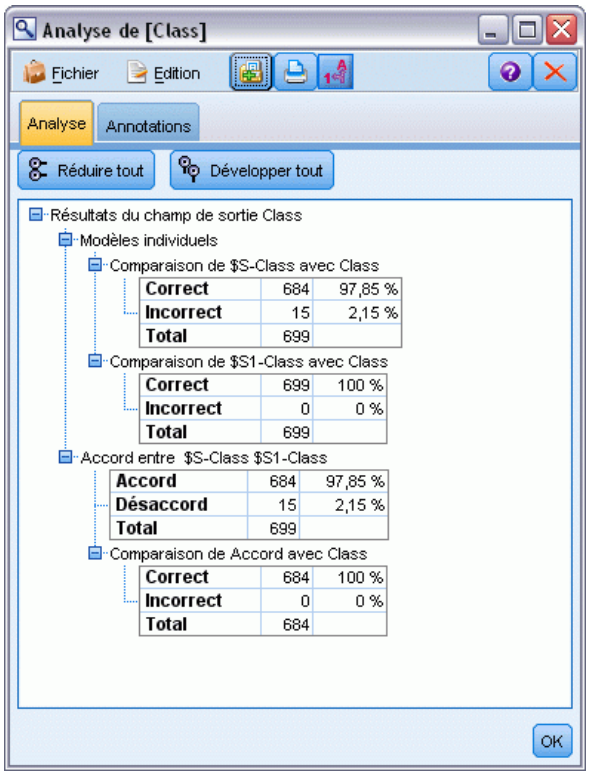
  Les champs générés pour le type de fonction Polynomiale sont appelés *$S1-Class* et *$SP1-Class*.

  Les résultats pour Polynomiale semblent bien meilleurs. La majorité des scores de propension sont de 0,995 ou plus ce qui est très encourageant.

- Pour confirmer l'amélioration dans le modèle, liez un noeud Analyse au nugget de modèle *class-poly*.

  Ouvrez le noeud Analyse, puis cliquez sur Exécuter.

Figure 25-12
nœud Analyse



Cette technique qui emploie le noeud Analyse vous permet de comparer au moins deux nuggets de modèle de même type. La sortie du noeud Analyse indique que la fonction RBF prédit correctement 97,85% des cas, ce qui est relativement bon. Mais cette sortie indique que la fonction Polynomiale a correctement prédit le diagnostique pour chacun des cas. En pratique, il est peu probable que vous soyez confronté à une précision de 100 %. Vous pouvez néanmoins utiliser le noeud Analyse pour déterminer si le modèle a une précision acceptable pour votre application.

En fait, aucun des autres types de fonction (Sigmoïdale ou Linéaire) n'a donné de résultats aussi bons que ceux de la fonction Polynomiale sur cet ensemble de données précis. Cependant, avec un autre ensemble de données, les résultats pourraient facilement être différents. Par conséquent, il est utile d'essayer toutes les options disponibles.

Récapitulatif

Vous avez utilisé différents types de fonctions du noyau SVM afin de prédire une classification à partir de plusieurs attributs. Vous avez vu la façon dont différents noyaux donnent différents résultats pour le même ensemble de données et comment mesurer l'amélioration d'un modèle par rapport à un autre.

Chapitre
26

Utilisation de la régression de Cox pour modéliser la durée jusqu'à l'attrition de la clientèle

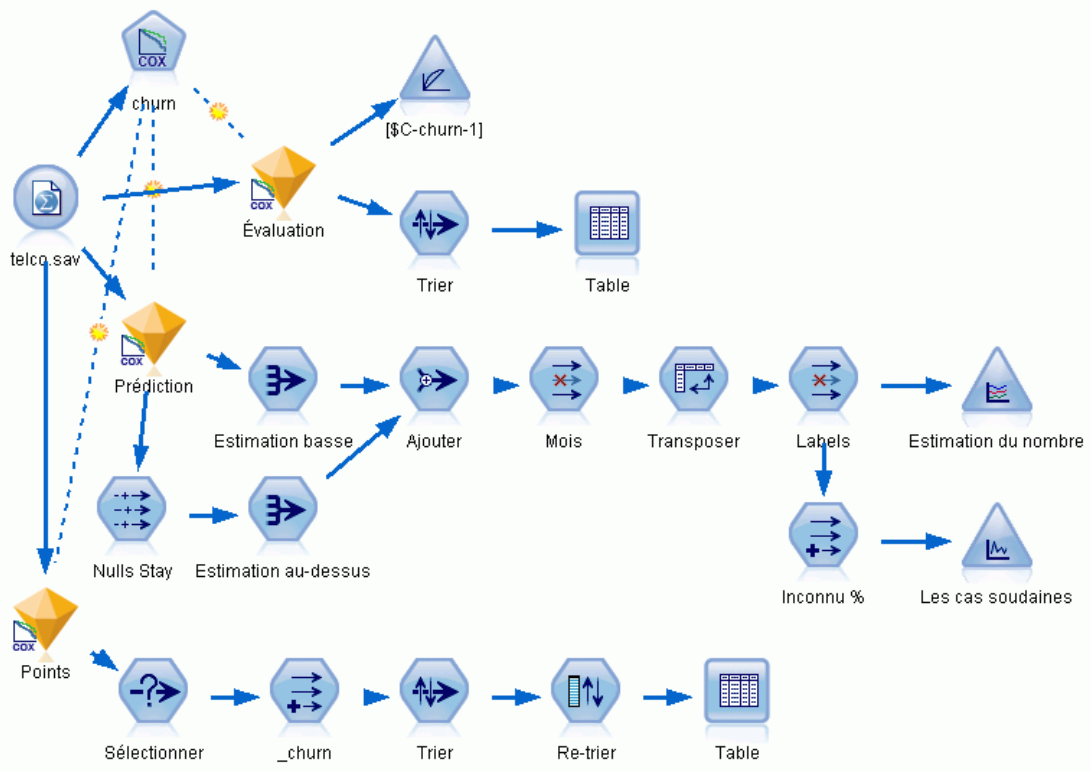
Dans ses efforts pour réduire l'attrition de la clientèle, une entreprise de télécommunication s'intéresse à la modélisation de la "durée jusqu'à l'attrition" afin de déterminer les facteurs associés aux clients qui changent rapidement de service. A cette fin, un échantillon aléatoire de clients est sélectionné et la période pendant laquelle ils ont été client, qu'ils soient encore des clients actifs ou non, ainsi que différents autres champs sont extraits de la base de données.

Cet exemple utilise le flux *telco_coxreg.str*, qui fait référence au fichier de données *telco.sav*. The data file is in the *Demos* folder and the stream file is in the *streams* subfolder. Pour plus d'informations, reportez-vous à la section Dossier Demos dans le chapitre 1 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.

Création d'un modèle adapté

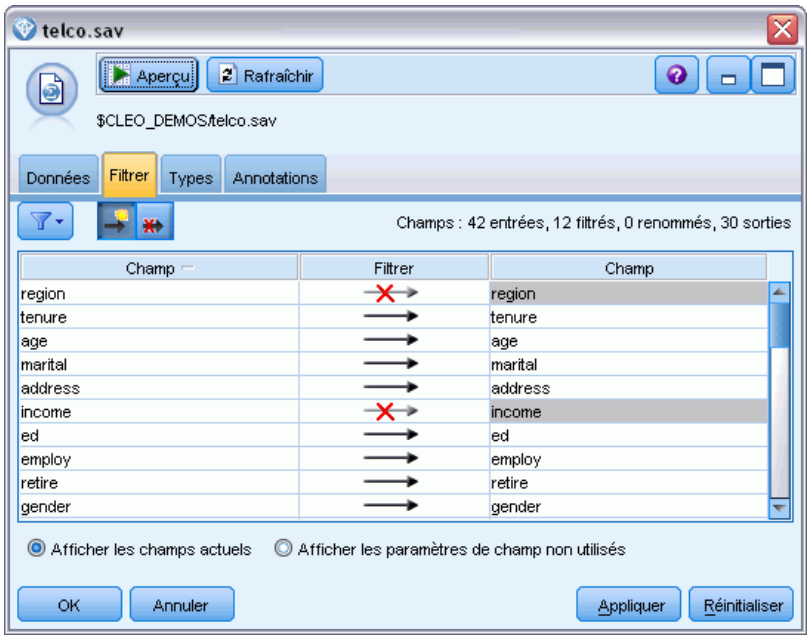
- Ajoutez un noeud source Fichier de statistiques pointant vers *telco.sav* dans le dossier *Demos*.

Figure 26-1
Exemple de flux pour l'analyse de la durée jusqu'à l'attrition



- Dans l'onglet Filtrer du noeud source, excluez les champs *region*, *income*, *longten* de *wireten*, et *loglong* à*logwire*.

Figure 26-2
Filtrage des champs inutiles

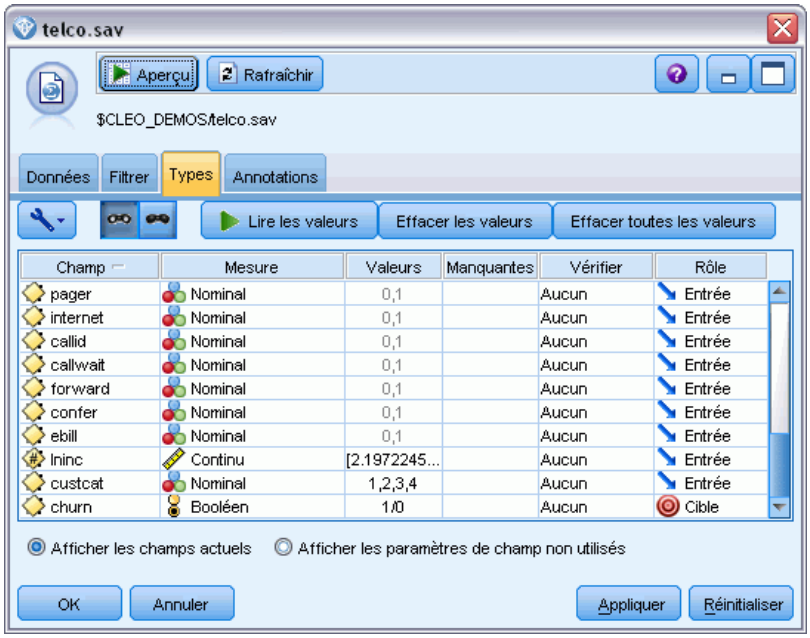


(Vous pouvez également régler le rôle de ces champs sur Aucun dans l'onglet Types au lieu de l'exclure ou sélectionner les champs que vous souhaitez utiliser dans le noeud de modélisation.)

- Dans l'onglet Types du noeud source, définissez le rôle du champ *attrition* sur Cible et son niveau de mesure sur Booléen. Le rôle de tous les autres champs doit être défini sur Entrée.

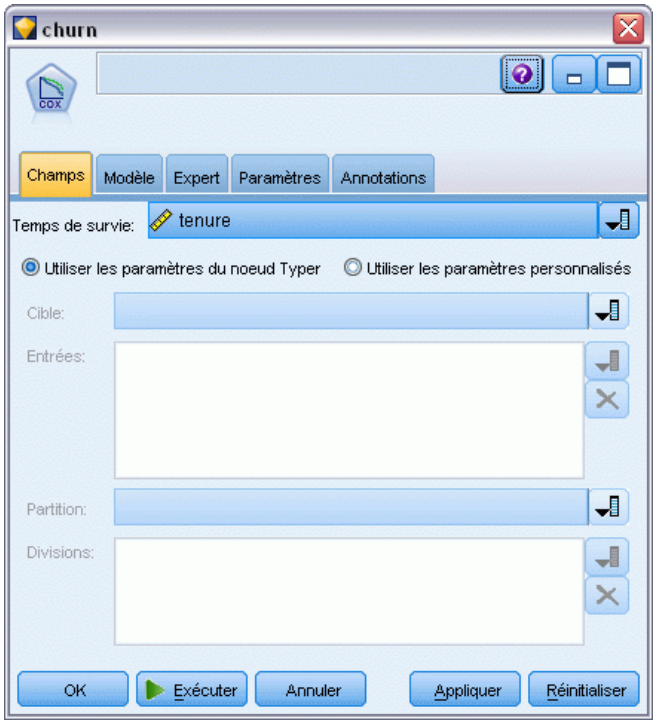
▶ Cliquez sur Lire les valeurs pour instancier les données.

Figure 26-3
Définition du rôle de champ



- Liez un noeud Cox au noeud source : dans l'onglet Champs, sélectionnez la variable de durée de survie *tenure*.

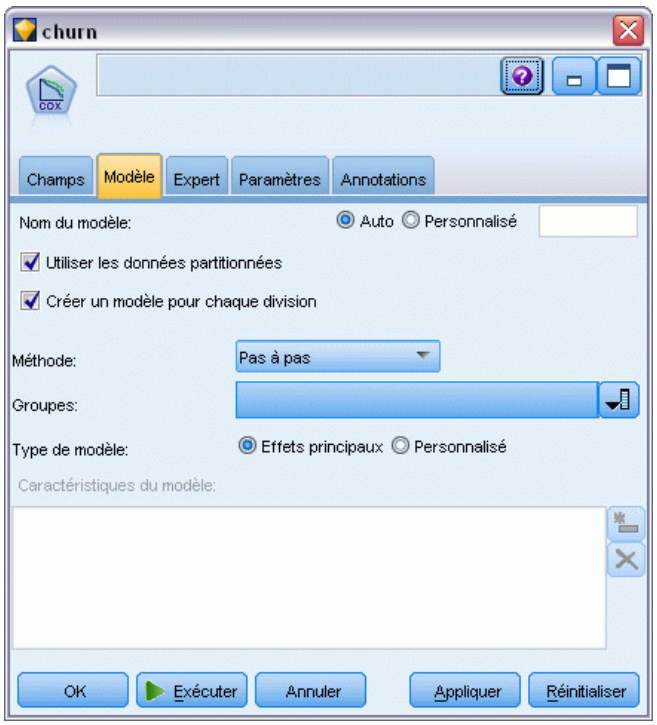
Figure 26-4
Sélection des options de champ



- Cliquez sur l'onglet Modèle.

▶ Sélectionnez la méthode de sélection de variables Pas à pas.

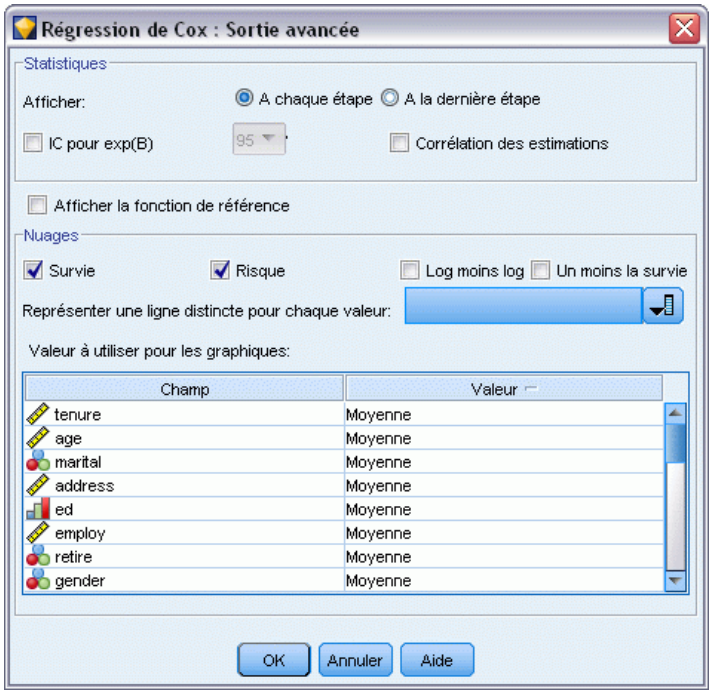
Figure 26-5
Choix des options de modèle



▶ Cliquez sur l'onglet Expert et sélectionnez Expert pour activer les options de modélisation expert.

▶ Cliquez sur Résultat.

Figure 26-6
Choix des options de sortie avancées



▶ Sélectionnez Survie et Risque comme nuages à créer puis cliquez sur OK.

▶ Cliquez sur Exécuter pour créer le nugget de modèle qui est ajouté au flux et à la palette Modèles en haut à droite. Pour en afficher les détails, double-cliquez sur le nugget dans le flux. Pour commencer, examinez l'onglet Sorties Avancées.

Observations censurées

Figure 26-7
Récapitulatif du traitement des observations

		N	Pourcentage
Observations disponibles dans l'analyse	**Evénement(a)**	274	27,4%
	Censurée	726	72,6%
	Total	1000	100,0%
Observations enlevées	**Observations avec valeurs manquantes**	0	,0%
	Observations avec durée négative	0	,0%
	Observations censurées avant l'événement le plus ancien dans une strate	0	,0%
	Total	0	,0%
Total		1000	100,0%

a. Variable dépendante : Months with service

La variable d'état identifie si l'évènement s'est produit pour une observation donnée. Lorsque l'évènement ne s'est pas produit, l'observation est dite censurée. Les observations censurées ne sont pas utilisées pour le calcul des coefficients de régression mais sont utilisées pour calculer le risque de référence. Le récapitulatif du traitement des observations affiche 726 observations censurées. Il s'agit des clients qui ne sont pas partis.

Codages de variables catégorielles

Figure 26-8
Codages de variables catégorielles

		Fréquence	(1)(s)	(2)	(3)	(4)
marital(t)	0=Unmarried	505	1			
	1=Married	495	0			
ed(t)	1=Did not complete high school	204	1	0	0	0
	2=High school degree	287	0	1	0	0
	3=Some college	209	0	0	1	0
	4=College degree	234	0	0	0	1
	5=Post-undergraduate degree	66	0	0	0	0
retire(t)	0=No	953	1			
	1=Yes	47	0			
gender(t)	0=Male	483	1			
	1=Female	517	0			
tollfree(t)	0=No	526	1			
	1=Yes	474	0			
equip(t)	0=No	614	1			
	1=Yes	386	0			
callcard(t)	0=No	322	1			
	1=Yes	678	0			
wireless(t)	0=No	704	1			
	1=Yes	296	0			
multline(t)	0=No	525	1			
	1=Yes	475	0			
voice(t)	0=No	696	1			
	1=Yes	304	0			
pager(t)	0=No	739	1			
	1=Yes	261	0			
internet(t)	0=No	632	1			
	1=Yes	368	0			
callid(t)	0=No	519	1			
	1=Yes	481	0			
callwait(t)	0=No	515	1			
	1=Yes	485	0			
forward(t)	0=No	507	1			
	1=Yes	493	0			
confer(t)	0=No	498	1			
	1=Yes	502	0			
ebill(t)	0=No	629	1			
	1=Yes	371	0			
custcat(t)	1=Basic service	266	1	0	0	
	2=E-service	217	0	1	0	
	3=Plus service	281	0	0	1	
	4=Total service	236	0	0	0	

Les codages de variables catégorielles sont une référence utile pour l'interprétation des coefficients de régression des covariables catégorielles, et particulièrement des variables dichotomiques. Par défaut, la catégorie de référence est la "dernière" catégorie. Ainsi, par exemple, bien que les clients dont le statut est *Marié* aient des valeurs de variable de 1 dans le fichier de données, ils sont codés comme 0 pour les besoins de la régression.

Sélection des variables

Figure 26-9
Tests composites

Etape	-2log-vraisemblance	Global (note)			Changement de l'étape précédente			Changement du bloc précédent		
		Khi-deux	ddl	Signif.	Khi-deux	ddl	Signif.	Khi-deux	ddl	Signif.
1(c)	3392,536	162,303	1	,000	133,828	1	,000	133,828	1	,000
2(d)	3087,314	249,392	2	,000	305,222	1	,000	439,050	2	,000
3(e)	3027,085	328,426	3	,000	60,229	1	,000	499,279	3	,000
4(f)	2990,790	347,197	4	,000	36,294	1	,000	535,574	4	,000
5(g)	2973,790	362,673	5	,000	17,000	1	,000	552,574	5	,000
6(h)	2958,796	376,140	6	,000	14,994	1	,000	567,568	6	,000
7(i)	2945,503	384,717	7	,000	13,293	1	,000	580,861	7	,000
8(j)	2936,993	417,341	8	,000	8,510	1	,004	589,371	8	,000
9(k)	2926,000	423,911	9	,000	10,994	1	,001	600,364	9	,000
10(l)	2917,551	428,078	10	,000	8,449	1	,004	608,813	10	,000
11(m)	2913,308	436,837	11	,000	4,243	1	,039	613,056	11	,000
12(n)	2908,078	440,158	12	,000	5,230	1	,022	618,286	12	,000

a. Bloc de départ numéro 0, fonction de log-vraisemblance initiale : -2log-vraisemblance : 3526,364

b. Bloc de départ numéro 1. Méthode = Ascendante pas à pas (rapport de vraisemblance)

c. Variable(s) entrée(s) à l'étape numéro 1: callcard

d. Variable(s) entrée(s) à l'étape numéro 2: longmon

e. Variable(s) entrée(s) à l'étape numéro 3: equip

f. Variable(s) entrée(s) à l'étape numéro 4: employ

g. Variable(s) entrée(s) à l'étape numéro 5: multline

h. Variable(s) entrée(s) à l'étape numéro 6: voice

i. Variable(s) entrée(s) à l'étape numéro 7: address

j. Variable(s) entrée(s) à l'étape numéro 8: equipmon

k. Variable(s) entrée(s) à l'étape numéro 9: ebill

l. Variable(s) entrée(s) à l'étape numéro 10: callid

m. Variable(s) entrée(s) à l'étape numéro 11: internet

n. Variable(s) entrée(s) à l'étape numéro 12: reside

Le processus de création de modèle utilise un algorithme Pas à pas ascendant. Les tests composites sont des mesures de la performance des modèles. La modification du Chi-deux de l'étape précédente est la différence entre le−2 log de vraisemblance du modèle à l'étape précédente et à l'étape actuelle. Si cette étape devait ajouter une variable, l'inclusion serait logique si la signification de cette modification était inférieure à 0,05. Si cette étape devait supprimer une variable, l'exclusion serait logique si la signification de cette modification était supérieure à 0,10. Au cours des douze étapes, douze variables sont ajoutées au modèle.

Figure 26-10
Variables de l'équation (étape 12 uniquement)

		B	E.S.	Wald	ddl	Signif.	Exp(B)
Etape 12	address	-,035	,009	14,543	1	,000	,966
	employ	-,051	,010	25,767	1	,000	,950
	reside	-,103	,046	5,037	1	,025	,902
	equip	-1,948	,381	26,180	1	,000	,143
	callcard	,777	,151	26,451	1	,000	2,175
	longmon	-,233	,022	115,619	1	,000	,792
	equipmon	-,042	,011	15,377	1	,000	,959
	multline	,612	,145	17,854	1	,000	1,844
	voice	-,501	,157	10,197	1	,001	,606
	internet	-,362	,160	5,114	1	,024	,697
	callid	-,464	,148	9,790	1	,002	,629
	ebill	-,399	,156	6,557	1	,010	,671

Le modèle final contient les variables *address*, *employ*, *reside*, *equip*, *callcard*, *longmon*, *equipmon*, *multline*, *voice*, *internet*, *callid*, et *ebill*. Pour comprendre les effets des variables indépendantes individuelles, examinez Exp(B) qui peut être interprété comme la modification prédite du risque d'augmentation d'unités dans la variable indépendante.

- La valeur de Exp(B) pour la variable *address* signifie que le risque d'attrition est réduit de 100%−(100%×0,966)=3,4% pour chaque année où le client a vécu à la même adresse. Le risque d'attrition pour un client ayant vécu à la même adresse pendant cinq ans est réduit de $100\%-(100\%\times 0{,}966^5)=15{,}88\%$.
- La valeur de Exp(B) pour la variable *callcard* signifie que le risque d'attrition pour un client ne s'étant pas abonné au service de carte téléphonique est 2,175 fois plus élevé que celui pour un client s'étant abonné à ce service. Souvenez-vous que selon les codages de variable catégorielle *Non* = 1 pour la régression.
- La valeur de Exp(B) pour la variable *internet* signifie que le risque d'attrition pour un client ne s'étant pas abonné au service Internet est 0,697 fois plus élevé que celui pour un client s'étant abonné à ce service. Ce chiffre est quelque peu inquiétant car il indique que les clients abonnés au service quittent plus rapidement l'entreprise que ceux n'étant pas abonnés.

Figure 26-11
Variables absentes du modèle (étape 12 uniquement)

		Score	ddl	Signif.
Etape 12	age	,122	1	,726
	marital	,648	1	,421
	ed	6,328	4	,176
	ed(1)	,007	1	,934
	ed(2)	,203	1	,652
	ed(3)	,835	1	,361
	ed(4)	5,773	1	,016
	retire	,013	1	,908
	gender	,214	1	,644
	tollfree	3,243	1	,072
	wireless	,668	1	,414
	tollmon	,000	1	,987
	cardmon	3,163	1	,075
	wiremon	1,084	1	,298
	pager	1,808	1	,179
	callwait	,266	1	,606
	forward	2,201	1	,138
	confer	2,568	1	,109
	lninc	2,853	1	,091
	custcat	,864	3	,834
	custcat(1)	,466	1	,495
	custcat(2)	,450	1	,502
	custcat(3)	,019	1	,889

Les variables absentes du modèle ont toutes des statistiques de score avec des valeurs de signification supérieures à 0,05. Cependant, les valeurs de signification de *tollfree* et de *cardmon*, bien qu'égales ou supérieures à 0,05, en sont relativement proches. Il serait intéressant d'étudier cette question plus avant.

Moyennes des covariables

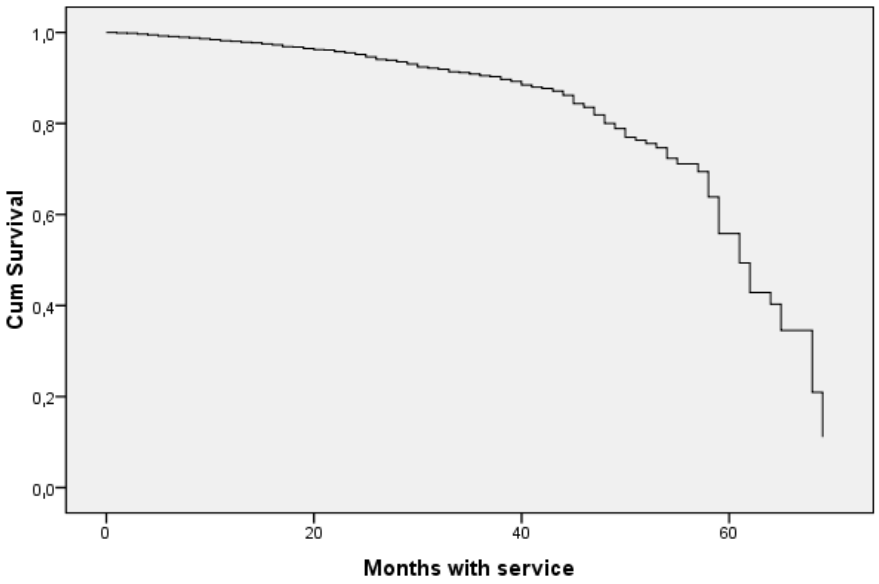
Figure 26-12
Moyennes des covariables

	Moyenne
age	41,684
marital	,505
address	11,551
ed(1)	,204
ed(2)	,287
ed(3)	,209
ed(4)	,234
employ	10,987
retire	,953
gender	,483
reside	2,331
tollfree	,526
equip	,614
callcard	,322
wireless	,704
longmon	11,723
tollmon	13,274
equipmon	14,220
cardmon	13,781
wiremon	11,584
multline	,525
voice	,696
pager	,739
internet	,632
callid	,519
callwait	,515
forward	,507
confer	,498
ebill	,629
lninc	3,957
custcat(1)	,266
custcat(2)	,217
custcat(3)	,281

Ce tableau affiche la valeur moyenne de chaque variable de prédiction. Ce tableau est une référence utile pour l'examen des nuages de survie qui sont construits pour les valeurs moyennes. Cependant, veuillez noter que le client "moyen" n'existe pas si l'on examine les moyennes des variables indicatrices pour les variables indépendantes catégorielles. Même avec toutes les variables indépendantes d'échelle, il est peu probable que vous trouviez un client dont les valeurs de covariables soient toutes proches de la moyenne. Si vous souhaitez visualiser la courbe de survie d'une observation spécifique, vous pouvez modifier les valeurs de covariables qui déterminent le tracé de la courbe de survie dans la boîte de dialogue Nuages. Si vous souhaitez visualiser la courbe de survie d'une observation spécifique, vous pouvez modifier les valeurs de covariables qui déterminent le tracé de la courbe de survie dans le groupe Nuages de la boîte de dialogue Sorties avancées.

Courbe de survie

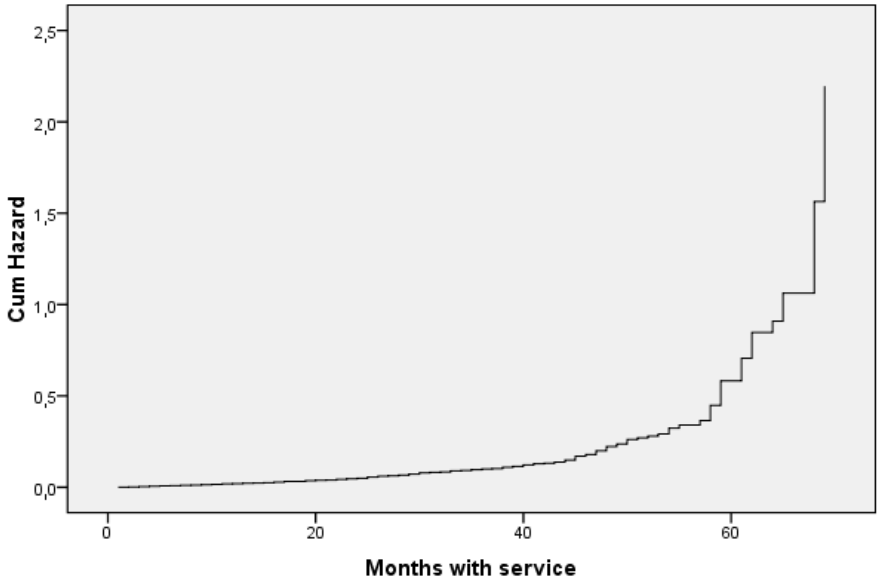
Figure 26-13
Courbe de survie pour le client "moyen"



La courbe de survie de base est un affichage visuel de la durée jusqu'à l'attrition pour le client "moyen" prédite par le modèle. L'axe horizontal indique la durée jusqu'à l'évènement. L'axe vertical indique la probabilité de survie. Ainsi, tous les points de la courbe de survie indiquent la probabilité que le client "moyen" reste client cette durée passée. Après 55 mois, la courbe de survie devient plus irrégulière. Il existe moins de clients restés clients de l'entreprise pendant aussi longtemps et les informations disponibles sont plus rares, ce qui crée une courbe en dents de scie.

Courbe de risque

Figure 26-14
Courbe de risque pour le client "moyen"

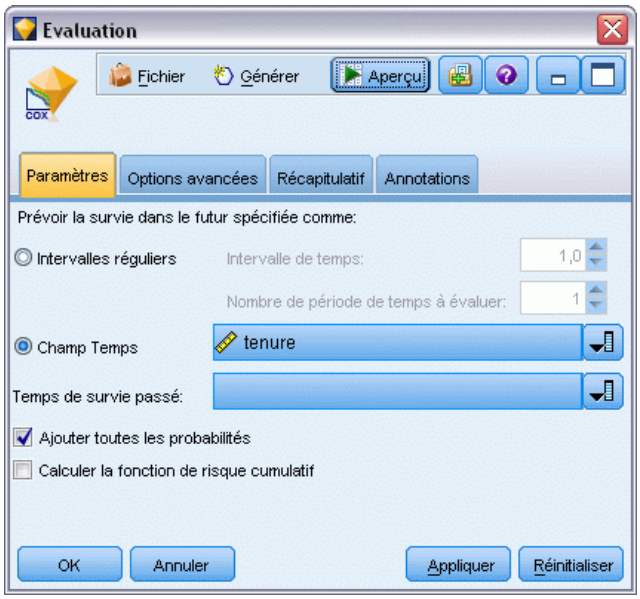


La courbe de risque de base est un affichage visuel des risques cumulatifs d'attrition pour le client "moyen" prédits par le modèle. L'axe horizontal indique la durée jusqu'à l'évènement. L'axe vertical indique les risques cumulatifs, égaux au log négatif de la probabilité de survie. Après 55 mois, la courbe de risque, comme la courbe de survie, devient plus irrégulière pour la même raison.

Evaluation

Les méthodes de sélection pas à pas garantissent que votre modèle n'ait que des variables indépendantes "significatives en termes de statistiques", mais ne garantissent pas que le modèle soit approprié pour prédire la cible. Pour ceci, vous devez analyser les enregistrements évalués.

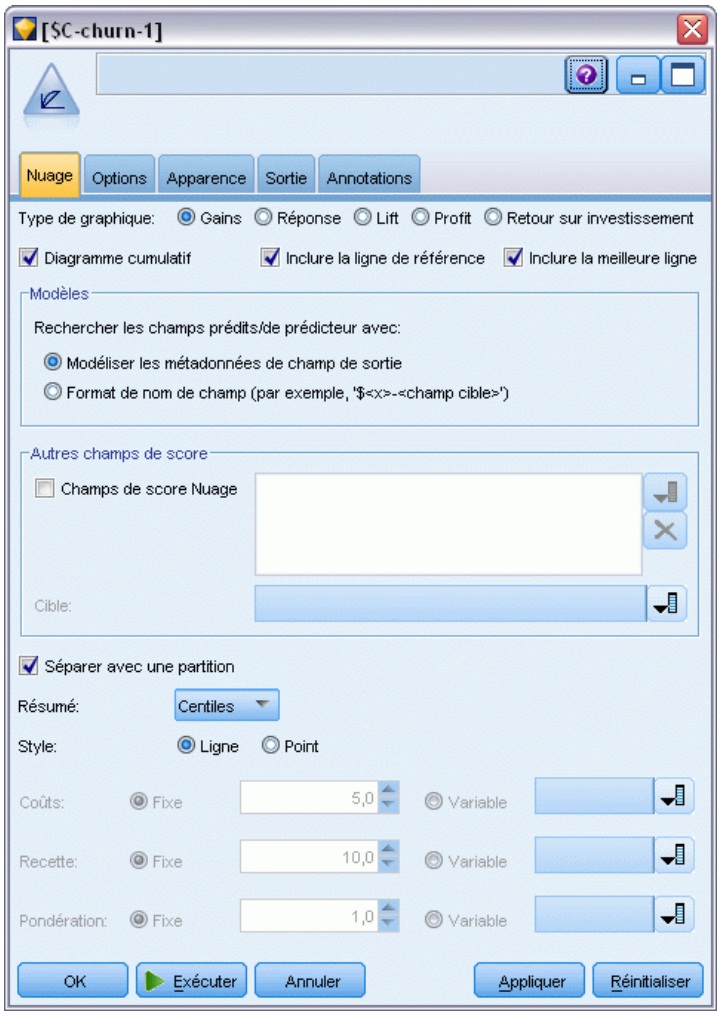
Figure 26-15
Nugget de Cox : onglet Paramètres



- Placez le nugget du modèle dans l'espace de travail et liez-le au noeud source, ouvrez le nugget et cliquez sur l'onglet Paramètres.
- Sélectionnez un Champ temporel et définissez la *durée d'affectation*. Chaque enregistrement sera évalué en fonction de sa durée d'affectation.
- Sélectionnez Ajouter toutes les probabilités.

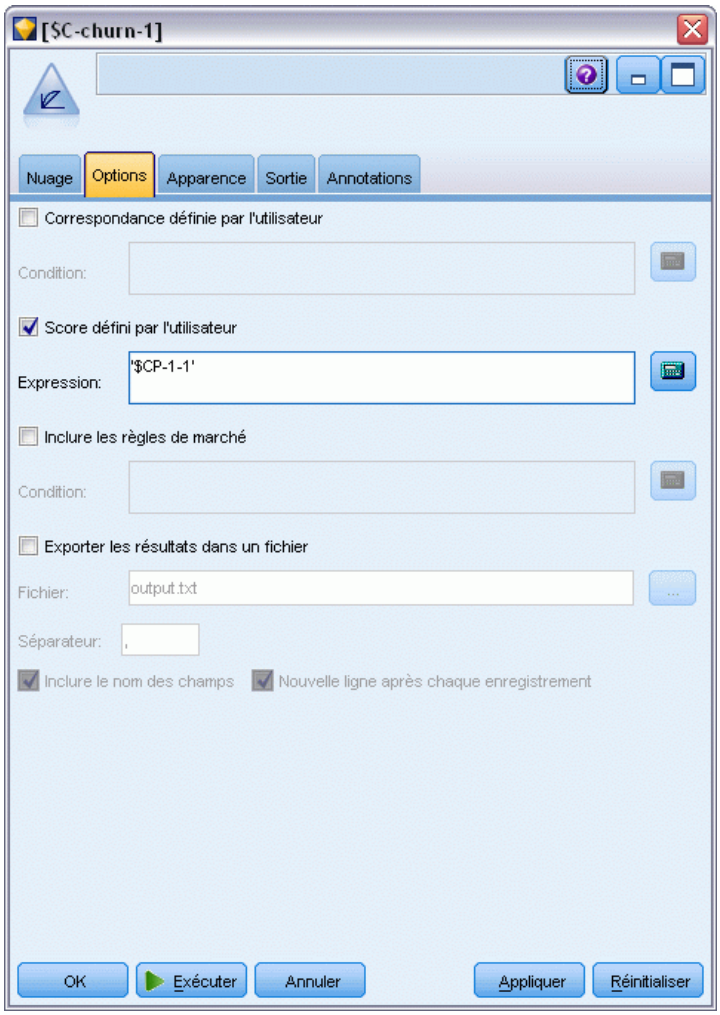
Cette action crée des évaluations qui utilisent 0,5 comme césure de l'attrition des clients ; si leur propension à quitter le service est supérieure à 0,5, ils sont évalués comme clients perdus. Ce chiffre n'est pas un chiffre magique et une césure différente peut offrir de meilleurs résultats. Utiliser le noeud Evaluation est une façon de choisir la césure.

Figure 26-16
Noeud Evaluation : onglet Nuage



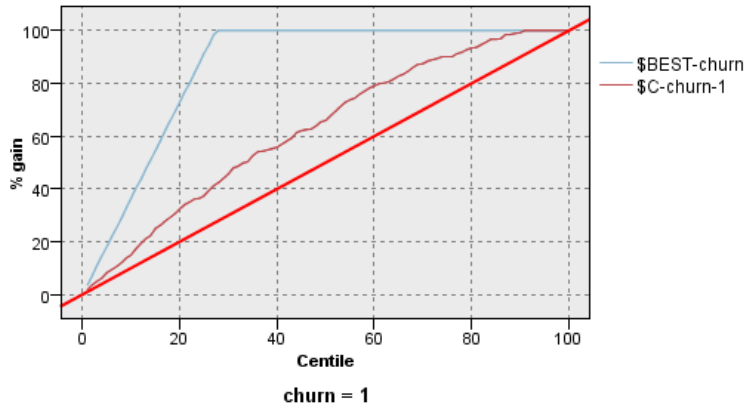
- Liez un noeud Evaluation au nugget de modèle ; dans l'onglet Nuage, sélectionnez Inclure la meilleure ligne.
- Cliquez sur l'onglet Options.

Figure 26-17
Noeud Evaluation : Onglet Options



- Sélectionnez Score défini par l'utilisateur et saisissez l'expression '$CP-1-1'. C'est un champ généré par le modèle qui correspond à la propension à l'attrition.
- Cliquez sur Exécuter.

Figure 26-18
Graphiques de gains

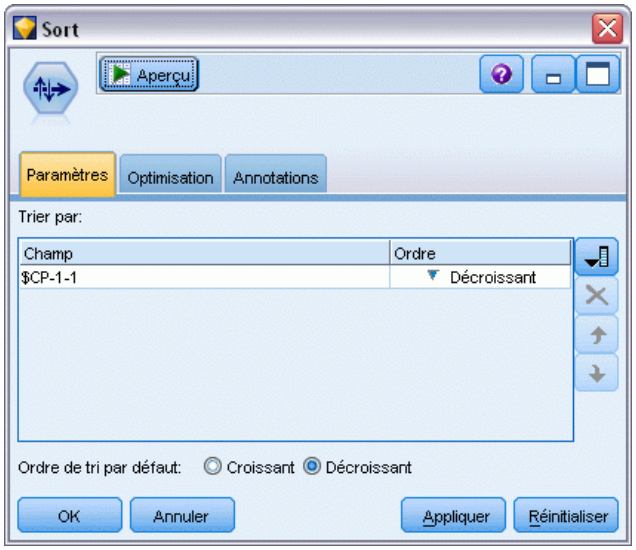


Le diagramme de gains cumulés montre le pourcentage du nombre total d'observations dans une modalité donnée obtenu en ciblant un pourcentage du nombre total d'observations. Par exemple, un point de la courbe est à (10%, 15%), ce qui signifie que si vous évaluez un ensemble de données avec le modèle et triez toutes les observations en fonction de la propension à l'attrition prédite, les premiers 10% devraient contenir environ 15% de toutes les observations qui correspondent à la catégorie *1* (clients perdus). De la même façon, les premiers 60% contiennent environ 79,2% des clients perdus. Si vous sélectionnez 100% de l'ensemble de données évalué, tous les clients perdus sont dans cet ensemble de données.

La diagonale est la courbe "de référence" ; si vous sélectionnez au hasard 20% des enregistrements de l'ensemble de données évalué, vous devriez "obtenir" environ 20% de tous les enregistrements qui correspondent à la catégorie *1*. Plus une courbe se situe au-dessus de la ligne de base, plus le gain est élevé. La "meilleure" ligne représente la courbe d'un modèle "parfait" qui attribue un score de propension à l'attrition plus élevé à tous les clients perdus plutôt qu'à tous les clients non perdus. Vous pouvez utiliser le diagramme de gains cumulés pour sélectionner une césure de classement en choisissant un pourcentage correspondant à un gain souhaitable, puis en associant ce pourcentage à la valeur de césure appropriée.

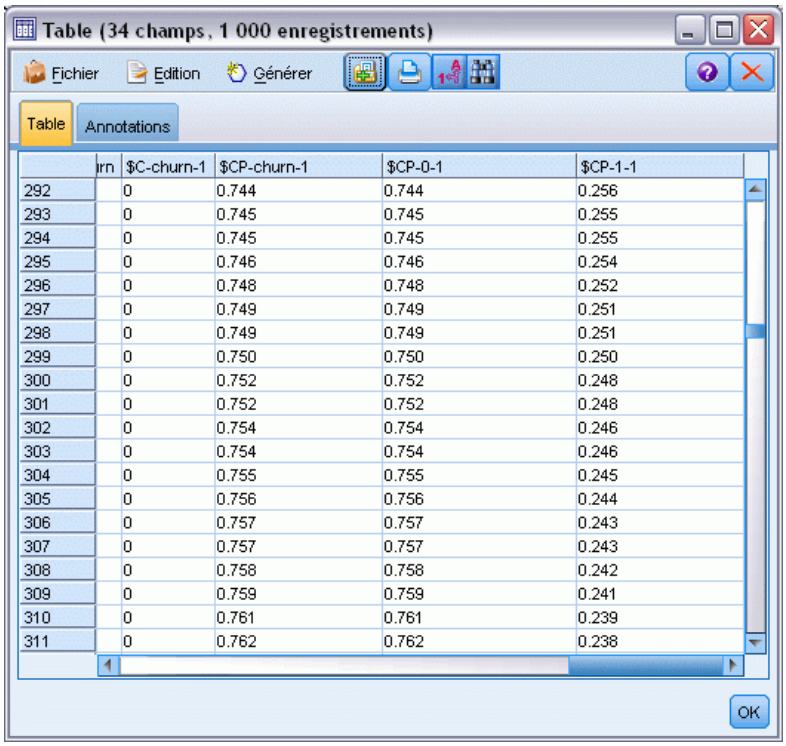
Ce qui constitue un gain « souhaitable » dépend du coût des erreurs de type I et de type II. C'est-à-dire, combien coûte le classement d'un client perdu en client retenu (Type I) ? Combien coûte le classement d'un client retenu en client perdu (Type II) ? Si la rétention du client est la préoccupation principale, il vous faut alors diminuer votre erreur de Type I ; sur le graphique des gains cumulatifs, cela peut correspondre à une augmentation de l'assistance clientèle pour les clients faisant partie des premiers 60% de la propension prédite de *1*, qui rassemblent 79,2% des clients perdus potentiels mais qui coûtent cher en temps et en ressources qui pourraient être utilisés pour acquérir de nouveaux clients. Si la baisse du coût du maintien de votre base de clientèle actuelle est votre priorité, diminuez alors votre erreur de Type II. Sur le graphique, cela peut correspondre à une augmentation de l'assistance clientèle pour les premiers 20% qui rassemblent 32,5% des clients perdus. Généralement, ces deux préoccupations sont importantes et il est nécessaire de choisir une règle de décision qui classifie les clients et qui offre le meilleur compromis entre la sensibilité et la spécificité.

Figure 26-19
Noeud Trier : onglet Paramètres



- Imaginons que vous avez décidé que 45,6% est un gain acceptable ce qui correspond à utiliser les premiers 30% des enregistrements. Pour trouver une césure de classification appropriée, liez un noeud Trier au nugget de modèle.
- Dans l'onglet Paramètres, choisissez de trier par *$CP-1-1* dans l'ordre décroissant puis cliquez sur OK.

Figure 26-20
Table



	rn	$C-churn-1	$CP-churn-1	$CP-0-1	$CP-1-1
292		0	0.744	0.744	0.256
293		0	0.745	0.745	0.255
294		0	0.745	0.745	0.255
295		0	0.746	0.746	0.254
296		0	0.748	0.748	0.252
297		0	0.749	0.749	0.251
298		0	0.749	0.749	0.251
299		0	0.750	0.750	0.250
300		0	0.752	0.752	0.248
301		0	0.752	0.752	0.248
302		0	0.754	0.754	0.246
303		0	0.754	0.754	0.246
304		0	0.755	0.755	0.245
305		0	0.756	0.756	0.244
306		0	0.757	0.757	0.243
307		0	0.757	0.757	0.243
308		0	0.758	0.758	0.242
309		0	0.759	0.759	0.241
310		0	0.761	0.761	0.239
311		0	0.762	0.762	0.238

- Reliez un noeud Table au noeud Trier.
- Ouvrez le noeud Table, puis cliquez sur Exécuter.

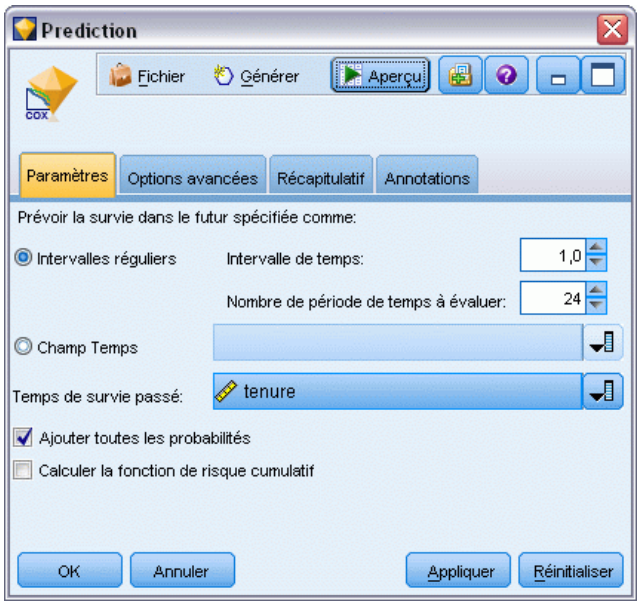
En faisant défiler les sorties vers le bas, vous pouvez voir que la valeur de *$CP-1-1* est de 0,248 pour le 300ème enregistrement. L'utilisation de 0.248 comme césure de classification devrait résulter approximativement en 30% des clients évalués comme clients perdus, capturant à peu près 45% du nombre total des clients perdus.

Suivi du nombre prévu de clients retenus

Lorsque le modèle vous convient, effectuez un suivi du nombre prévu de clients de l'ensemble de données qui sont retenus pendant les deux prochaines années. Les valeurs nulles, qui correspondent à des clients dont la durée d'affectation totale (temps futur + *durée d'affectation*) est en-dehors de la plage des durées de survie pour les données utilisées pour former le modèle, représentent un défi intéressant. Une façon de traiter ces valeurs est de créer deux ensembles de prédictions, un dans lequel on considère que les valeurs nulles ont été perdues et l'autre dans

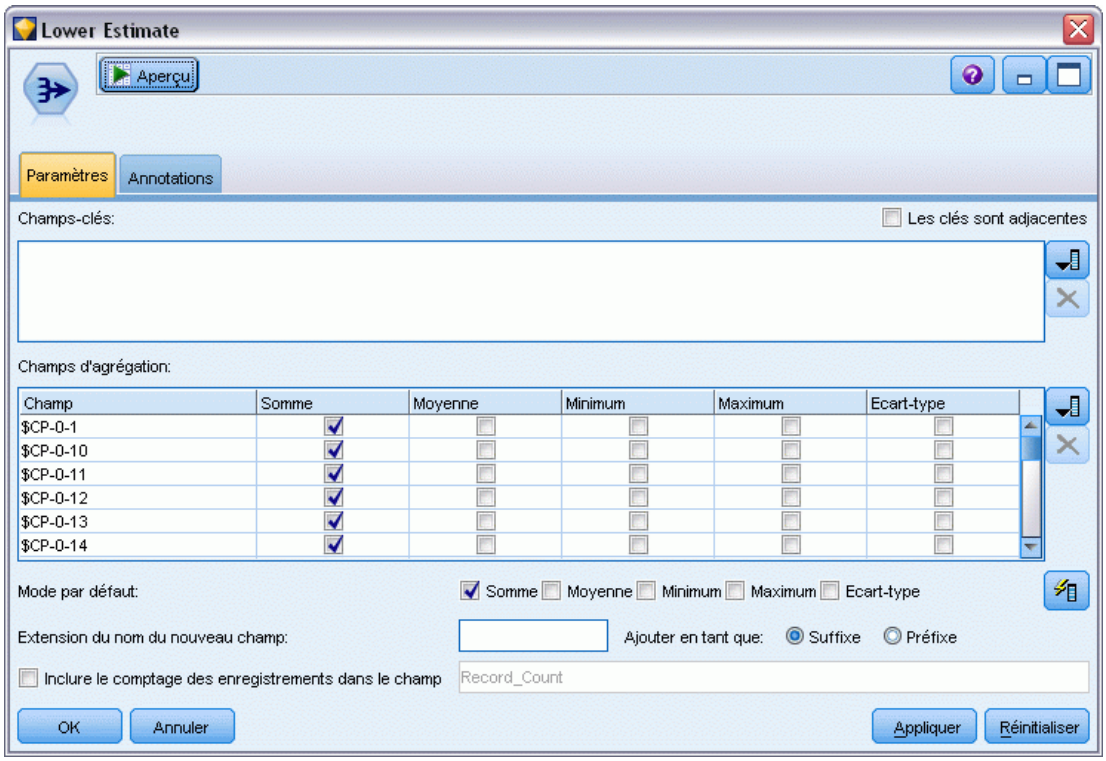
lequel on considère que ces valeurs ont été retenues. Ainsi, vous pouvez établir les limites supérieures et inférieures du nombre prévu de clients retenus.

Figure 26-21
Nugget de Cox : onglet Paramètres



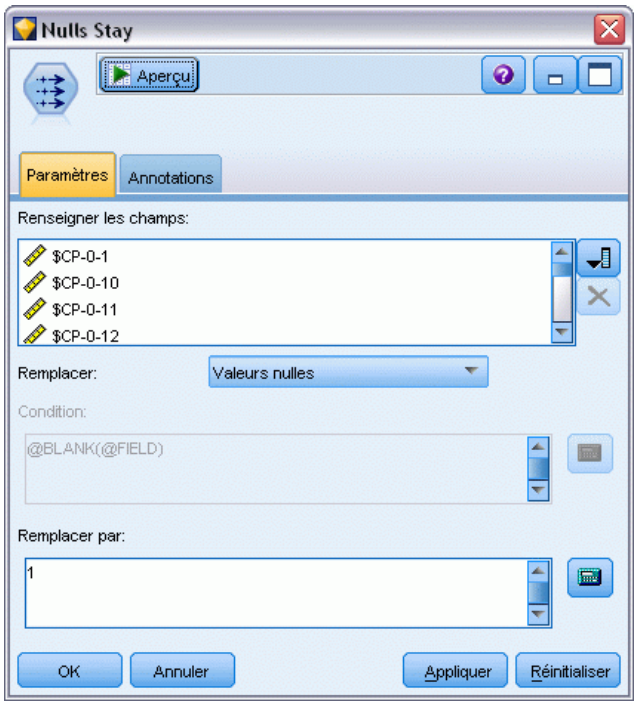
- Double-cliquez sur le nugget de modèle dans la palette Modèles (ou copiez-collez le nugget sur l'espace de travail du flux) et joignez le nouveau nugget au noeud Source.
- Ouvrez le nugget dans l'onglet Paramètres.
- Vérifiez que Intervalles réguliers est sélectionné et spécifiez 1,0 comme intervalle de temps et 24 comme nombre de périodes à évaluer. Chaque enregistrement sera ainsi évalué pendant chacun des 24 mois suivants.
- Sélectionnez le champ *durée d'affectation* pour spécifier la durée de survie passée. L'algorithme d'évaluation prendra en compte la durée de survie de chaque client comme client de l'entreprise.
- Sélectionnez Ajouter toutes les probabilités.

Figure 26-22
Noeud Agréger : onglet Paramètres



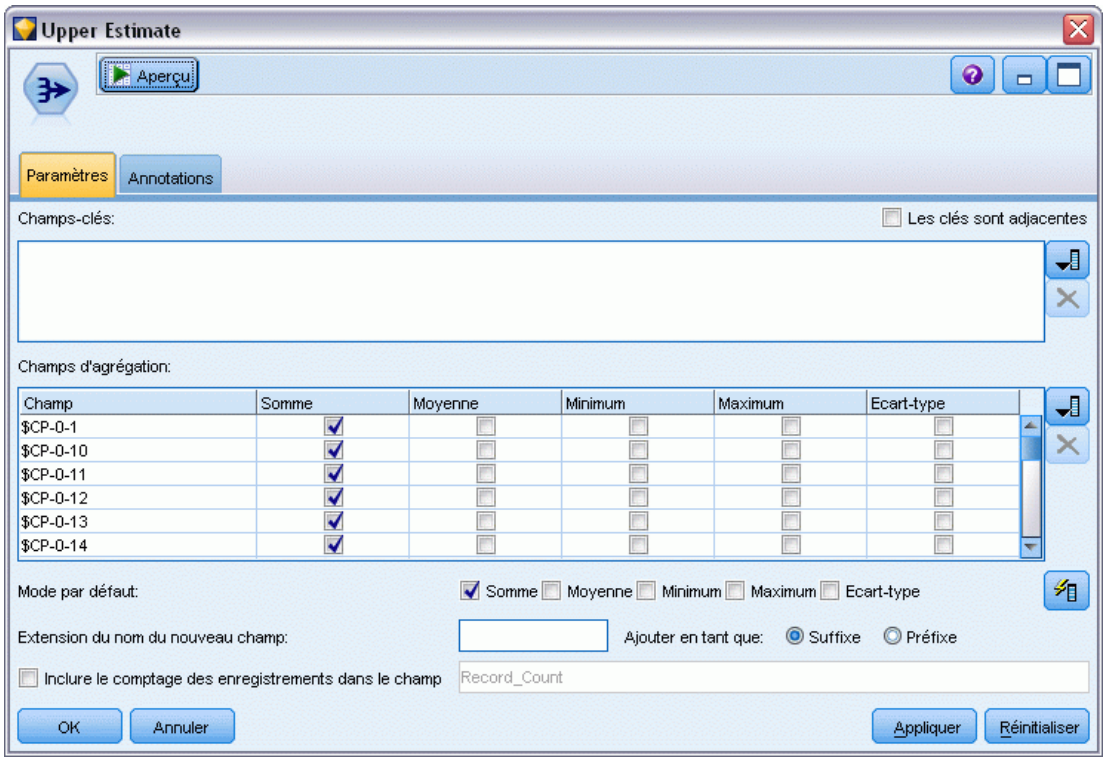
- Liez un noeud Agréger au nugget de modèle ; dans l'onglet Paramètres, désélectionnez le mode par défaut Moyenne.
- Sélectionnez *$CP-0-1* à *$CP-0-24*, les champs de forme *$CP-0-n*, étant les champs à agréger. Pour faciliter cette action, triez les champs par nom (c'est-à-dire par ordre alphabétique) dans la boîte de dialogue Sélectionner les champs.
- Désélectionnez Inclure le comptage des enregistrements dans le champ.
- Cliquez sur OK. Ce noeud crée les prévisions de "limite inférieure".

Figure 26-23
Noeud Remplacer : onglet Paramètres



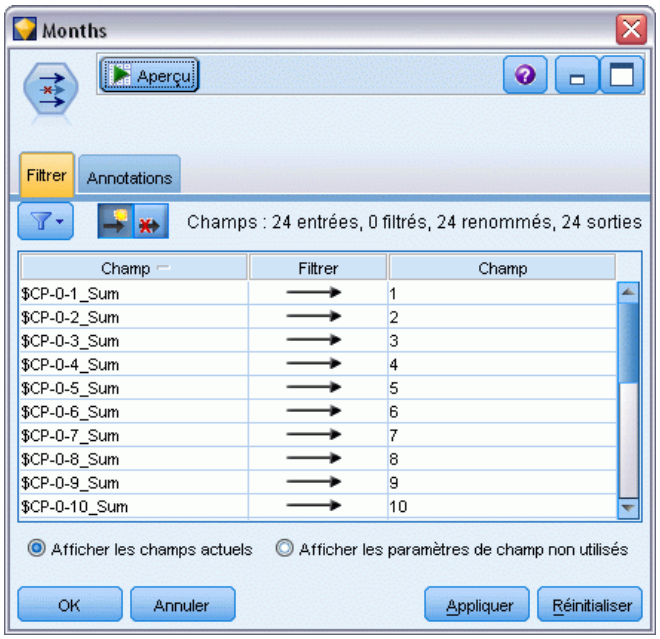
- Liez un noeud Remplacer au nugget Coxreg auquel le noeud Agréger vient d'être lié ; dans l'onglet Paramètres, sélectionnez *$CP-0-1* à *$CP-0-24*, les champs de forme *$CP-0-n*, étant les champs à remplir. Pour faciliter cette action, triez les champs par nom (c'est-à-dire par ordre alphabétique) dans la boîte de dialogue Sélectionner les champs.
- Choisissez de remplacer les Valeurs nulles par la valeur 1.
- Cliquez sur OK.

Figure 26-24
Noeud Agréger : onglet Paramètres



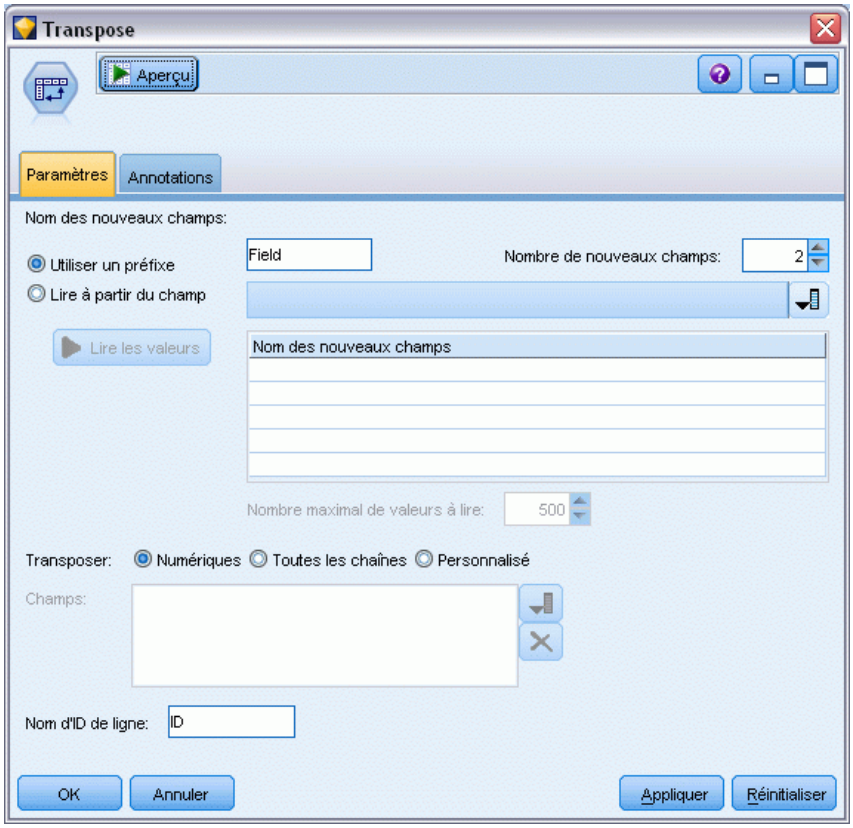
- Liez un noeud Agréger au noeud Remplacer ; dans l'onglet Paramètres, désélectionnez le mode par défaut Moyenne.
- Sélectionnez *$CP-0-1* à *$CP-0-24*, les champs de forme *$CP-0-n*, étant les champs à agréger. Pour faciliter cette action, triez les champs par nom (c'est-à-dire par ordre alphabétique) dans la boîte de dialogue Sélectionner les champs.
- Désélectionnez Inclure le comptage des enregistrements dans le champ.
- Cliquez sur OK. Ce noeud crée les prévisions de "limite supérieure".

Figure 26-25
Noeud Filtrer : onglet Paramètres



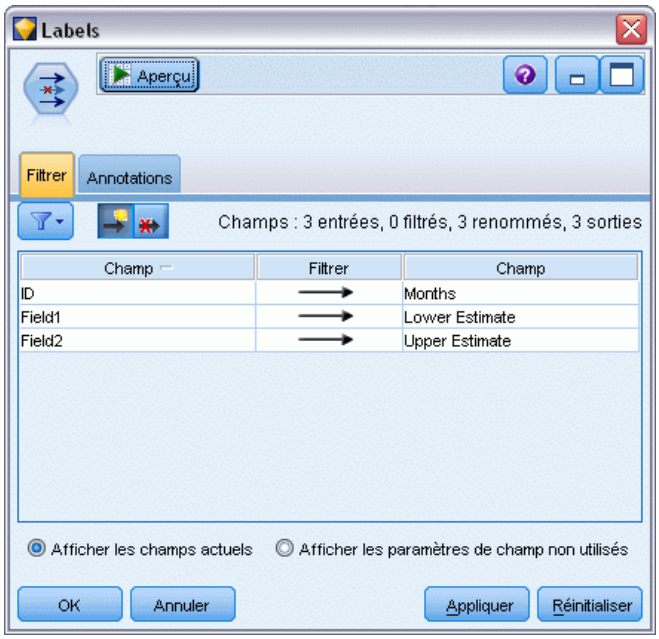
- Reliez un noeud Ajouter aux deux noeuds Agréger puis reliez un noeud Filtrer au noeud Ajouter.
- Dans l'onglet Paramètres du noeud Filtrer, renommez les champs de *1* à *24*. En utilisant un noeud Transposer, les noms de ces champs deviendront des valeurs de l'axe *x* dans les graphiques en aval.

Figure 26-26
Noeud Transposer : onglet Paramètres



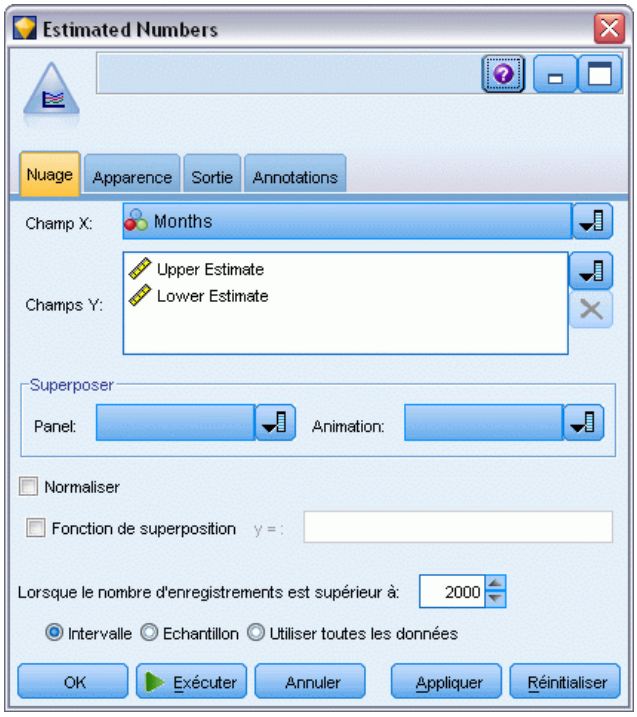
- Reliez un noeud Transposer au noeud Filtrer.
- Saisissez 2 comme nombre des nouveaux champs.

Figure 26-27
Noeud Filtrer : Onglet Filtrer



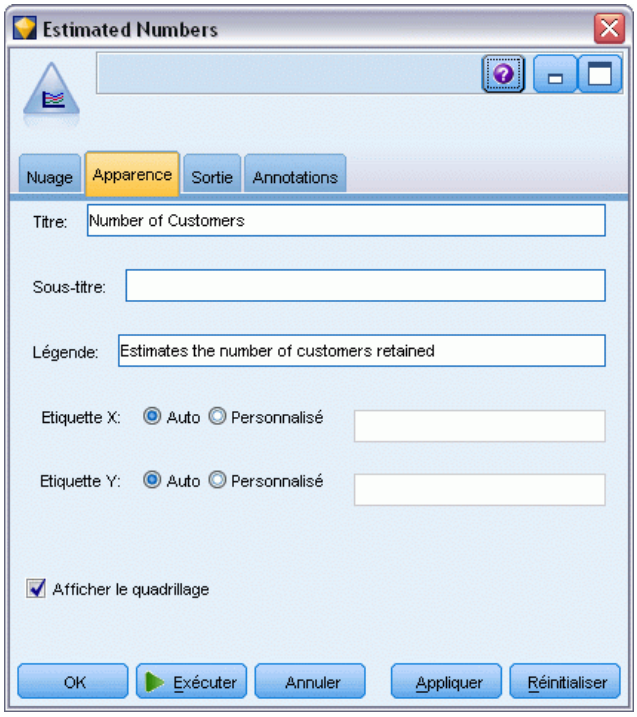
- Reliez un noeud Filtrer au noeud Transposer.
- Dans l'onglet Paramètres du noeud Filtrer, changez le nom de *ID* en *Mois*, *Champ1* en *Estimation inférieure*, et *Champ2* en *Estimation supérieure*.

Figure 26-28
Noeud Courbes : onglet Nuage



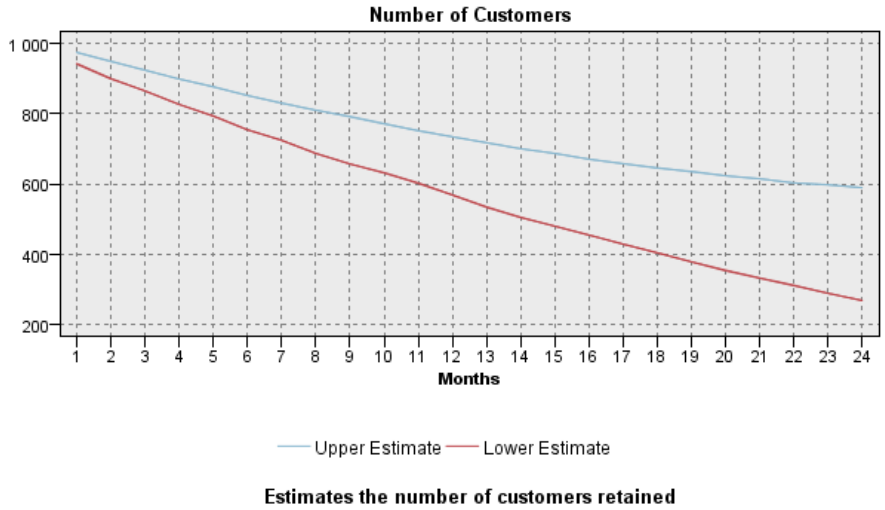
- Reliez un noeud Courbes au noeud Filtrer.
- Dans l'onglet Nuage, sélectionnez *Mois* comme champ X, *Estimation inférieure* et *Estimation supérieure* comme champs Y.

Figure 26-29
Noeud Courbes : onglet Apparence



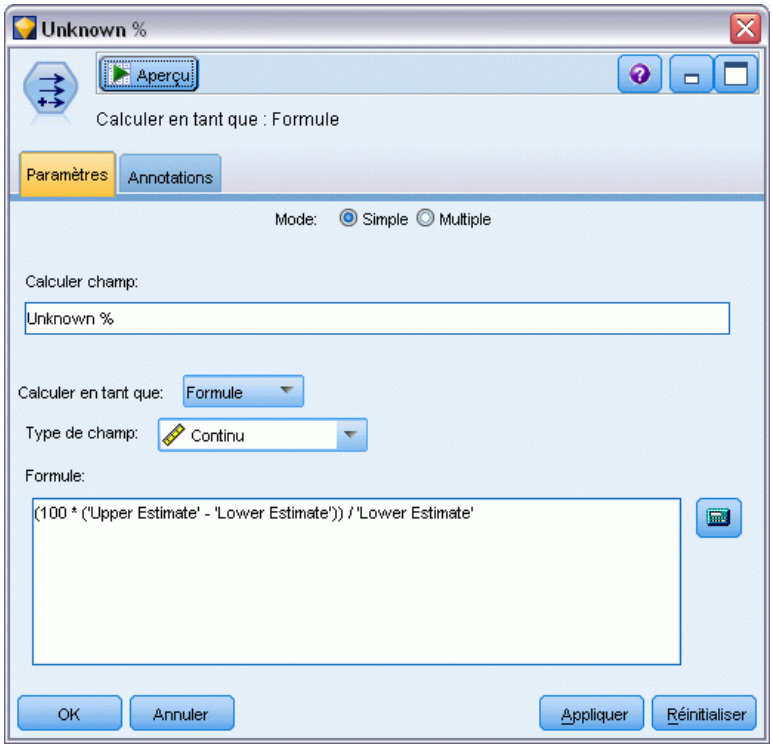
- Cliquez sur l'onglet Apparence.
- Saisissez Nombre de clients comme titre.
- Saisissez Estimations du nombre de clients retenus comme légende.
- Cliquez sur Exécuter.

Figure 26-30
Courbes estimant le nombre de clients retenus



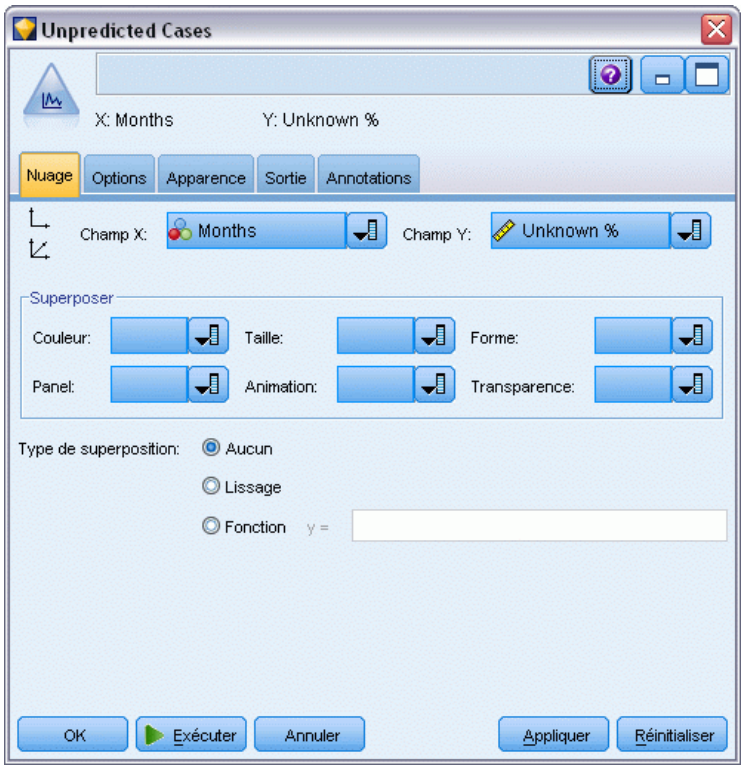
Les limites supérieures et inférieures du nombre estimé de clients retenus sont représentées. La différence entre les deux lignes est le nombre de clients évalués comme nuls et, par conséquent, dont l'état est fortement incertain. Au fil du temps, le nombre de ces clients augmente. Après 12 mois, vous devriez retenir entre 601 et 735 des clients d'origine de l'ensemble de données ; après 24 mois, entre 288 et 597.

Figure 26-31
Noeud Calculer : onglet Paramètres



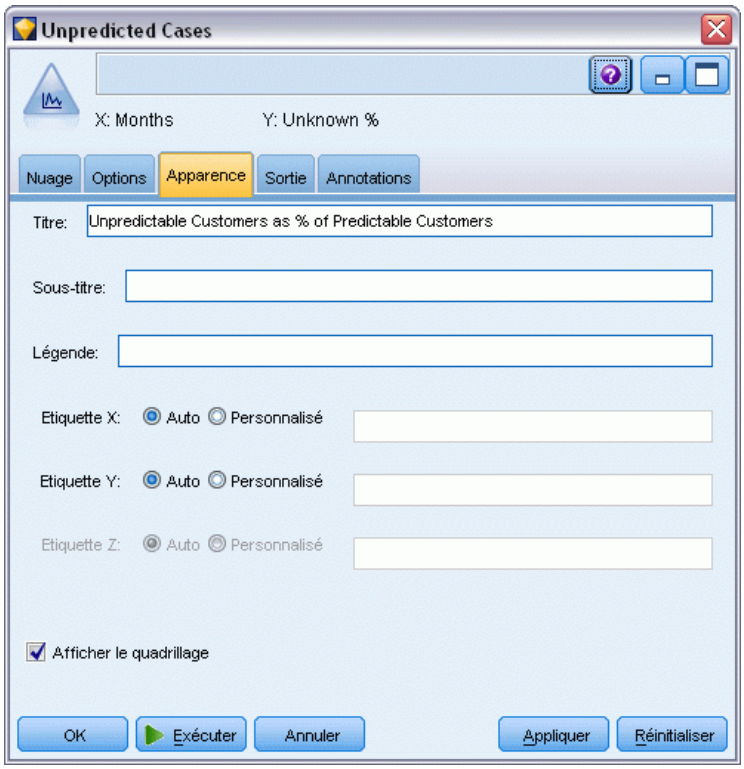
- Pour examiner de nouveau combien les estimations du nombre de clients retenus sont incertaines, reliez un noeud Calculer au noeud Filtrer.
- Dans l'onglet Paramètres du noeud Calculer, saisissez *% inconnu* comme champ de calcul.
- Sélectionnez Continu comme type de champ.
- Saisissez la formule (100 * ('Estimation supérieure' - 'Estimation inférieure')) / 'Estimation inférieure'. *% inconnnu* est le nombre de clients "dans le doute" sous la forme d'un pourcentage de l'estimation inférieure.
- Cliquez sur OK.

Figure 26-32
Noeud Nuage : onglet Nuage



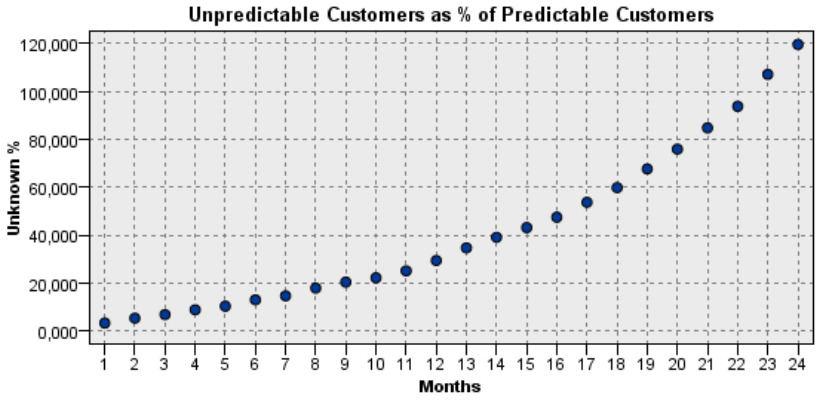
- Reliez un noeud Nuage au noeud Calculer.
- Dans l'onglet Nuage du noeud Nuage, sélectionnez *Mois* comme champ X et *% inconnu* comme champ Y.
- Cliquez sur l'onglet Apparence.

Figure 26-33
Noeud Nuage : onglet Apparence



- Saisissez le titre Clients imprévisibles comme % des clients prévisibles.
- Exécutez le noeud.

Figure 26-34
Nuage des clients imprévisibles

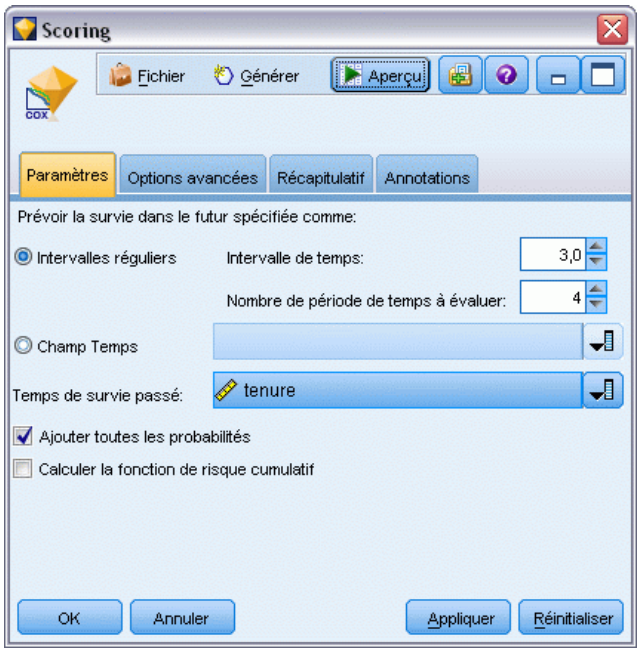


Pendant la première année, le pourcentage des clients imprévisibles augmente assez régulièrement, mais explose pendant la deuxième année jusqu'à ce que, le 23ème mois, le nombre de clients avec des valeurs nulles dépasse le nombre prévu de clients retenus.

Évaluation

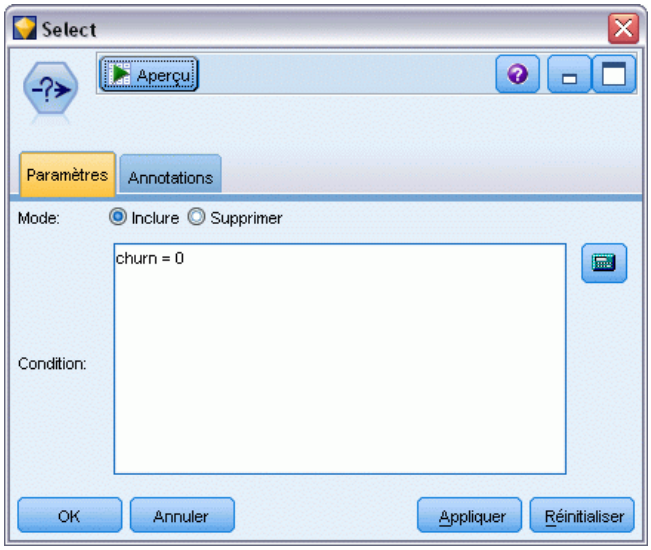
Lorsque votre modèle vous convient, évaluez les clients afin d'identifier les individus les plus susceptibles d'attrition l'année suivante, par trimestre.

Figure 26-35
Nugget Coxreg : onglet Paramètres



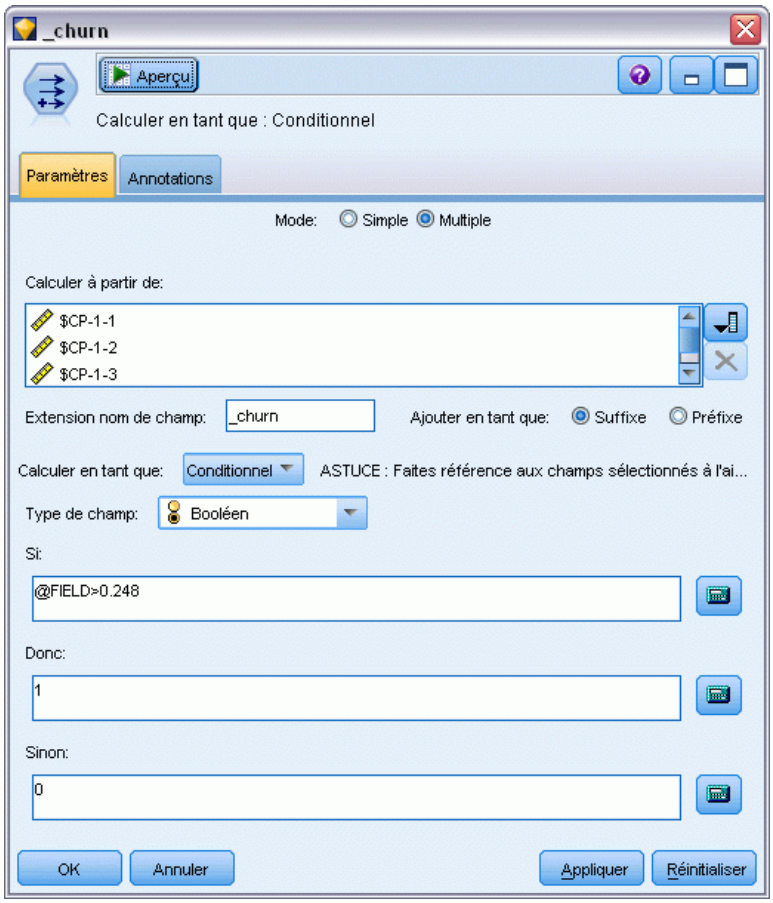
- Joignez un troisième nugget de modèle au noeud Source et ouvrez le nugget de modèle.
- Vérifiez que Intervalles réguliers est sélectionné et spécifiez 3.0 comme intervalle de temps et 4 comme nombre de périodes à évaluer. Chaque enregistrement sera ainsi évalué pendant les quatre trimestres suivants.
- Sélectionnez le champ *durée d'affectation* pour spécifier la durée de survie passée. L'algorithme d'évaluation prendra en compte la durée de survie de chaque client comme client de l'entreprise.
- Sélectionnez Ajouter toutes les probabilités. Ces champs supplémentaires faciliteront le tri des enregistrements et leur affichage dans un tableau.

Figure 26-36
Noeud Sélectionner : onglet Paramètres



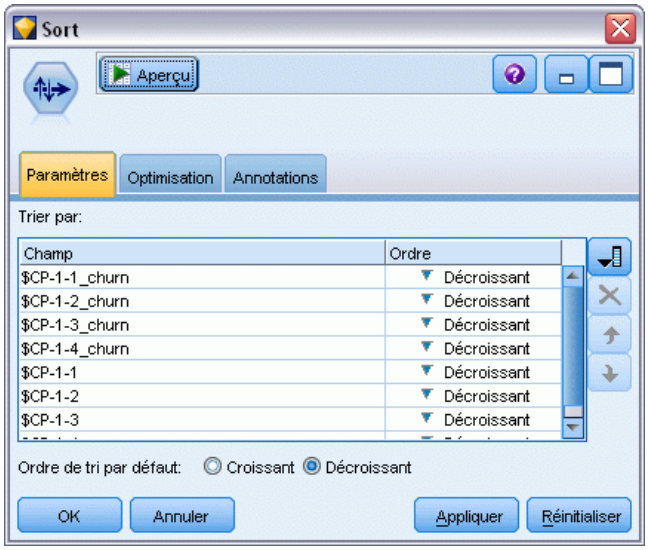
- Reliez un noeud Sélectionner au nugget de modèle ; dans l'onglet Paramètres, saisissez la condition attrition=0. Cette action supprime les clients déjà perdus du tableau de résultats.

Figure 26-37
Noeud Calculer : onglet Paramètres



- Reliez un noeud Calculer au noeud Sélectionner ; dans l'onglet Paramètres, sélectionnez le mode Multiple.
- Choisissez de calculer les champs *$CP-1-1* à *$CP-1-4*, les champs de forme *$CP-1-n*, et saisissez le suffixe _attrition à ajouter. Pour faciliter cette action, triez les champs par nom (c'est-à-dire par ordre alphabétique) dans la boîte de dialogue Sélectionner les champs.
- Choisissez de calculer le champ comme champ Conditionnel.
- Sélectionnez Booléen comme niveau de mesure.
- Saisissez @FIELD>0.248 comme condition If (Si). Veuillez noter qu'il s'agit là de la césure de classification identifiée pendant l'évaluation.
- Saisissez 1 comme expression Then (Donc).
- Saisissez 0 comme expression Else (Sinon).
- Cliquez sur OK.

Figure 26-38
Noeud Trier : onglet Paramètres



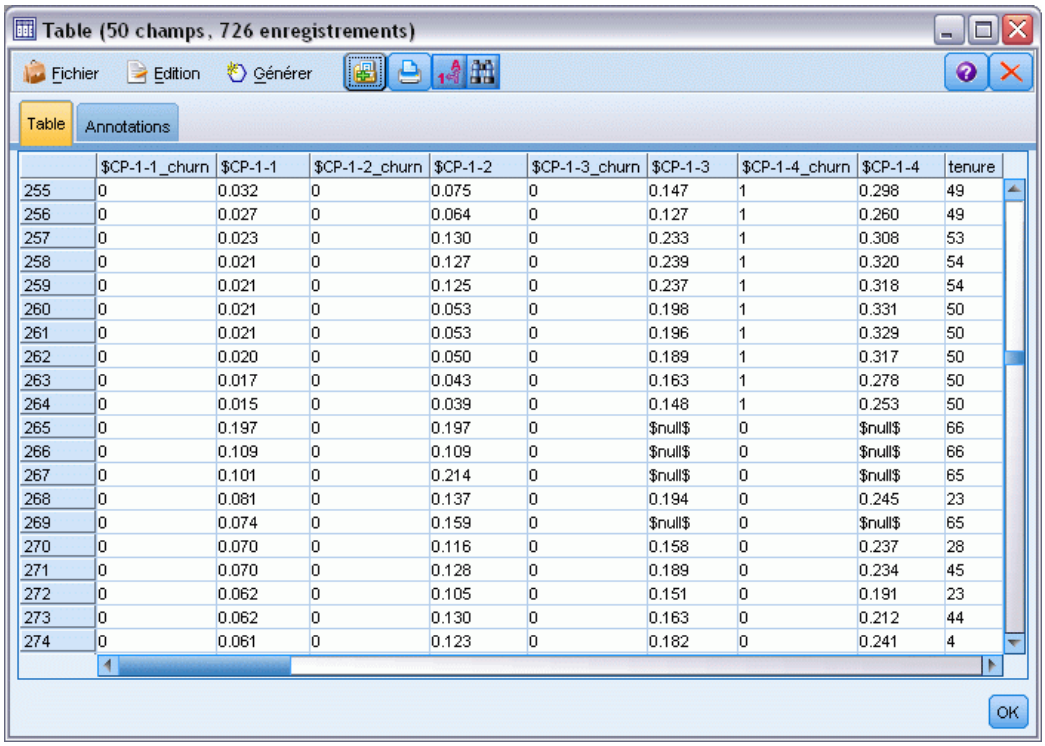
- Reliez un noeud Trier au noeud Calculer ; dans l'onglet Paramètres, choisissez de trier par *$CP-1-1_attrition* à *$CP-1-4-attrition* puis par *$CP-1-1* à *$CP-1-4*, dans l'ordre décroissant. Les clients dont l'attrition a été prévue figureront en premier.

Figure 26-39
Noeud Réorganiser : onglet Réorganiser

- Reliez un noeud Réorganiser au noeud Trier ; dans l'onglet Réorganiser, choisissez de placer les champs *$CP-1-1_attrition* à *$CP-1-4* devant les autres champs. Ceci est une option qui permet

simplement de faciliter la lecture du tableau de résultats. Utilisez les boutons pour déplacer les champs comme indiqué dans le schéma.

Figure 26-40
Tableau présentant les scores des clients

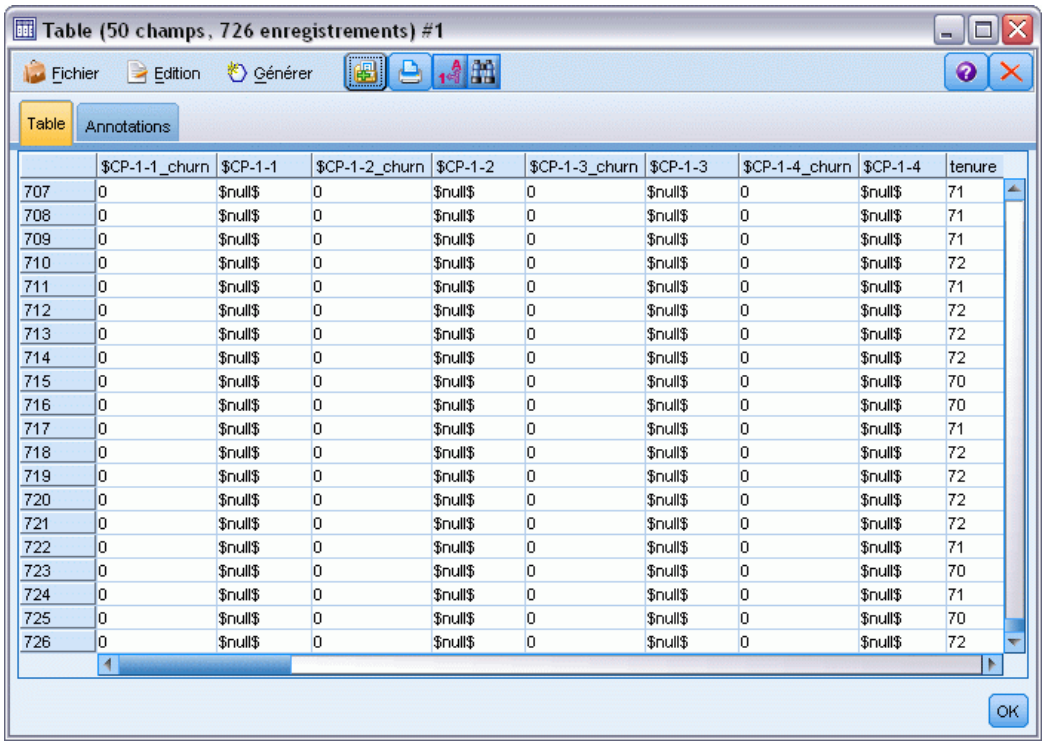


	$CP-1-1_churn	$CP-1-1	$CP-1-2_churn	$CP-1-2	$CP-1-3_churn	$CP-1-3	$CP-1-4_churn	$CP-1-4	tenure
255	0	0.032	0	0.075	0	0.147	1	0.298	49
256	0	0.027	0	0.064	0	0.127	1	0.260	49
257	0	0.023	0	0.130	0	0.233	1	0.308	53
258	0	0.021	0	0.127	0	0.239	1	0.320	54
259	0	0.021	0	0.125	0	0.237	1	0.318	54
260	0	0.021	0	0.053	0	0.198	1	0.331	50
261	0	0.021	0	0.053	0	0.196	1	0.329	50
262	0	0.020	0	0.050	0	0.189	1	0.317	50
263	0	0.017	0	0.043	0	0.163	1	0.278	50
264	0	0.015	0	0.039	0	0.148	1	0.253	50
265	0	0.197	0	0.197	0	$null$	0	$null$	66
266	0	0.109	0	0.109	0	$null$	0	$null$	66
267	0	0.101	0	0.214	0	$null$	0	$null$	65
268	0	0.081	0	0.137	0	0.194	0	0.245	23
269	0	0.074	0	0.159	0	$null$	0	$null$	65
270	0	0.070	0	0.116	0	0.158	0	0.237	28
271	0	0.070	0	0.128	0	0.189	0	0.234	45
272	0	0.062	0	0.105	0	0.151	0	0.191	23
273	0	0.062	0	0.130	0	0.163	0	0.212	44
274	0	0.061	0	0.123	0	0.182	0	0.241	4

▶ Reliez un noeud Table au noeud Réorganiser et exécutez-le.

L'attrition de 264 clients est prévue pour la fin de l'année, 184 à la fin du troisième trimestre, 103 à la fin du deuxième et 31 à la fin du premier. Veuillez noter que sur deux clients, celui avec la plus forte propension à l'attrition pendant le premier trimestre n'a pas nécessairement la plus forte propension à l'attrition pendant les trimestres suivants ; examinez par exemple les enregistrements 256 et 260. Ceci est probablement dû à la forme de la fonction des risques des mois suivant la durée d'affectation actuelle du client ; par exemple, les clients qui ont rejoint l'entreprise en raison d'une promotion sont plus susceptibles de partir plus rapidement que les clients ayant rejoint l'entreprise sur recommandation personnelle, mais s'ils ne partent pas, ils peuvent alors être plus fidèles pour leur durée d'affectation restante. Réorganiser de nouveau les clients pour obtenir des vues différentes des clients les plus susceptibles de quitter.

Figure 26-41
Tableau présentant les clients avec des valeurs nulles



	$CP-1-1_churn	$CP-1-1	$CP-1-2_churn	$CP-1-2	$CP-1-3_churn	$CP-1-3	$CP-1-4_churn	$CP-1-4	tenure
707	0	$null$	0	$null$	0	$null$	0	$null$	71
708	0	$null$	0	$null$	0	$null$	0	$null$	71
709	0	$null$	0	$null$	0	$null$	0	$null$	71
710	0	$null$	0	$null$	0	$null$	0	$null$	72
711	0	$null$	0	$null$	0	$null$	0	$null$	71
712	0	$null$	0	$null$	0	$null$	0	$null$	72
713	0	$null$	0	$null$	0	$null$	0	$null$	72
714	0	$null$	0	$null$	0	$null$	0	$null$	72
715	0	$null$	0	$null$	0	$null$	0	$null$	70
716	0	$null$	0	$null$	0	$null$	0	$null$	70
717	0	$null$	0	$null$	0	$null$	0	$null$	71
718	0	$null$	0	$null$	0	$null$	0	$null$	72
719	0	$null$	0	$null$	0	$null$	0	$null$	72
720	0	$null$	0	$null$	0	$null$	0	$null$	72
721	0	$null$	0	$null$	0	$null$	0	$null$	72
722	0	$null$	0	$null$	0	$null$	0	$null$	71
723	0	$null$	0	$null$	0	$null$	0	$null$	70
724	0	$null$	0	$null$	0	$null$	0	$null$	71
725	0	$null$	0	$null$	0	$null$	0	$null$	70
726	0	$null$	0	$null$	0	$null$	0	$null$	72

Les clients avec des valeurs nulles prédites se trouvent au bas de la table. Il s'agit de clients dont la durée d'affectation totale (temps futur + *durée d'affectation*) se trouve en-dehors de la plage des durées de survie des données utilisées pour former le modèle.

Récapitulatif

L'utilisation de la régression de Cox vous a permis de trouver un modèle approprié pour la durée jusqu'à l'attrition, de représenter le nombre prévu de clients retenus pendant les deux prochaines années et d'identifier les clients individuels les plus susceptibles de quitter au cours de l'année suivante. Remarque : même si ce modèle semble acceptable, il n'est peut-être pas le meilleur modèle. Idéalement, vous devriez comparer ce modèle, obtenu à partir de la méthode Pas à pas ascendante, à un modèle qui utilise la méthode Pas à pas descendante.

Vous trouverez des explications sur le fondement mathématique des méthodes de modélisation utilisées dans IBM® SPSS® Modeler dans le *Guide des Algorithmes SPSS Modeler*.

Chapitre
27

Analyse d'un panier de courses (Induction de règle/C5.0)

Cet exemple se base sur des données fictives décrivant le contenu d'un panier à provisions (c'est-à-dire, un ensemble d'articles achetés en même temps) et sur les données personnelles de l'acheteur, lesquelles peuvent être collectées via un programme de fidélité. L'objectif est d'identifier des ensembles de consommateurs effectuant des achats similaires et pouvant être regroupés selon des caractéristiques démographiques telles que l'âge, les revenus, etc.

Cet exemple illustre deux phases du processus de Data mining :

- La modélisation de règles d'association et l'affichage des relations mettent en évidence les liens entre les articles achetés.
- L'induction d'une règle C5.0 permet d'établir un portrait des acheteurs des groupes de produits identifiés.

Remarque : Cette application ne fait pas directement appel à la modélisation prédictive, c'est pourquoi la précision des modèles générés n'est pas mesurée et la distinction apprentissage/test n'est pas effectuée au cours du processus de Data mining.

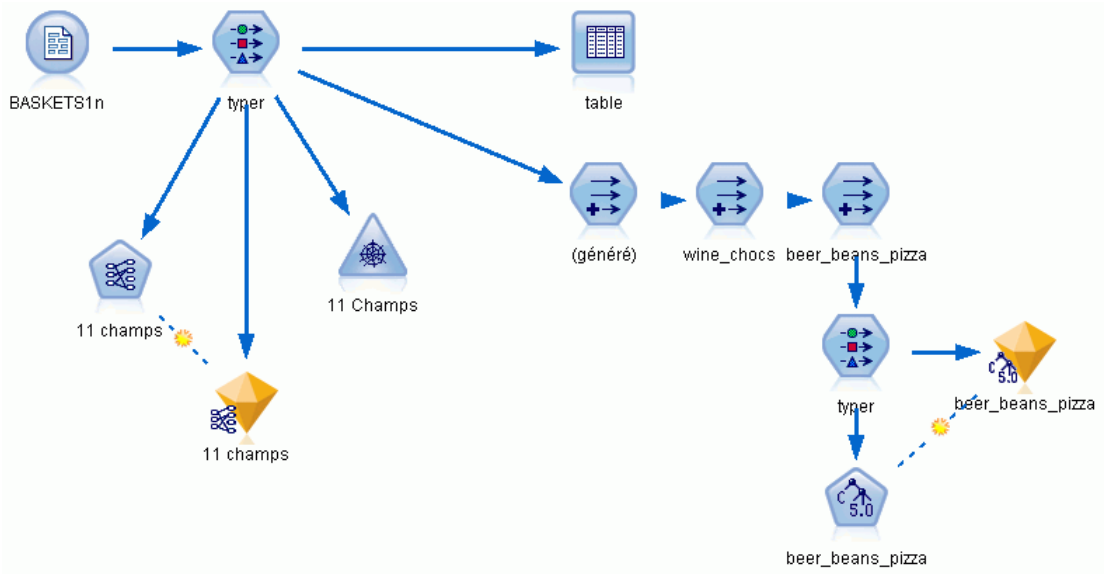
Cet exemple utilise le flux nommé *baskrule* qui fait référence au fichier de données *BASKETS1n*. Ces fichiers sont disponibles dans le répertoire *Demos* de n'importe quelle installation de IBM® SPSS® Modeler. Ce dossier est accessible à partir du groupe de programmes IBM® SPSS® Modeler dans le menu Démarrer de Windows. Le fichier *baskrule* se trouve dans le répertoire des *flux*.

Accès aux données

Connectez un noeud Délimité à l'ensemble de données *Panier* et choisissez de lire le nom des champs à partir du fichier. Connectez un nœud Typer à la source de données, puis connectez-le à un nœud Table. Définissez le niveau de mesure du champ *No carte* sur *Sans type* (un numéro de carte de fidélité n'apparaissant qu'une fois dans l'ensemble de données, son utilisation ne présente pas d'intérêt particulier pour la modélisation). Sélectionnez *Nominal* comme niveau de

mesure du champ *sexe* (afin que l'algorithme de modélisation Apriori ne considère pas le champ *sexe* comme un booléen).

Figure 27-1
Flux baskrule (panier)



Exécutez le flux pour instancier le nœud Typer et affichez le tableau. L'ensemble de données contient 18 champs, chacun représentant un panier.

Les 18 champs sont présentés sous les en-têtes suivants.

Récapitulatif du panier :

- *No carte*. Numéro de carte de fidélité de l'acheteur.
- *montant*. Prix total des articles du panier.
- *paiement*. Méthode de paiement.

Informations personnelles sur le détenteur de la carte :

- *sexe*
- *locataire*. Indique si le détenteur de la carte est locataire.
- *revenu*
- *âge*

Catégories de produits contenues dans le panier :

- *fruits & légumes*
- *boucherie*
- *produits laitiers*
- *conserves légumes*
- *conserves viande*

- *frozenmeal*
- *beer*
- *vin*
- *softdrink*
- *fish*
- *confiseries*

Identification des analogies entre les articles du panier

Vous devez tout d'abord obtenir un aperçu des analogies (associations) existant entre les articles du panier. Pour ce faire, utilisez un algorithme Apriori afin de produire des règles d'association. Sélectionnez les champs à utiliser au cours du processus de modélisation en définissant, dans le noeud Typer, le rôle de toutes les catégories de produits sur *Les deux* et toutes les autres rôles sur *Aucun*. (*Les deux* signifie que le champ peut figurer en tant qu'entrée ou en tant que sortie du modèle résultant.)

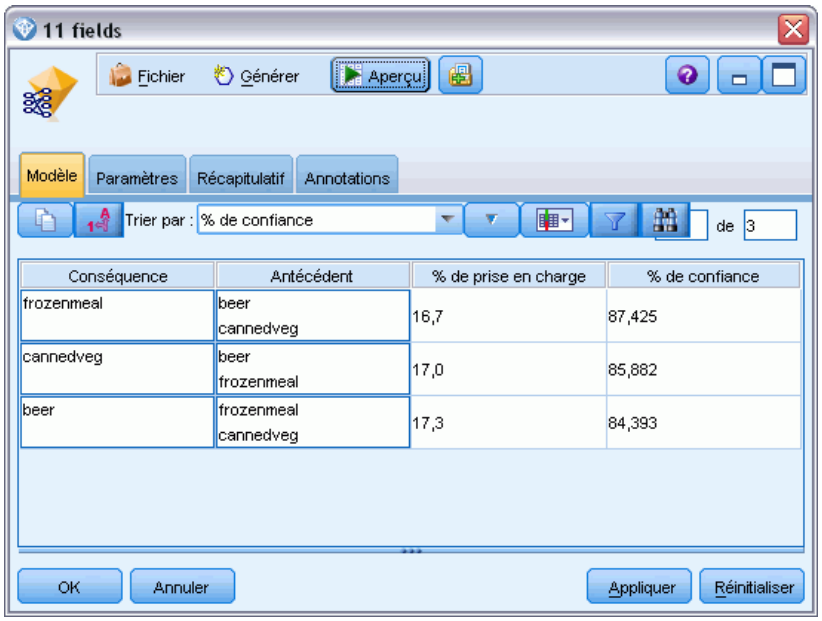
Remarque : vous pouvez définir les options de plusieurs champs. Pour ce faire, maintenez la touche Maj enfoncée tout en sélectionnant les champs, puis spécifiez une option à partir des colonnes.

Figure 27-2
Sélection des champs pour la modélisation

Une fois les champs destinés à la modélisation spécifiés, connectez le noeud Apriori au noeud Typer, modifiez-le, sélectionnez l'option Uniquement valeurs vraies pour booléens, puis exécutez le noeud Apriori. Un modèle apparaît sur l'onglet Modèles situé dans la partie supérieure droite de

la fenêtre des gestionnaires. Il contient des règles d'association que vous pouvez consulter en utilisant le menu contextuel et l'option Parcourir.

Figure 27-3
Règles d'association



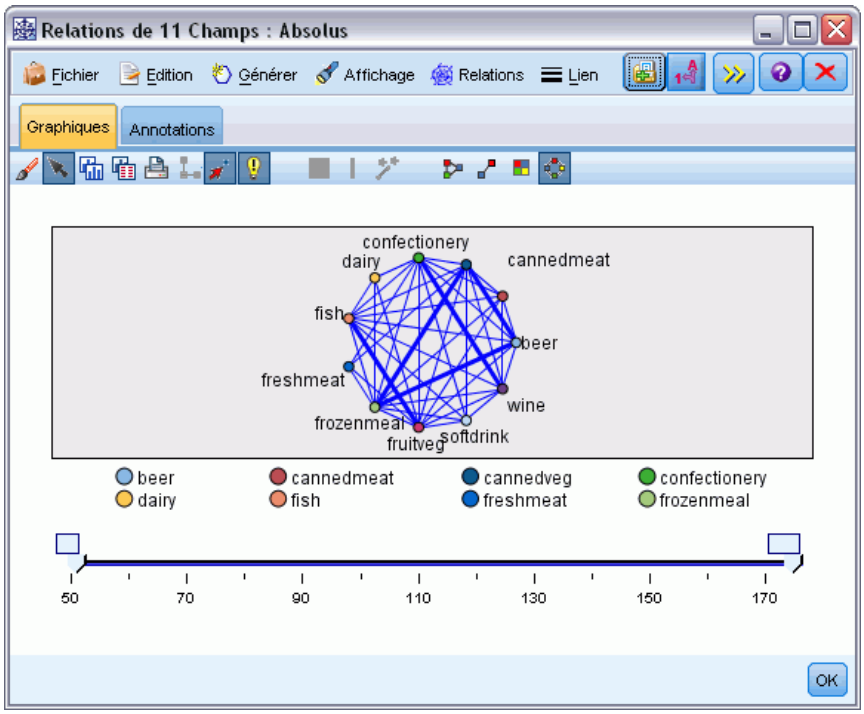
Ces règles montrent différentes associations entre les produits surgelés, les légumes en conserve et la bière. La présence de règles d'association d'ordre 2 du type :

surgelés -> bière
bière - >surgelés

indique que l'affichage des relations (qui ne présente que ce type d'associations) pourrait permettre de dégager certaines tendances à partir de ces données.

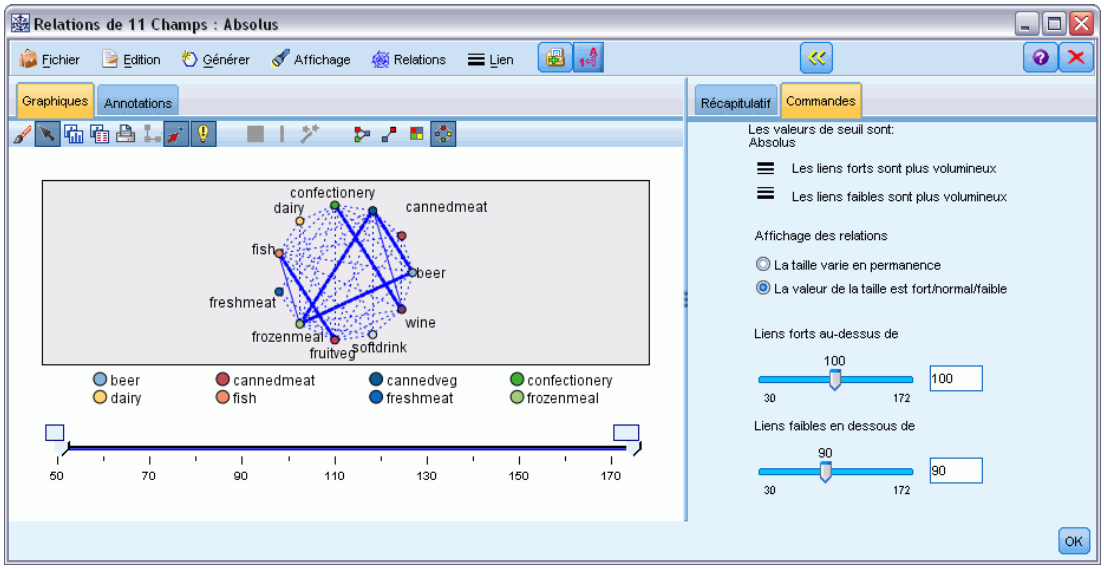
Connectez un nœud Relations au nœud Typer, éditez le nœud Relations, sélectionnez tous les champs correspondant au contenu du panier, cochez la case Afficher uniquement les booléens ayant une valeur vraie et exécutez le nœud Relations.

Figure 27-4
Affichage des relations des associations entre les produits



La plupart des combinaisons de catégories de produits étant présentes dans plusieurs paniers, les liens forts figurant dans cette relation sont trop nombreux pour permettre de repérer les groupes d'acheteurs identifiés par le modèle.

Figure 27-5
Affichage des relations restreint



- Pour spécifier les connexions faibles et les connexions fortes, cliquez sur la double flèche jaune de la barre d'outils. Ce bouton permet d'agrandir la boîte de dialogue, et d'afficher le récapitulatif et les commandes de sortie de la relation.
- Sélectionnez La valeur de la taille est fort/normal/faible.
- Définissez les liens faibles au-dessous de 90.
- Définissez les liens forts au-dessus de 100.

Cet affichage généré met en évidence les groupes d'acheteurs suivants :

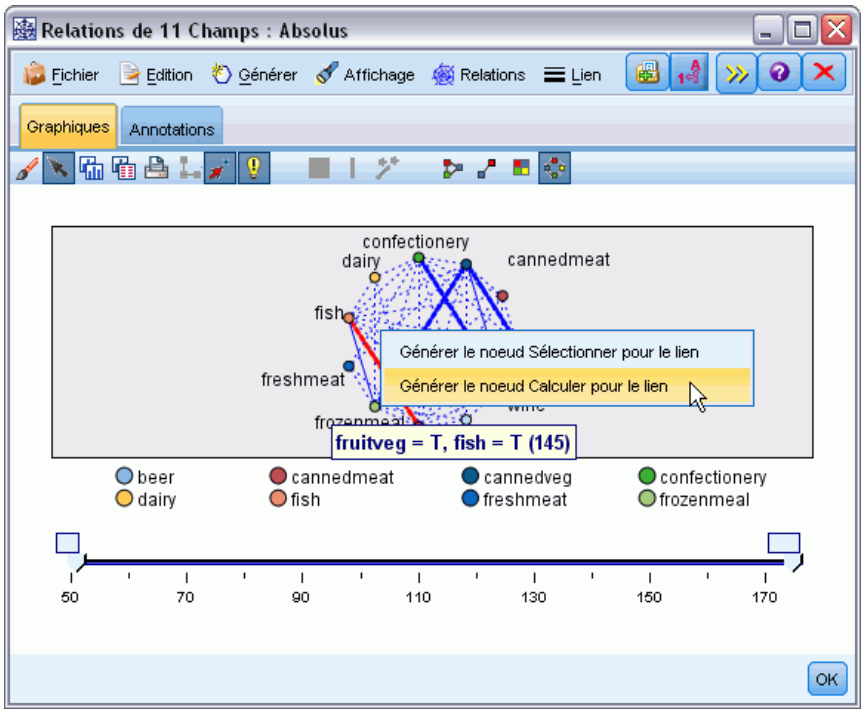
- Ceux qui achètent du poisson, des fruits et légumes (le groupe « santé » dans notre exemple).
- Ceux qui achètent du vin et des confiseries
- Ceux qui achètent de la bière, des plats surgelés et des légumes en conserve (« bière, petits pois et pizza »)

Portrait des groupes d'acheteurs

Vous avez maintenant mis en évidence trois types de consommateur, regroupés en fonction des produits qu'ils achètent. Vous allez à présent les identifier plus en détail, en établissant leur profil démographique. Pour ce faire, vous pouvez associer chacun d'eux à un booléen correspondant au groupe auquel il appartient, puis utiliser une règle C5.0 pour définir le profil de ces booléens.

Vous devez tout d'abord calculer un booléen pour chaque groupe. Il peut être généré automatiquement en utilisant l'affichage des relations que vous venez de créer. A l'aide du bouton droit de la souris, cliquez sur le lien qui relie *fruits & légumes* et *poisson* pour le mettre en évidence; puis cliquez avec le bouton droit de la souris et sélectionnez Générer le nœud Calculer pour le lien.

Figure 27-6
Calcul d'un booléen pour chaque groupe d'acheteurs



Dans le nœud Calculer généré, éditez le nom du champ et choisissez *santé*. Répétez l'opération avec le lien reliant *vin* à *confiseries*, et nommez le champ Calculer résultant *vin_choco*.

Pour le troisième groupe (impliquant trois liens), assurez-vous d'abord qu'aucun lien n'est sélectionné. Cliquez ensuite sur tous les liens du triangle *conserves légumes*, *bière* et *surgelés* pour les sélectionner tout en maintenant la touche MAJ enfoncée. (Assurez-vous d'être en mode interactif et non en mode d'édition.) Ensuite, choisissez les options de menu suivantes à partir de l'affichage des relations :

Générer > Noeud Calculer (Et)

Modifiez le nom du champ Calculer résultant en *bière_pizza_petits_pois*.

Pour établir le profil de vos groupes d'acheteurs, connectez le noeud Typer existant à ces trois noeuds Calculer en série, puis connectez un autre noeud Typer. Dans le nouveau nœud Typer, définissez le rôle *Aucun* pour tous les champs, sauf pour *montant*, *paiement*, *sexe*, *locataire*, *revenu* et *âge*, qui doivent être définis sur la valeur *Entrée*, et le groupe d'acheteurs défini (par exemple, *bière_pizza_petits_pois*), qui doit être défini sur la valeur *Cible*. Connectez un nœud C5.0,

définissez le type de sortie sur Ensemble de règles, puis exécutez le nœud. Le modèle généré (pour *bière_pizza_petits_pois*) contient un profil démographique clair pour ce groupe d'acheteurs :

```
Rule 1 for T:
if sex = M
and income <= 16,900
then T
```

Pour appliquer la même méthode aux autres boléens des groupes d'acheteurs, sélectionnez-les comme sortie pour le second nœud Typer. D'autres profils peuvent être générés si vous utilisez Apriori au lieu de C5.0 dans ce contexte. Apriori permet également d'établir simultanément le profil de tous les booléens du groupe de clients, car il n'est pas limité à un seul champ de sortie.

Récapitulatif

Cet exemple illustre la manière dont IBM® SPSS® Modeler peut être utilisé pour mettre en évidence des associations, ou des liens, entre les éléments d'une base de données, par modélisation (à l'aide d'Apriori) et par visualisation (à l'aide de l'affichage des relations). Ces liens correspondent à des regroupements d'observations effectuées dans les données ; les groupes identifiés peuvent être analysés en détail et leur profil peut être établi par modélisation (à l'aide d'ensembles de règles C5.0).

Dans le domaine de la vente au détail, ces groupes peuvent permettre, par exemple, de cibler des offres spéciales afin d'obtenir de meilleurs résultats en termes de réponse au publipostage direct, ou de personnaliser la gamme de produits stockés par un magasin afin de répondre aux besoins de la clientèle, identifiée en fonction de caractéristiques démographiques.

Chapitre
28

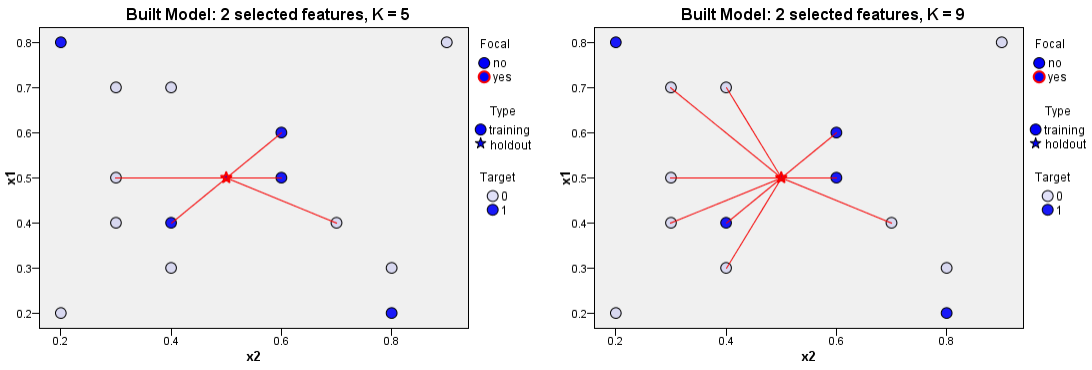
Estimation des offres de nouveaux véhicules (KNN)

L'analyse d'agrégation suivant le saut minimum (ou du plus proche voisin) est une méthode de classification des observations basée sur la similarité des observations entre elles. Dans le domaine de l'apprentissage automatique, elle a été développée comme un moyen de reconnaître des patrons de données sans nécessiter une correspondance exacte à une observation ou à un patron enregistré. Les observations semblables sont proches les unes des autres et les observations dissemblables sont éloignées les unes des autres. Ainsi la distance entre deux observations est une mesure de leur dissimilarité.

Les observations proches les unes des autres sont appelées « voisins » . Lorsqu'une nouvelle observation (de rétention) est présentée, sa distance de chaque observation du modèle est calculée. Les classifications des observations les plus similaires « les plus proches voisins » sont mesurées et la nouvelle observation est placée dans la catégorie qui contient le plus grand nombre de voisins les plus proches.

Vous pouvez spécifier le nombre de voisins les plus proches à examiner, cette valeur est appelée *k*. Les images montrent comment une nouvelle observation est classifiée en utilisant deux valeurs différentes de *k*. Lorsque *k* = 5, la nouvelle observation est placée dans la catégorie *1* car une majorité de voisins les plus proches appartiennent à la catégorie *1*. Toutefois, lorsque *k* = 9, la nouvelle observation est placée dans la catégorie *0* car une majorité de voisins les plus proches appartiennent à la catégorie *0*.

Figure 28-1
Les effets de k sur la classification



L'analyse d'agrégation suivant le saut minimum peut aussi être utilisée pour calculer les valeurs d'une cible continue. Dans cette situation, la valeur cible moyenne ou médiane des voisins les plus proches est utilisée pour obtenir la valeur prédite de la nouvelle observation.

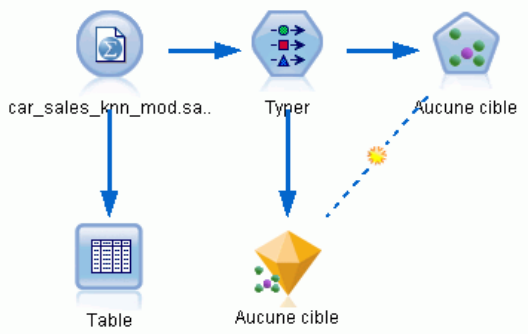
Un constructeur automobile a développé des prototypes pour deux nouveaux véhicules, une voiture et un camion. Avant de présenter les nouveaux modèles dans sa gamme, le constructeur souhaite déterminer les véhicules existant sur le marché qui sont les plus semblables aux prototypes (c'est-à-dire, les véhicules qui sont « les plus proches voisins ») et ainsi déterminer les modèles avec lesquels ils seront en concurrence.

Le constructeur a collecté des données concernant les modèles existants dans plusieurs catégories et a ajouté les détails de ses prototypes. Les catégories dans lesquelles les modèles doivent être comparés comprennent le prix en milliers (*price*), la taille du moteur (*engine_s*), la puissance en chevaux (*horsepow*), l'empattement (*wheelbas*), la largeur (*width*), la longueur (*length*), le poids total (*curb_wgt*), la capacité du réservoir (*fuel_cap*) et le rendement énergétique (*mpg*).

Cet exemple utilise le flux nommé *car_sales_knn.str*, disponible dans le dossier *Demos* du sous-dossier des *flux*. Le fichier de données est *car_sales_knn_mod.sav*. Pour plus d'informations, reportez-vous à la section Dossier Demos dans le chapitre 1 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.

Création du flux

Figure 28-2
Exemple de flux de modélisation KNN

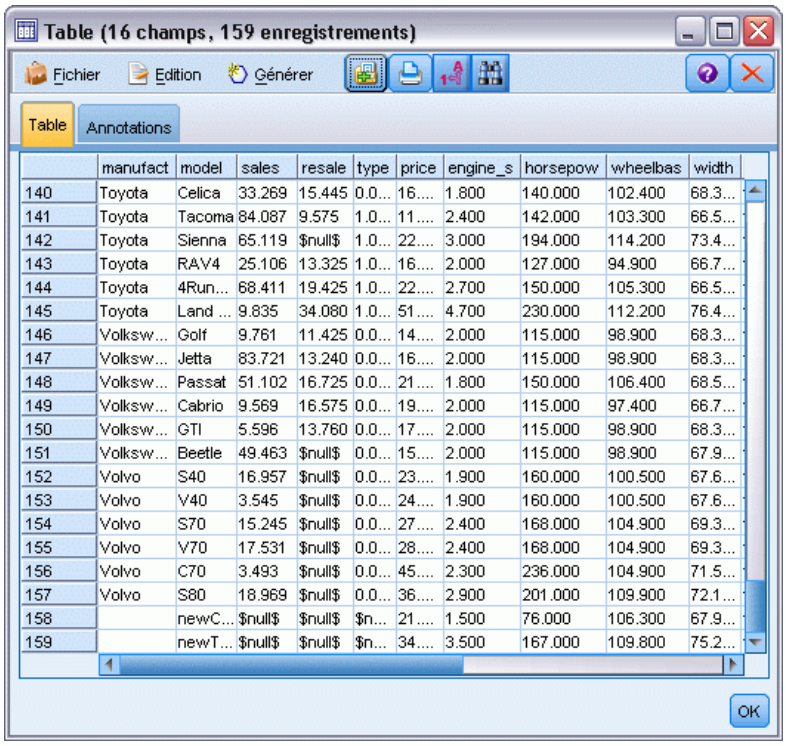


Créez un nouveau flux et ajoutez un noeud source Fichier de statistiques pointant vers *car_sales_knn_mod.sav* dans le dossier *Demos* de votre installation de IBM® SPSS® Modeler.

En premier lieu, examinons les données collectées par le constructeur.

- Reliez un noeud Table au noeud source Fichier de statistiques.
- Ouvrez le noeud Table, puis cliquez sur Exécuter.

Figure 28-3
Données source pour les automobiles et les camions



	manufact	model	sales	resale	type	price	engine_s	horsepow	wheelbas	width
140	Toyota	Celica	33.269	15.445	0.0...	16....	1.800	140.000	102.400	68.3...
141	Toyota	Tacoma	84.087	9.575	1.0...	11....	2.400	142.000	103.300	66.5...
142	Toyota	Sienna	65.119	$null$	1.0...	22....	3.000	194.000	114.200	73.4...
143	Toyota	RAV4	25.106	13.325	1.0...	16....	2.000	127.000	94.900	66.7...
144	Toyota	4Run...	68.411	19.425	1.0...	22....	2.700	150.000	105.300	66.5...
145	Toyota	Land ...	9.835	34.080	1.0...	51....	4.700	230.000	112.200	76.4...
146	Volksw...	Golf	9.761	11.425	0.0...	14....	2.000	115.000	98.900	68.3...
147	Volksw...	Jetta	83.721	13.240	0.0...	16....	2.000	115.000	98.900	68.3...
148	Volksw...	Passat	51.102	16.725	0.0...	21....	1.800	150.000	106.400	68.5...
149	Volksw...	Cabrio	9.569	16.575	0.0...	19....	2.000	115.000	97.400	66.7...
150	Volksw...	GTI	5.596	13.760	0.0...	17....	2.000	115.000	98.900	68.3...
151	Volksw...	Beetle	49.463	$null$	0.0...	15....	2.000	115.000	98.900	67.9...
152	Volvo	S40	16.957	$null$	0.0...	23....	1.900	160.000	100.500	67.6...
153	Volvo	V40	3.545	$null$	0.0...	24....	1.900	160.000	100.500	67.6...
154	Volvo	S70	15.245	$null$	0.0...	27....	2.400	168.000	104.900	69.3...
155	Volvo	V70	17.531	$null$	0.0...	28....	2.400	168.000	104.900	69.3...
156	Volvo	C70	3.493	$null$	0.0...	45....	2.300	236.000	104.900	71.5...
157	Volvo	S80	18.969	$null$	0.0...	36....	2.900	201.000	109.900	72.1...
158		newC...	$null$	$null$	$n...	21....	1.500	76.000	106.300	67.9...
159		newT...	$null$	$null$	$n...	34....	3.500	167.000	109.800	75.2...

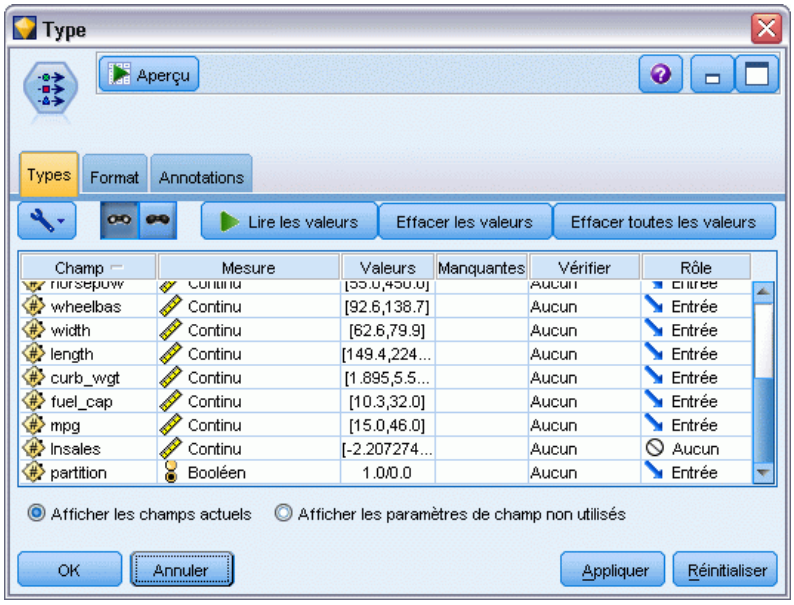
Les détails des deux prototypes, nommés *newCar* et *newTruck*, ont été ajoutés à la fin du fichier.

Nous pouvons voir à partir des données source que le constructeur utilise la classification « camion » (valeur 1 dans la colonne *type*) de manière très globale pour signifier tout type de véhicule non automobile.

La dernière colonne, *partition*, est nécessaire car les deux prototypes peuvent être désignés comme des ensembles de rétention lorsque nous sommes amenés à identifier leurs voisins les plus proches. De cette manière, leur données n'influencent pas les calculs car il s'agit du reste du marché que nous souhaitons prendre en considération. Le réglage de la valeur de la *partition* des deux enregistrements de rétention sur 1, alors que tous les autres enregistrements ont une valeur de 0 dans ce champ, nous permettra d'utiliser ce champ par la suite, lorsque nous réglerons les enregistrements centraux (les enregistrements pour lesquels nous souhaitons calculer les voisins les plus proches).

Laissons la fenêtre des sorties du tableau ouverte pour le moment, car nous la consulterons par la suite.

Figure 28-4
Paramètres du noeud Typer

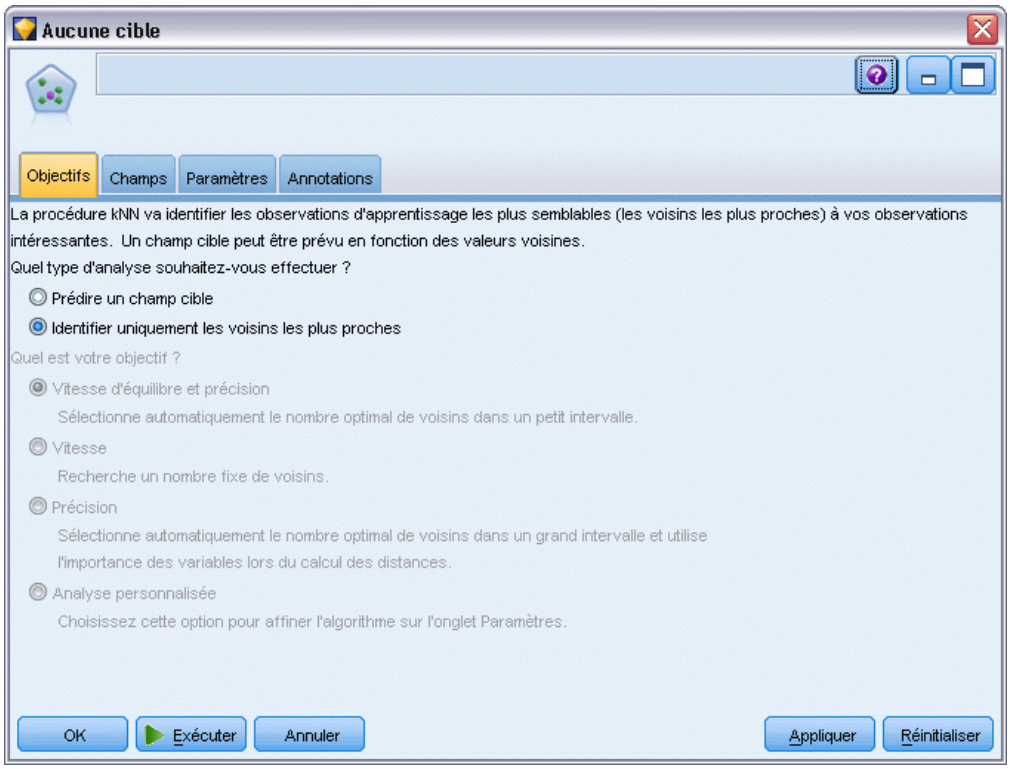


- Ajoutez un noeud Typer au flux.
- Reliez un noeud Typer au noeud source Fichier de statistiques.
- Ouvrez le noeud Typer.

  Nous souhaitons effectuer la comparaison uniquement sur les champs *price* à *mpg*, aussi laissons-nous le rôle de tous ces champs configuré sur Entrée.

- Configurez le rôle de tous les autres champs (*manufact* à *type*, plus *lnsales*) sur Aucun.
- Configurez le niveau de mesure du dernier champ, *partition*, sur Booléen. Vérifiez que son rôle est configuré sur Entrée.
- Cliquez sur Lire les valeurs pour lire les valeurs de données dans le flux.
- Cliquez sur OK.

Figure 28-5
Choisir d'identifier les voisins les plus proches

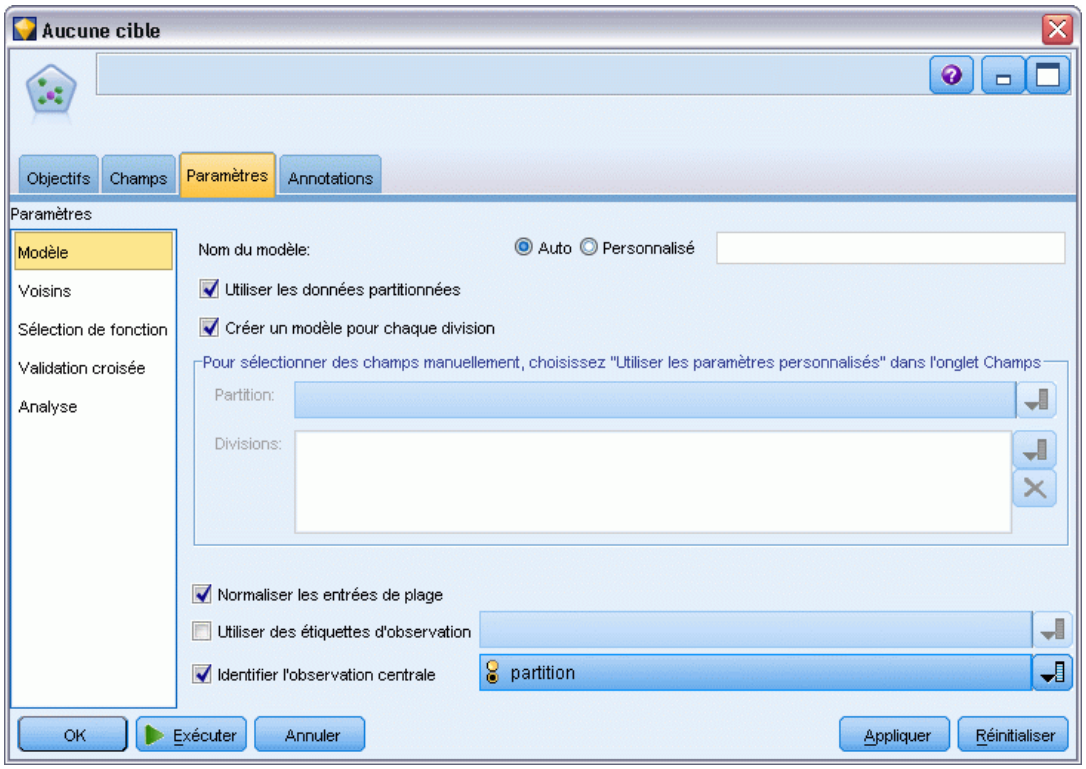


- Reliez un noeud KNN au noeud Typer.
- Ouvrez le noeud KNN.

  Cette fois-ci, nous n'allons pas prédire un champ cible, car nous souhaitons seulement trouver les voisins les plus proches de nos deux prototypes.

- Dans l'onglet Objectifs, sélectionnez Identifier uniquement les voisins les plus proches.
- Cliquez sur l'onglet Paramètres.

Figure 28-6
Utilisation du champ de partition pour identifier les enregistrements centraux



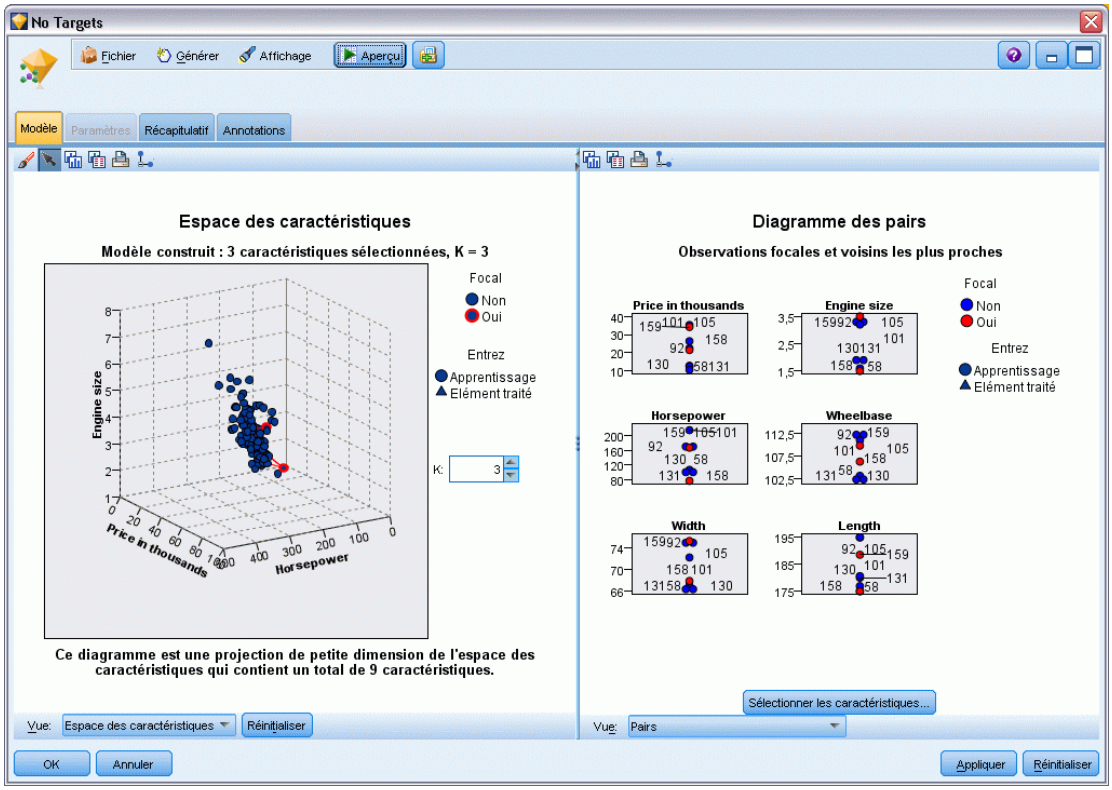
Nous pouvons maintenant utiliser le champ *partition* pour identifier les enregistrements centraux (les enregistrements pour lesquelles nous souhaitons identifier les voisins les plus proches). En utilisant un champ booléen, nous nous assurons que les enregistrements dont la valeur de ce champ est configurée sur 1 deviennent nos enregistrements centraux.

Comme nous l'avons vu, les seuls enregistrements qui possèdent une valeur de 1 pour ce champ sont *newCar* et *newTruck*, aussi ces derniers seront-ils nos enregistrements centraux.

- Dans le panneau Modèle de l'onglet Paramètres, cochez la case Identifier un enregistrement central.
- Dans la liste déroulante de ce champ, sélectionnez partition.
- Cliquez sur le bouton Exécuter.

Examen des sorties

Figure 28-7
La fenêtre Visualiseur de modèles



Un nugget de modèle a été créé dans l'espace de travail du flux et dans la palette Modèles. Ouvrez l'un des nuggets pour afficher le Visualiseur de modèles qui dispose d'une fenêtre à deux panneaux :

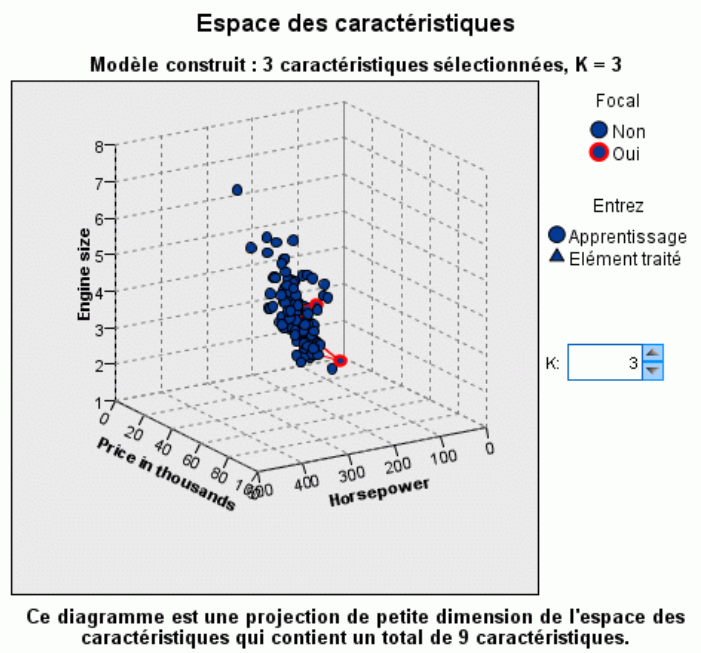
- Le premier affiche une présentation du modèle, appelée vue principale. La vue principale du modèle Voisin le plus proche est aussi appelée **espace du variable indépendante**.
- Le second affiche un des deux types de vues :

  Une vue de modèle auxiliaire affiche davantage d'informations sur le modèle, mais n'est pas focalisée sur le modèle lui-même.

  Un vue liée est un affichage montrant les détails d'une caractéristique du modèle lorsque vous faites défiler une partie de la vue principale.

Espace du variable indépendante

Figure 28-8
Diagramme de l'espace du variable indépendante



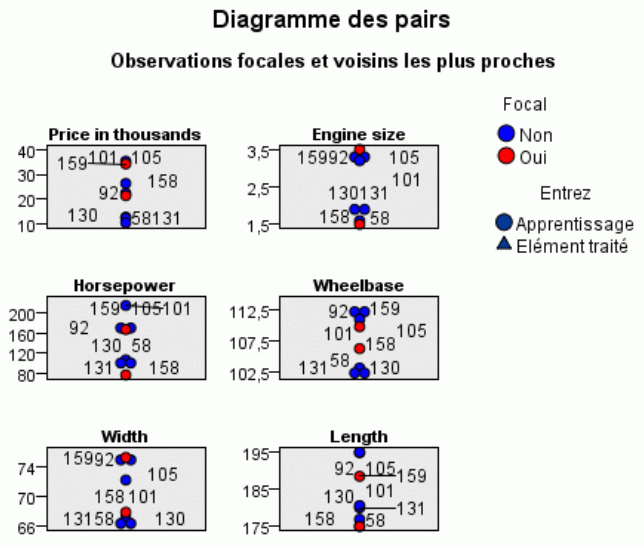
Le diagramme de l'espace du variable indépendante est un diagramme 3D interactif qui représente les points des données pour les trois caractéristiques (les trois premiers champs d'entrée des données source) représentant le prix, la taille du moteur et la puissance.

Nos deux enregistrements centraux sont mis en surbrillance en rouge, avec des lignes qui les relient à leurs *k* voisins les plus proches.

En cliquant sur le diagramme et en le faisant glisser, vous pouvez le faire pivoter et obtenir une meilleure vue de la distribution des points dans l'espace du variable indépendante. Cliquez sur le bouton Réinitialiser pour rétablir la vue par défaut.

Diagramme des pairs

Figure 28-9
Diagramme des pairs



La vue auxiliaire par défaut est le diagramme des pairs qui met en évidence les deux enregistrements centraux sélectionnés dans l'espace du variable indépendante ainsi que leurs *k* voisins les plus proches pour chacune des six caractéristiques (les six premiers champs d'entrée des données source).

Les véhicules sont représentés par leur numéro d'enregistrement dans les données source. Nous devons maintenant établir les sorties à partir du noeud Table afin de les identifier.

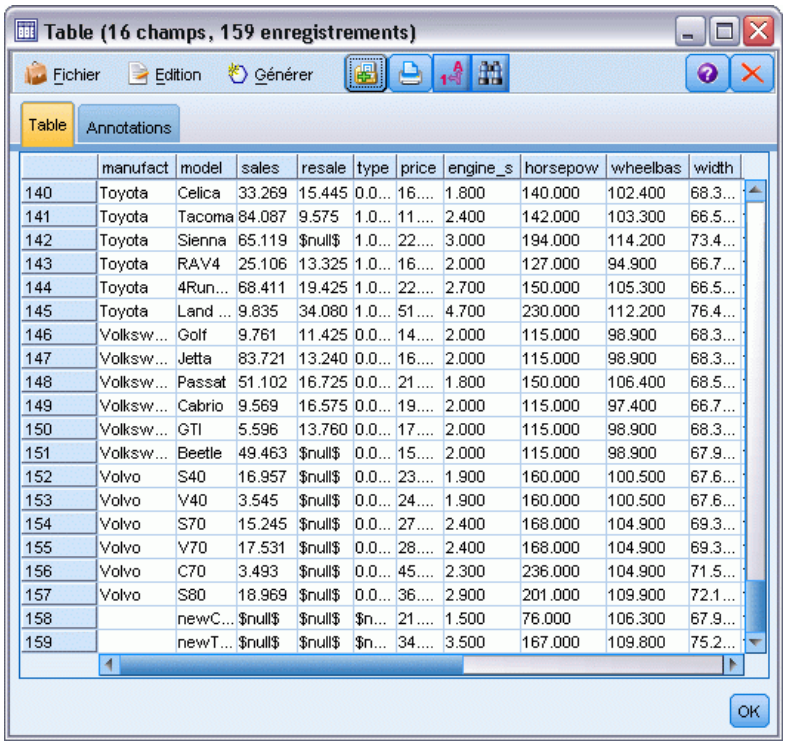
Si la sortie du noeud Table est encore disponible :

- Cliquez sur l'onglet Sorties du panneau du gestionnaire, en haut à droite de la fenêtre principale de IBM® SPSS® Modeler.
- Double-cliquez sur l'entrée Table (16 champs, 159 enregistrements).

Si la sortie de la table n'est plus disponible :

- Dans la fenêtre principale de SPSS Modeler, cliquez sur le noeud Table.
- Cliquez sur Exécuter.

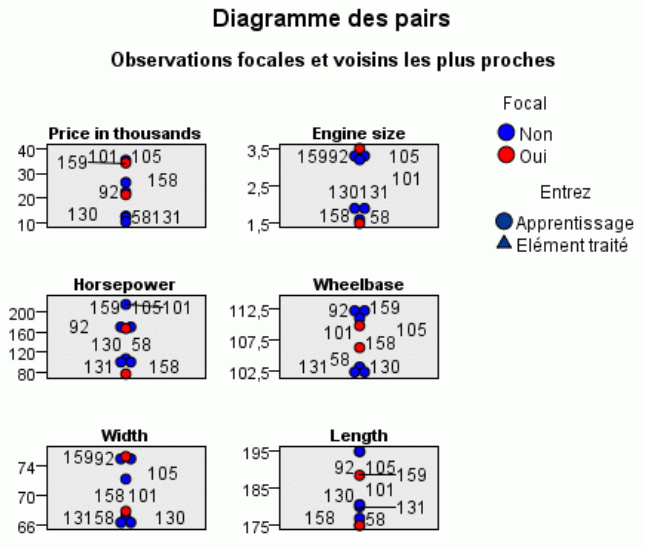
Figure 28-10
Identification des enregistrements par numéro d'enregistrement



	manufact	model	sales	resale	type	price	engine_s	horsepow	wheelbas	width
140	Toyota	Celica	33.269	15.445	0.0...	16....	1.800	140.000	102.400	68.3...
141	Toyota	Tacoma	84.087	9.575	1.0...	11....	2.400	142.000	103.300	66.5...
142	Toyota	Sienna	65.119	$null$	1.0...	22....	3.000	194.000	114.200	73.4...
143	Toyota	RAV4	25.106	13.325	1.0...	16....	2.000	127.000	94.900	66.7...
144	Toyota	4Run...	68.411	19.425	1.0...	22....	2.700	150.000	105.300	66.5...
145	Toyota	Land ...	9.835	34.080	1.0...	51....	4.700	230.000	112.200	76.4...
146	Volksw...	Golf	9.761	11.425	0.0...	14....	2.000	115.000	98.900	68.3...
147	Volksw...	Jetta	83.721	13.240	0.0...	16....	2.000	115.000	98.900	68.3...
148	Volksw...	Passat	51.102	16.725	0.0...	21....	1.800	150.000	106.400	68.5...
149	Volksw...	Cabrio	9.569	16.575	0.0...	19....	2.000	115.000	97.400	66.7...
150	Volksw...	GTI	5.596	13.760	0.0...	17....	2.000	115.000	98.900	68.3...
151	Volksw...	Beetle	49.463	$null$	0.0...	15....	2.000	115.000	98.900	67.9...
152	Volvo	S40	16.957	$null$	0.0...	23....	1.900	160.000	100.500	67.6...
153	Volvo	V40	3.545	$null$	0.0...	24....	1.900	160.000	100.500	67.6...
154	Volvo	S70	15.245	$null$	0.0...	27....	2.400	168.000	104.900	69.3...
155	Volvo	V70	17.531	$null$	0.0...	28....	2.400	168.000	104.900	69.3...
156	Volvo	C70	3.493	$null$	0.0...	45....	2.300	236.000	104.900	71.5...
157	Volvo	S80	18.969	$null$	0.0...	36....	2.900	201.000	109.900	72.1...
158		newC...	$null$	$null$	$n...	21....	1.500	76.000	106.300	67.9...
159		newT...	$null$	$null$	$n...	34....	3.500	167.000	109.800	75.2...

Si vous accédez en bas de la table, vous pouvez constater que *newCar* et *newTruck* sont les deux derniers enregistrements des données avec les numéros 158 et 159, respectivement.

Figure 28-11
Comparaison des caractéristiques dans le diagramme des pairs



À partir de ceci, nous pouvons voir dans le diagramme des pairs, par exemple, que *newTruck* (159) possède une plus grande taille de moteur que tous ses voisins les plus proches, alors que *newCar* (158) possède un moteur plus petit que tous *ses* voisins les plus proches.

Pour chacune des six caractéristiques, vous pouvez déplacer la souris sur les points individuels afin d'afficher la valeur réelle de chacune des caractéristiques de cette observation spécifique.

Mais quels sont les véhicules qui sont les voisins les plus proches de *newCar* et de *newTruck* ?

Comme le diagramme des pairs est un peu encombré, simplifions la vue.

- Cliquez sur la liste déroulante Affichage en bas du diagramme des pairs (l'entrée nommée Pairs).
- Sélectionnez Table des voisins et des distances.

Table des voisins et des distances

Figure 28-12
table des voisins et des distances

k voisins les plus proches et distances

Affiché pour les observations focales initiales

Observation focale	Voisins les plus proches			Distances les plus proches		
	1	2	3	1	2	3
158	131	130	58	0,979	0,990	1,011
159	105	92	101	0,580	0,634	0,644

Ceci est mieux. Nous pouvons maintenant voir les trois modèles dont nos deux prototypes sont les voisins les plus proches sur le marché.

Pour *newCar* (enregistrement central 158) il s'agit de la Saturn SC (131), de la Saturn SL (130) et de la Honda Civic (58).

Ceci n'est pas vraiment surprenant (les trois sont des berlines de taille moyenne, donc *newCar* correspond bien, en particulier en ce qui concerne son excellente efficacité énergétique).

Pour *newTruck* (enregistrement central 159), les voisins les plus proches seront la Nissan Quest (105), la Mercury Villager (92) et la Mercedes M-Class (101).

Comme nous l'avons vu plus tôt, il ne s'agit pas nécessairement de camions au sens propre, mais simplement de véhicules qui ne sont pas classés comme des automobiles. Si nous observons la sortie du noeud Table afin de rechercher les voisins les plus proches, nous constatons que *newTruck* est relativement cher tout en étant l'un des plus lourds de sa catégorie. Cependant, l'efficacité énergétique est une fois encore meilleure que ses plus proches rivaux, ce qui devrait jouer en sa faveur.

Récapitulatif

Nous avons vu comment vous pouvez utiliser l'analyse des voisins les plus proches pour comparer un ensemble étendu de caractéristiques d'observations à partir d'un ensemble de données particulier. Nous avons aussi calculé, pour deux enregistrements de rétention très différents, les observations qui ressemblent le plus étroitement à ces ensembles de rétention.

Annexe

A

Remarques

Ces informations ont été développées pour les produits et services offerts dans le monde.

Il est possible qu'IBM n'offre pas dans les autres pays les produits, services et fonctionnalités décrits dans ce document. Contactez votre représentant local IBM pour obtenir des informations sur les produits et services actuellement disponibles dans votre région. Toute référence à un produit, programme ou service IBM n'implique pas que les seuls les produits, programmes ou services IBM peuvent être utilisés. Tout produit, programme ou service de fonctionnalité équivalente qui ne viole pas la propriété intellectuelle IBM peut être utilisé à la place. Cependant l'utilisateur doit évaluer et vérifier l'utilisation d'un produit, programme ou service non IBM.

IBM peut posséder des brevets ou des applications de brevet en attente qui couvrent les sujets décrits dans ce document. L'octroi de ce document n'équivaut aucunement à celui d'une licence pour ces brevets. Vous pouvez envoyer par écrit des questions concernant la licence à :

IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785, États-Unis

Pour obtenir des informations de licence concernant la configuration de caractères codés sur deux octets (DBCS), veuillez contacter dans votre pays le département chargé de la propriété intellectuelle chez IBM ou envoyez vos commentaires par écrit à :

Intellectual Property Licensing, Legal and Intellectual Property Law, IBM Japan Ltd., 1623-14, Shimotsuruma, Yamato-shi, Kanagawa 242-8502 Japon.

Le paragraphe suivant ne s'applique pas au Royaume-Uni ni à aucun pays dans lequel ces dispositions sont contraires au droit local : INTERNATIONAL BUSINESS MACHINES FOURNIT CETTE PUBLICATION « EN L'ETAT » SANS GARANTIE D'AUCUNE SORTE, IMPLICITE OU EXPLICITE, Y COMPRIS, MAIS SANS ETRE LIMITE AUX GARANTIES IMPLICITES DE NON VIOLATION, DE QUALITE MARCHANDE OU D'ADAPTATION POUR UN USAGE PARTICULIER. Certains états n'autorisent pas l'exclusion de garanties explicites ou implicites lors de certaines transactions, par conséquent, il est possible que cet énoncé ne vous concerne pas.

Ces informations peuvent contenir des erreurs techniques ou des erreurs typographiques. Ces informations sont modifiées de temps en temps ; ces modifications seront intégrées aux nouvelles versions de la publication. IBM peut apporter des améliorations et/ou modifications des produits et/ou des programmes décrits dans cette publications à tout moment sans avertissement préalable.

Toute référence dans ces informations à des sites Web autres qu'IBM est fournie dans un but pratique uniquement et ne sert en aucun cas de recommandation pour ces sites Web. Le matériel contenu sur ces sites Web ne fait pas partie du matériel de ce produit IBM et l'utilisation de ces sites Web se fait à vos propres risques.

IBM peut utiliser ou distribuer les informations que vous lui fournissez, de la façon dont il le souhaite, sans encourir aucune obligation envers vous.

Les personnes disposant d'une licence pour ce programme et qui souhaitent obtenir des informations sur celui-ci pour activer : (i) l'échange d'informations entre des programmes créés de manière indépendante et d'autres programmes (notamment celui-ci) et (ii) l'utilisation mutuelle des informations qui ont été échangées, doivent contacter :

IBM Software Group, Attention: Licensing, 233 S. Wacker Dr., Chicago, IL 60606, États-Unis.

Ces informations peuvent être disponibles, soumises à des conditions générales, et dans certains cas payantes.

Le programme sous licence décrit dans ce document et toute la documentation sous licence disponible pour ce programme sont fournis par IBM en conformité avec les conditions de l'accord du client IBM, avec l'accord de licence du programme international IBM et avec tout accord équivalent entre nous.

Toutes les données sur les performances contenues dans le présent document ont été obtenues dans un environnement contrôlé. Par conséquent, les résultats obtenus dans d'autres environnements d'exploitation peuvent varier de manière significative. Certaines mesures peuvent avoir été effectuées sur des systèmes en cours de développement et il est impossible de garantir que ces mesures seront les mêmes sur les systèmes commercialisés. De plus, certaines mesures peuvent avoir été estimées par extrapolation. Les résultats réels peuvent être différents. Les utilisateurs de ce document doivent vérifier les données applicables à leur environnement spécifique.

les informations concernant les produits autres qu'IBM ont été obtenues auprès des fabricants de ces produits, leurs annonces publiques ou d'autres sources publiques disponibles. IBM n'a pas testé ces produits et ne peut confirmer l'exactitude de leurs performances, leur compatibilité ou toute autre fonctionnalité associée à des produits autres qu'IBM. Les questions sur les capacités de produits autres qu'IBM doivent être adressées aux fabricants de ces produits.

Toutes les déclarations concernant la direction ou les intentions futures d'IBM peuvent être modifiées ou retirées sans avertissement préalable et représentent uniquement des buts et des objectifs.

Ces informations contiennent des exemples de données et de rapports utilisés au cours d'opérations quotidiennes standard. Pour les illustrer le mieux possible, ces exemples contiennent des noms d'individus, d'entreprises, de marques et de produits. Tous ces noms sont fictifs et toute ressemblance avec des noms et des adresses utilisés par une entreprise réelle ne serait que pure coïncidence.

Si vous consultez la version papier de ces informations, il est possible que certaines photographies et illustrations en couleurs n'apparaissent pas.

Marques commerciales

IBM, le logo IBM, ibm.com et SPSS sont des marques commerciales d'IBM Corporation, déposées dans de nombreuses juridictions du monde entier. Une liste à jour des marques IBM est disponible sur Internet à l'adresse *http://www.ibm.com/legal/copytrade.shtml*.

Intel, le logo Intel, Intel Inside, le logo Intel Inside, Intel Centrino, le logo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium, et Pentium sont des marques commerciales ou des marques déposées de Intel Corporation ou de ses filiales aux Etats-Unis et dans d'autres pays.

Linux est une marque déposée de Linus Torvalds aux Etats-Unis et/ou dans d'autres pays.

Microsoft, Windows, Windows NT et le logo Windows sont des marques commerciales de Microsoft Corporation aux Etats-Unis et/ou dans d'autres pays.

UNIX est une marque déposée de The Open Group aux Etats-Unis et dans d'autres pays.

Java et toutes les marques et logos Java sont des marques commerciales de Sun Microsystems, Inc. aux Etats-Unis et/ou dans d'autres pays.

Les autres noms de produits et de services peuvent être des marques d'IBM ou d'autres sociétés.

Bibliographie

Asuncion, A., et D. Newman. 2007. "Référentiel d'apprentissage automatique UCI." Available at http://mlearn.ics.uci.edu/MLRepository.html.

Index