Руководство по приложениям IBM SPSS Modeler 18.1.1

Примечание

Прежде чем использовать эту информацию и продукт, описанный в ней, прочтите сведения в разделе “Уведомления” на стр. 359.

Информация о продукте

Это издание применимо к версии 18, выпуск 1, модификация 1 IBM SPSS Modeler и ко всем последующим версиям и модификациям до тех пор, пока в новых изданиях не будет указано иное.

Содержание

Глава 1. О программе IBM SPSS Modeler

IBM® SPSS Modeler - это комплект инструментов исследования данных, при помощи которого можно быстро разрабатывать прогнозные модели, использующие деловые знания и опыт, и внедрять их в деловые операции для усовершенствования процесса принятия решений. Разработанный на основе модели промышленного стандарта CRISP-DM, IBM SPSS Modeler поддерживает весь процесс исследования данных, от обработки исходных данных до получения лучших деловых результатов.

IBM SPSS Modeler предлагает ряд методов моделирования, взятых из таких областей, как обучение машин, искусственный интеллект и статистика. При помощи методов, доступных на палитре Моделирование, можно извлечь новую информацию из данных и разработать прогнозные модели. У каждого из методов есть свои сильные стороны и типы задач, для решения которых он лучше всего подходит.

SPSS Modeler можно приобрести как отдельный продукт или использовать как клиент в сочетании с SPSS Modeler Server. Кроме того, доступен ряд дополнительных возможностей, сводка которых дается в следующих разделах. Дополнительную информацию смотрите по ссылке https://www.ibm.com/analytics/us/en/technology/spss/.

Продукты IBM SPSS Modeler

В семейство продуктов IBM SPSS Modeler и связанные с этим семейством программы входят следующие продукты:

- IBM SPSS Modeler
- IBM SPSS Modeler Server
- IBM SPSS Modeler Administration Console (включено в IBM SPSS Deployment Manager)
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- Адаптеры IBM SPSS Modeler Server для IBM SPSS Collaboration and Deployment Services

IBM SPSS Modeler

SPSS Modeler - это полнофункциональная версия продукта, устанавливаемая и запускаемая на персональном компьютере. SPSS Modeler можно запустить в локальном режиме, как автономный продукт, или в распределенном режиме вместе с IBM SPSS Modeler Server, чтобы повысить производительность на больших наборах данных.

Используя SPSS Modeler, можно быстро и интуитивно строить точные прогнозные модели, не прибегая к программированию. Используя уникальный визуальный интерфейс, можно легко визуализировать процесс анализа данных. В продукт встроены расширенные функции аналитики, при поддержке которых можно обнаруживать в данных скрытые структуры и тенденции. Можно моделировать результаты и выяснять, какие факторы на них влияют, чтобы полностью использовать деловые возможности и ограничивать риски.

SPSS Modeler доступен в двух версиях: SPSS Modeler Professional и SPSS Modeler Premium. Дополнительную информацию смотрите в разделе “Выпуски IBM SPSS Modeler” на стр. 2.

IBM SPSS Modeler Server

SPSS Modeler пользуется архитектурой клиент - сервер, чтобы распределять требования ресурсоемких операций по мощным серверным программам, что повышает производительность для больших наборов данных.

SPSS Modeler Server - это отдельно лицензируемый продукт, который непрерывно работает в режиме распределенного анализа на хосте сервера совместно с одной или несколькими установками IBM SPSS Modeler. При этом SPSS Modeler Server обеспечивает высокую производительность для больших наборов данных, поскольку ресурсоемкие операции можно выполнять на сервере без скачивания данных на компьютер клиента. Кроме того, IBM SPSS Modeler Server обеспечивает поддержку для возможностей оптимизации SQL и моделирования в базе данных, что дает дополнительный выигрыш в производительности и автоматизации.

IBM SPSS Modeler Administration Console

Modeler Administration Console - это графический пользовательский интерфейс для управления многочисленными опциями конфигурации SPSS Modeler Server; их можно также конфигурировать и посредством файла опций. Консоль, входящая в состав IBM SPSS Deployment Manager, может использоваться для отслеживания и конфигурирования установок SPSS Modeler Server; она доступна без дополнительной оплаты для действующих заказчиков SPSS Modeler Server. Эту прикладную программу можно установить только на компьютерах Windows; однако она может управлять сервером на любой поддерживаемой платформе.

IBM SPSS Modeler Batch

Хотя обычно исследование данных - интерактивный процесс, можно также запустить SPSS Modeler из командной строки, не открывая графический интерфейс. Например, у вас могут быть продолжительные или повторяющиеся задачи, которые желательно выполнить без участия пользователя. SPSS Modeler Batch - это особая версия продукта, предоставляющая поддержку всех аналитических возможностей SPSS Modeler без вызова обычного пользовательского интерфейса. SPSS Modeler Server необходим для использования SPSS Modeler Batch.

IBM SPSS Modeler Solution Publisher

SPSS Modeler Solution Publisher - это инструмент, при помощи которого можно создать пакетную версию потока SPSS Modeler; такую версию можно запускать внешним механизмом времени выполнения или встроить во внешнюю прикладную программу. Этим способом можно публиковать и внедрять полные потоки SPSS Modeler для использования в средах, где SPSS Modeler не установлен. SPSS Modeler Solution Publisher распространяется в составе службы IBM SPSS Collaboration and Deployment Services - Scoring, для которой требуется отдельная лицензия. С этой лицензией вы получаете SPSS Modeler Solution Publisher Runtime, при помощи которого можете запускать опубликованные потоки.

Дополнительную информацию о SPSS Modeler Solution Publisher смотрите в документации IBM SPSS Collaboration and Deployment Services. Центр знаний IBM SPSS Collaboration and Deployment Services содержит разделы "IBM SPSS Modeler Solution Publisher" и "IBM SPSS Analytics Toolkit".

Адаптеры IBM SPSS Modeler Server для IBM SPSS Collaboration and Deployment Services

Для IBM SPSS Collaboration and Deployment Services доступен ряд адаптеров, при посредстве которых SPSS Modeler и SPSS Modeler Server могут взаимодействовать с репозиторием IBM SPSS Collaboration and Deployment Services. При этом поток SPSS Modeler, внедренный в репозиторий, доступен для совместного использования несколькими пользователями или для обращения из прикладной программы IBM SPSS Modeler Advantage тонкого клиента. Адаптер устанавливается в той системе, в которой находится репозиторий.

Выпуски IBM SPSS Modeler

SPSS Modeler доступен в следующих выпусках.

SPSS Modeler Professional

SPSS Modeler Professional содержит все инструменты, необходимые для работы с большинством типов структурированных данных, таких как трассировка поведения и взаимодействия в системах CRM, демографии, поведения покупателей и данных о продажах.

SPSS Modeler Premium

SPSS Modeler Premium - это отдельно лицензируемый продукт, расширяющий SPSS Modeler Professional для работы со специализированными данными и с неструктурированными текстовыми данными. SPSS Modeler Premium включает в себя IBM SPSS Modeler Text Analytics:

IBM SPSS Modeler Text Analytics использует новейшие лингвистические технологии и обработку естественного языка (NLP) для быстрой обработки самых разнообразных неструктурированных текстовых данных, для извлечения и организации ключевых понятий и группирования этих понятий в категории. Извлеченные понятия и категории можно сочетать с существующими структурированными данными, такими как демографические, и применять к моделированию при помощи полного комплекта инструментов исследования данных IBM SPSS Modeler для получения более качественных и специализированных решений.

IBM SPSS Modeler Subscription

IBM SPSS Modeler Subscription обеспечивает все предсказательные аналитические возможности традиционного клиента IBM SPSS Modeler. В выпуске с подпиской вы можете регулярно скачивать обновления продукта.

Документация

К документации можно обратиться из меню Справка в SPSS Modeler. При этом открывается Центр знаний, общедоступный извне продукта.

Полная документация для каждого продукта (в том числе инструкции по установке) доступна также в формате PDF в нескольких сжатых папках как часть скачиваемого образа продукта. Кроме этого, можно скачать документы PDF с веб-сайта http://www.ibm.com/support/docview.wss?uid=swg27046871.

Документация к SPSS Modeler Professional

В комплект документации SPSS Modeler Professional (включая указания по установке) входят:

- **IBM SPSS Modeler Руководство пользователя.** Общее введение в использование SPSS Modeler, в том числе о создании потоков данных, обработке пропущенных значений, построению выражений CLEM работе с проектами и отчетами и составлению пакетов потоков для внедрения в IBM SPSS Collaboration and Deployment Services или IBM SPSS Modeler Advantage.
- **Узлы источников, обработки и вывода IBM SPSS Modeler.** Описания всех узлов, служащих для чтения, обработки и вывода данных в различных форматах. По существу это все узлы, кроме узлов моделирования.
- **Узлы моделирования IBM SPSS Modeler.** Описания всех узлов, служащих для создания моделей исследования данных. IBM SPSS Modeler предлагает ряд методов моделирования, взятых из таких областей, как обучение машин, искусственный интеллект и статистика.
- **Руководство по прикладным программам IBM SPSS Modeler.** Примеры в этом руководстве служат кратким специализированным введением к тем или иным методам и технологиям моделирования. Это руководство доступно также в электронном виде в меню Справка. Дополнительную информацию смотрите в разделе “Примеры прикладных программ” на стр. 4.
- **Сценарии и автоматизация Python IBM SPSS Modeler.** Информация об автоматизации системы путем создания сценариев Python, включая сценарии свойств, которые могут использоваться для работы с узлами и потоками.

- **Руководство по внедрению IBM SPSS Modeler .** Информация о выполнении IBM SPSS Modeler потоков как шагов обработки заданий под управлением IBM SPSS Deployment Manager.
- **Руководство разработчика IBM SPSS Modeler CLEF .** CLEF предоставляет возможности интеграции с программами других производителей, таких как подпрограммы обработки данных или алгоритмы моделирования, как с узлами в IBM SPSS Modeler.
- **Руководство по исследованию данных в базе данных IBM SPSS Modeler.** Информация о том, как использовать мощности вашей базы данных для повышения производительности и расширения диапазона возможностей анализа с привлечением алгоритмов от сторонних производителей.
- **Руководство администратора и руководство по производительности IBM SPSS Modeler Server .** Информация о том, как сконфигурировать и администрировать IBM SPSS Modeler Server.
- **IBM SPSS Deployment Manager Руководство пользователя.** Информация об использовании пользовательского интерфейса административной консоли включено в прикладную программу Deployment Manager для мониторинга и конфигурирования сервера IBM SPSS Modeler.
- **Руководство по CRISP-DM IBM SPSS Modeler.** Пошаговое руководство к использованию методологии CRISP-DM для исследования данных SPSS Modeler.
- **IBM SPSS Modeler Batch Руководство пользователя.** Полное руководство по использованию IBM SPSS Modeler в пакетном режиме, включая подробности выполнения в пакетном режиме и аргументы командной строки. Это руководство доступно только в формате PDF.

Документация SPSS Modeler Premium

В комплект документации SPSS Modeler Premium (включая указания по установке) входят:

- **SPSS Modeler Text Analytics Руководство пользователя.** Информация об использовании аналитики текстов совместно с SPSS Modeler, в том числе по узлам исследования текстов, интерактивной инструментальной среде, шаблонам и другим ресурсам.

Примеры прикладных программ

Инструменты исследования данных в SPSS Modeler помогают разрешить широкий спектр деловых и организационных проблем, а примеры прикладных программ предоставляют краткие, целевые введения в конкретные методы и способы моделирования. Используемые здесь наборы данных намного меньше огромных складов данных, которыми управляют некоторые исследователи данных, но применяемые понятия и методы должны масштабироваться до реальных прикладных программ.

Чтобы обратиться к примерам, выберите **Примеры прикладных программ** в меню Справка в SPSS Modeler.

Файлы данных и потоки примеров устанавливаются в папке `Demos` в каталоге установки продукта. Дополнительную информацию смотрите в разделе “Папка demos”.

Примеры моделирования баз данных. Смотрите эти примеры в руководстве *IBM SPSS Modeler: Руководство по исследованию данных в базе данных*.

Примеры сценариев. Смотрите эти примеры в руководстве *IBM SPSS Modeler Scripting and Automation Guide*.

Папка demos

Файлы данных и примеры потоков, используемые с примерами прикладных программ, устанавливаются в папке `Demos` в каталоге установки продукта (например: `C:\Program Files\IBM\SPSS\Modeler\<версия>\Demos`). К этой папке можно также обратиться из группы программ IBM SPSS Modeler в меню Пуск Windows или, щелкнув по `Demos` в списке недавно использовавшихся каталогов в диалоговом окне **Файл > Открыть поток**.

Отслеживание лицензий

При работе с SPSS Modeler использование лицензий отслеживается и записывается в журнал через регулярные интервалы времени. В журнал записываются показатели лицензирования *AUTHORIZED_USER* и *CONCURRENT_USER*; тип записываемого в журнал показателя зависит от типа лицензии, которая у вас есть для SPSS Modeler.

Генерируемые файлы журналов могут обрабатываться инструментом IBM License Metric Tool, из которого вы можете сгенерировать отчеты об использовании лицензий.

Файлы журналов лицензирования создаются в том же каталоге, куда записываются и файлы журналов клиента SPSS Modeler (по умолчанию `%ALLUSERSPROFILE%/9IBM/SPSS/Modeler/<версия>/log`).

Глава 2. Обзор продукта

Начинаем работу

Как прикладная программа исследования данных, IBM SPSS Modeler предлагает стратегический подход к поиску полезных взаимосвязей в больших наборах данных. В отличии от более традиционных статистических методов, здесь от вас не требуется заранее знать, что вы ищите. Вам дается возможность изучать данные, пробуя различные модели и исследуя разного рода взаимосвязи, пока вы не найдете полезную информацию.

Запуск IBM SPSS Modeler

Чтобы запустить прикладную программу, выберите:

Пуск > [Все] программы > IBM SPSS Modeler <version> > IBM SPSS Modeler <version>

Через несколько секунд откроется главное окно.

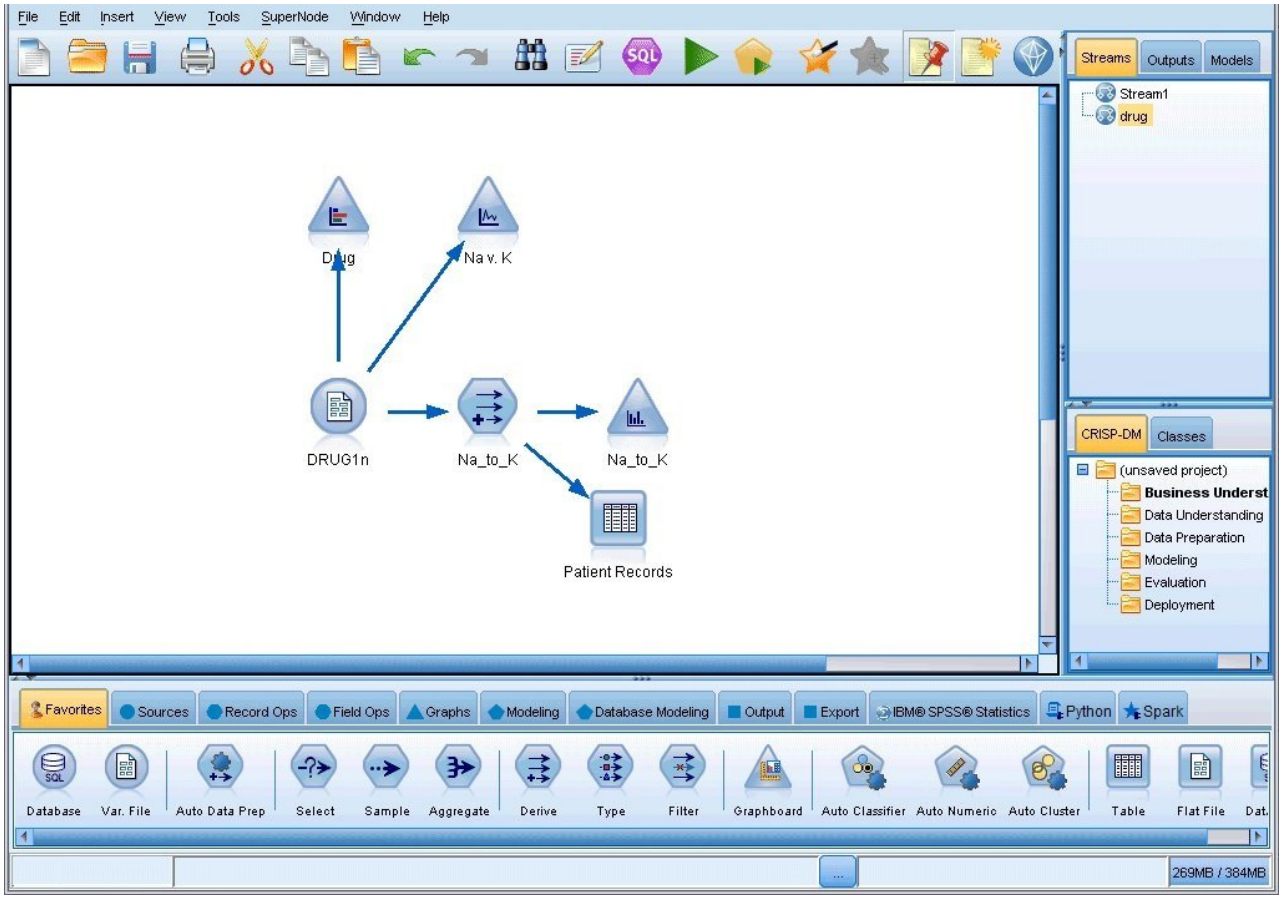


Рисунок 1. Главное окно прикладных программ IBM SPSS Modeler

Запуск из командной строки

Для запуска IBM SPSS Modeler можно использовать командную строку операционной системы, как описано ниже.

1. Откройте окно DOS (окно командной строки) на компьютере с IBM SPSS Modeler.
2. Для запуска интерфейса IBM SPSS Modeler в интерактивном режиме введите команду `modelerclient` с нужными аргументами, например:

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

Доступные аргументы (флаги) позволяют подключаться к серверу, загружать потоки, выполнять сценарии и указывать при необходимости прочие параметры выполнения.

Соединение с IBM SPSS Modeler Server

IBM SPSS Modeler можно запускать как автономную прикладную программу или как клиент, подключенный непосредственно к IBM SPSS Modeler Server или к IBM SPSS Modeler Server или кластеру сервера через подключаемый модуль координатора процессов в IBM SPSS Collaboration and Deployment Services. В нижней части окна IBM SPSS Modeler слева будет выведено текущее состояние соединения.

Каждый раз, когда требуется соединиться с сервером, можно вручную ввести имя сервера, к которому вы хотите подключиться, или выбрать заранее заданное вами имя. Однако если используется IBM SPSS Collaboration and Deployment Services, можно выполнить поиск в списке серверов или кластеров сервера в диалоговом окне Регистрация на сервере. Возможность просмотра через службы Statistics, работающие в сети, обеспечивается координатором процессов.

Чтобы соединиться с сервером:

1. В меню Инструменты выберите **Регистрация на сервере**. Откроется диалоговое окно Регистрация на сервере. Другой вариант - дважды щелкните по области состояния соединения в окне IBM SPSS Modeler.
2. В диалоговом окне задайте опции соединения с компьютером локального сервера или выберите соединение в таблице.
   - Нажмите кнопку **Добавить** или **Изменить**, чтобы добавить или изменить соединение. Дополнительную информацию смотрите в разделе “Добавление и изменение соединений с IBM SPSS Modeler Server” на стр. 9.
   - Нажмите кнопку **Поиск** для доступа к серверу или кластеру сервера в координаторе процессов. Дополнительную информацию смотрите в разделе “Поиск серверов в IBM SPSS Collaboration and Deployment Services” на стр. 9.

   Таблица серверов. Эта таблица содержит набор заданных соединений с серверами. В таблице выводятся соединение по умолчанию, имя сервера, описание и номер порта. Можно вручную добавить новое соединение, а также выбрать или найти существующее соединение. Чтобы задать конкретный сервер в качестве соединения по умолчанию, включите переключатель в столбце По умолчанию таблицы этого соединения.

   Путь к данным по умолчанию. Задайте путь для данных на компьютере сервера. Нажмите кнопку с многоточием **(...)** для просмотра нужного положения.

   Установите учетные записи. Оставьте этот переключатель выключенным, чтобы разрешить возможность **единой регистрации**, которая пытается зарегистрироваться на сервере, используя локальное имя пользователя и пароль на вашем компьютере. Если единая регистрация невозможна или отключена этим переключателем (например, для входа в учетную запись администратора), для ввода параметров регистрации будут доступны следующие поля.

   ID пользователя. Введите имя пользователя для входа на сервер.

   Пароль. Укажите пароль, связанный с указанным имени пользователя.

   Домен. Задайте домен, который будет использоваться для входа на сервер. Доменное имя требуется, только если компьютер сервера находится в другом домене Windows по отношению к компьютеру клиента.
3. Нажмите кнопку **OK**, чтобы завершить подключение.

Чтобы отсоединиться от сервера:

1. В меню Инструменты выберите **Регистрация на сервере**. Откроется диалоговое окно Регистрация на сервере. Другой вариант - дважды щелкните по области состояния соединения в окне IBM SPSS Modeler.
2. В диалоговом окне выберите Локальный сервер и нажмите кнопку **OK**.

Добавление и изменение соединений с IBM SPSS Modeler Server

Можно вручную изменить или добавить соединение с сервером в диалоговом окне Регистрация на сервере. При нажатии кнопки Добавить открывается пустое диалоговое окно Добавить/Изменить сервер, в которое можно ввести подробности соединения с сервером. После выбора существующего соединения и нажатия кнопки Изменить в диалоговом окне Подключение к серверу откроется диалоговое окно Добавить/Изменить сервер со сведениями об этом соединении, где можно внести необходимые изменения.

Примечание: Соединение с сервером, добавленное из IBM SPSS Collaboration and Deployment Services, нельзя изменить, поскольку имя, номер порта и другая информация заданы в IBM SPSS Collaboration and Deployment Services. Рекомендуется использовать одни и те же порты для связи с IBM SPSS Collaboration and Deployment Services и с клиентом SPSS Modeler. Они могут быть заданы как `max_server_port` и `min_server_port` в файле options.cfg.

Чтобы добавить соединения с сервером

1. В меню Инструменты выберите **Регистрация на сервере**. Откроется диалоговое окно Регистрация на сервере.
2. В этом диалоговом окне нажмите кнопку **Добавить**. Откроется диалоговое окно Добавить/Изменить сервер.
3. Введите сведения о соединении с сервером и нажмите кнопку **OK**, чтобы сохранить это соединение и вернуться в диалоговое окно Регистрация на сервере.

- **Сервер.** Задайте доступный сервер или выберите его из списка. Компьютер сервера задается алфавитно-цифровым именем (например, *myserver*) или IP-адресом, назначенным компьютеру сервера, (например, 202.123.456.78).
- **Порт.** Укажите номер порта, на котором сервер ожидает сообщений. Если номер по умолчанию не работает, попросите у администратора системы правильный номер порта.
- **Описание.** Введите необязательное описание для этого соединения с сервером.
- **Обеспечить защищенное соединение (использовать SSL).** Укажите, нужно ли использовать подключение по протоколу SSL (**Secure Sockets Layer**). SSL - обычно используемый протокол для защиты данных, пересылаемых по сети. Для доступа к этой функции необходимо разрешить SSL на сервере, где запущен IBM SPSS Modeler Server. При необходимости обращайтесь к вашему администратору для получения более полной информации.

Чтобы изменить соединения с сервером

1. В меню Инструменты выберите **Регистрация на сервере**. Откроется диалоговое окно Регистрация на сервере.
2. В этом диалоговом окне выберите соединение, которое нужно изменить, и нажмите кнопку **Изменить**. Откроется диалоговое окно Добавить/Изменить сервер.
3. Измените сведения о соединении с сервером и нажмите кнопку **OK**, чтобы сохранить это соединение и вернуться в диалоговое окно Регистрация на сервере.

Поиск серверов в IBM SPSS Collaboration and Deployment Services

Вместо того, чтобы вводить соединение с сервером вручную, можно выбрать в сети сервер или кластер серверов через координатор процессов, доступный в IBM SPSS Collaboration and Deployment Services. Кластер серверов - это группа серверов, внутри которой координатор процессов определяет сервер, лучше всего подходящий для ответов на запрос обработки.

Хотя можно и вручную добавлять серверы в диалоговом окне Подключение к серверу, поиск имеющихся серверов позволяет соединяться с серверами без необходимости знания правильного имени сервера и номера

порта. Эта информация предоставляется автоматически. Однако при этом необходима правильная информация об учетной записи, такая как имя пользователя, домен и пароль.

Примечание: Если у вас нет доступа к функции координатора процессов, можно вручную ввести имя сервера с которым вы хотите соединиться, или выбрать имя, заданное вами ранее. Дополнительную информацию смотрите в разделе “Добавление и изменение соединений с IBM SPSS Modeler Server” на стр. 9.

Чтобы найти серверы и кластеры:

1. В меню Инструменты выберите **Регистрация на сервере**. Откроется диалоговое окно Регистрация на сервере.
2. В диалоговом окне нажмите кнопку **Поиск**, чтобы открыть диалоговое окно Поиск серверов. Если вы еще не вошли в IBM SPSS Collaboration and Deployment Services при попытке просмотра координатора процессов, вам предложат сделать это.
3. Выберите сервер или кластер серверов в списке.
4. Нажмите кнопку **OK**, чтобы закрыть диалоговое окно и добавить это соединение в диалоговое окно Подключение к серверу.

Соединение с Analytic Server

Если у вас есть несколько доступных механизмов Analytic Server, то с помощью диалога Соединение с сервером Analytic Server укажите более одного сервера для использования в IBM SPSS Modeler. Ваш администратор мог уже задать Analytic Server по умолчанию в файле `<путь_установки_Modeler>/config/options.cfg`. Но можно использовать и другие доступные серверы после того, как они будут определены. Например, при использовании узлов Источник и Экспорт Analytic Server, возможно, имеет смысл применять разные соединения Analytic Server в разных ветвях потока, так что каждая ветвь будет использовать свой собственный Analytic Server и данные не будут извлекаться в IBM SPSS Modeler Server. Заметим, что если ветвь содержит более одного соединения Analytic Server, то данные будут извлекаться из Analytic Server в IBM SPSS Modeler Server. Дополнительную информацию, в том числе об ограничениях, смотрите в разделе Свойства потока Analytic Server.

Чтобы создать новое соединение Analytic Server, перейдите в **Инструменты > Соединения сервера Analytic Server** и введите необходимую информацию в следующих разделах диалога.

Подключение

URL. Введите URL для Analytic Server в формате `https://имя_хоста:порт/корень_контекста`, где `имя_хоста` - это IP-адрес или имя хоста Analytic Server, `порт` - это номер порта, а `корень_контекста` - это корень контекста для Analytic Server.

Арендатор. Введите имя арендатора, элемент которого - IBM SPSS Modeler Server Если вы не знаете арендатора, обратитесь к своему системному администратору.

Аутентификация

Режим. Выберите из следующих режимов аутентификации.

- В поле **Имя пользователя и пароль** требуется ввести имя пользователя и пароль.
- В поле **Хранимые идентификационные данные** требуется ввести идентификационные данные из IBM SPSS Collaboration and Deployment Services Repository.
- В поле **Kerberos** требуется ввести имя принципала службы и путь к файлу конфигурации. Если у вас нет этой информации, обратитесь к своему системному администратору.

Имя пользователя. Введите имя пользователя Analytic Server.

Пароль. Введите пароль Analytic Server.

Соединиться. Нажмите кнопку **Соединиться**, чтобы проверить новое соединение.

Подключения

Когда вы укажете информацию выше и щелкнете по **Соединиться**, соединение будет добавлено в эту таблицу соединений. Если надо удалить соединение, то выберите его и щелкните по **Удалить**.

Если ваш администратор задал соединение с Analytic Server по умолчанию в файле `options.cfg`, то можно выбрать **Добавить соединение по умолчанию**, чтобы добавить и его к списку ваших доступных соединений. Вас попросят ввести имя пользователя и пароль.

Изменение каталога temp

Для некоторых операций, выполняемых IBM SPSS Modeler Server, требуется создавать временные файлы. По умолчанию IBM SPSS Modeler создает временные файлы в каталоге временных файлов системы. Вы можете изменить положение каталога временных файлов, как описано ниже.

1. Создайте новый каталог с именем `spss` и подкаталог с именем *servertemp*.
2. Откройте для редактирования файл *options.cfg* в подкаталоге */config* в каталоге установки IBM SPSS Modeler. Отредактируйте в этом файле параметр `temp_directory`, задав: `temp_directory, "C:/spss/servertemp"`.
3. После этого нужно перезапустить службу IBM SPSS Modeler Server. Для этого можно щелкнуть по вкладке **Службы** на Панели управления Windows. Просто остановите эту службу и запустите ее еще раз, чтобы активировать внесенные изменения. Кроме того, служба перезапустится при перезапуске компьютера.

Теперь все временные файлы будут записываться в этот новый каталог.

Примечание:

- Надо использовать наклонные черты вправо.
- При выполнении потоков оценки посредством запуска заданий IBM SPSS Collaboration and Deployment Services параметр `temp_directory` не применяется. При запуске такого задания создается временный файл. По умолчанию этот файл сохраняется в каталоге установки сервера IBM SPSS Modeler. Папку данных по умолчанию, в которой сохраняются временные файлы, можно изменить при создании соединения с сервером IBM SPSS Modeler в IBM SPSS Modeler.

Запуск нескольких сеансов IBM SPSS Modeler

Если нужно запускать несколько сеансов IBM SPSS Modeler одновременно, требуются некоторые изменения в настройке IBM SPSS Modeler и Windows. Например, это нужно, если у вас две серверных лицензии и вы хотите выполнять два потока на двух серверах с одного компьютера клиента.

Чтобы разрешить несколько сеансов IBM SPSS Modeler:

1. Выберите:

   Пуск > [Все] Программы > IBM SPSS Modeler
2. Щелкните правой кнопкой по ярлыку (значку) IBM SPSS Modeler и выберите **Свойства**.
3. В текстовом поле **Назначение** добавьте `-noshare` в конце строки.
4. В Проводнике Windows выберите:

   Сервис > Свойства папки...
5. На вкладке Типы файлов выберите пункт Поток IBM SPSS Modeler и нажмите кнопку **Дополнительно**.
6. В диалоговом окне Изменений свойств типа файлов выберите пункт Открыть в IBM SPSS Modeler и нажмите кнопку **Изменить**.
7. В текстовом поле **Приложение, исполняющее действие** добавьте `-noshare` перед аргументом **-stream**.

Беглый взгляд на интерфейс IBM SPSS Modeler

В каждой точке процесса исследования данных простой в использовании интерфейс IBM SPSS Modeler привлекает ваш конкретный опыт в бизнесе. При этом мощные и точные модели обеспечиваются применением таких алгоритмов моделирования, как прогнозные алгоритмы, а также алгоритмы классификации, сегментации и обнаружения связей. Результаты моделей легко можно встроить и передать в базы данных, IBM SPSS Statistics и широкий круг других прикладных программ.

Работа с инструментом IBM SPSS Modeler включает в себя три этапа работы с данными.

- Во-первых, данные передаются в IBM SPSS Modeler.
- Во-вторых, данные пропускаются через последовательность операций.
- Наконец, данные отправляются в объект назначения.

Эта последовательность операций называется **поток данных**, поскольку данные запись за записью текут потоком из источника, проходят через все операции и, наконец, стекают в объект назначения -- модель или один из типов вывода данных.

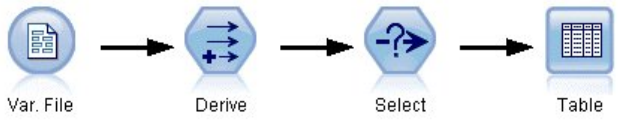


Рисунок 2. Простой поток

Холст потока IBM SPSS Modeler

Холст потока - самая большая область окна IBM SPSS Modeler, в которой строят потоки данных и работают с ними.

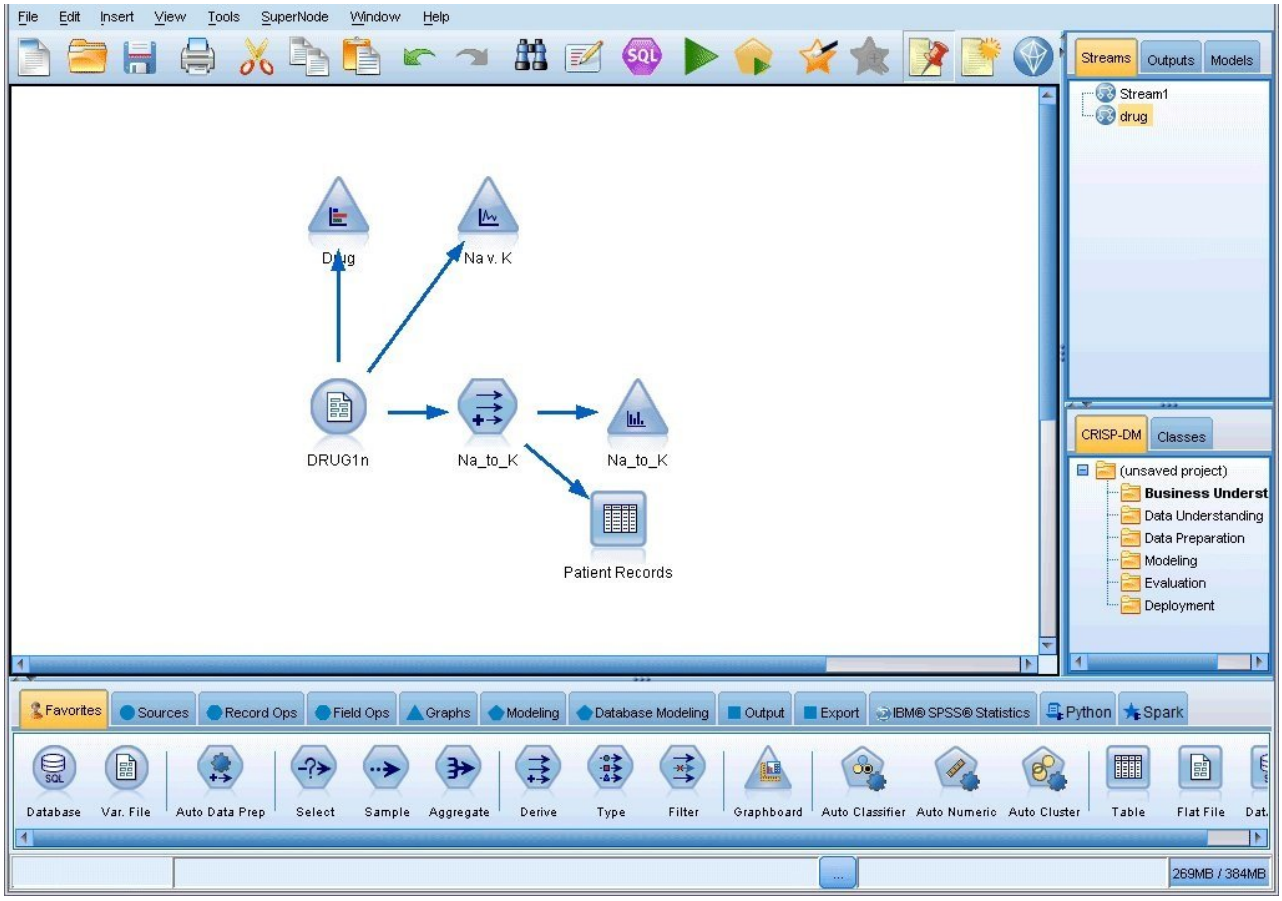


Рисунок 3. Рабочее пространство IBM SPSS Modeler (представление по умолчанию)

Потоки создаются на холсте интерфейса путем построения диаграмм, содержащих нужные операции над данными. Каждая операция представлена значком, или **узлом**, и узлы соединены стрелками в **поток**, представляющий путь данных через все операции.

В IBM SPSS Modeler можно одновременно работать с несколькими потоками, как на одном холсте, так и открыв новый холст потока. В течение сеанса потоки хранятся в менеджере потоков в верхнем правом углу окна IBM SPSS Modeler.

Примечание: При использовании MacBook с включенной опцией встроенного трекпада **Force Click and haptic feedback** перетаскивание узлов с палитры на холст потока может привести к дублированию добавляемых на холст узлов. Чтобы избежать этой проблемы, мы рекомендуем отключать системное предпочтение трекпада **Force Click and haptic feedback**.

Палитра узлов

Большинство инструментов данных и моделирования в SPSS Modeler доступны на *палитре узлов* в нижней полосе окна под холстом потока.

Например, вкладка **Операции над записями** палитры узлов содержит узлы, которые можно использовать для таких операций над *записями* данных, как отбор, объединение и присоединение.

Чтобы добавить узлы на холст, щелкайте дважды по значкам на палитре узлов или перетаскивайте их на холст. Затем соедините их, чтобы создать *поток*, по которому будут передаваться данные.

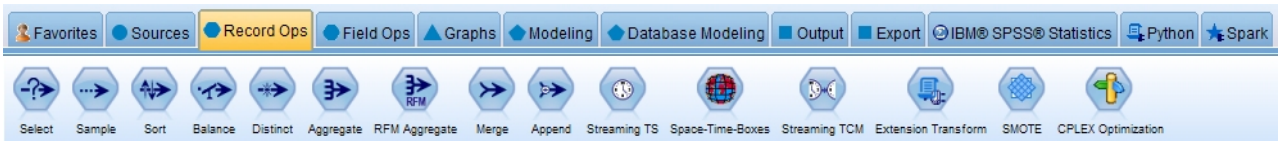


Рисунок 4. Вкладка Операции над записями на палитре узлов

Каждая вкладка на палитре узлов содержит подборку узлов, применяемых на том или ином этапе потока, а именно:

- Узлы **Источники** поставляют данные в SPSS Modeler.
- Узлы **Операции над записями** выполняют такие операции над *записями* данных, как отбор, объединение и присоединение.
- Узлы **Операции над полями** выполняют такие операции над *полями* данных, как фильтрация, вычисление новых полей и выяснение типа измерений для данных полей.
- Узлы **Диаграммы** выводят наглядные представления данных до и после моделирования. Доступны такие диаграммы, как графики, гистограммы, веб-узлы и диаграммы оценки.
- Узлы **Моделирование** используют алгоритмы моделирования, доступные в SPSS Modeler, такие как нейросети, деревья решений, алгоритмы кластеризации и секвенирование данных.
- Узлы **Моделирование баз данных** используют алгоритмы моделирования, доступные в базах данных Microsoft SQL Server, IBM Db2, Oracle и Netezza.
- Узлы **Вывод** создают различные представления данных, диаграмм и результатов моделирования, которые можно просматривать в SPSS Modeler.
- Узлы **Экспорт** создают различные представления для просмотра во внешних прикладных программах, таких как IBM SPSS Data Collection или Excel.
- Узлы **IBM SPSS Statistics** импортируют данные из IBM SPSS Statistics и импортируют их сюда, а также запускают процедуры IBM SPSS Statistics.
- Узлы **Python** можно использовать для запуска алгоритмов Python.
- Узлы **Spark** можно использовать для запуска алгоритмов Spark.

Познакомившись ближе с SPSS Modeler, вы можете настроить содержимое палитры, как вам удобнее.

В левой части палитры узлов можно отфильтровать выводимые узлы, выбрав Supervised, Связывание или Сегментация.

Ниже Палитры узлов находится панель отчетов, которая показывает ход выполнения различных операций, например, чтения данных в поток данных. Также ниже Палитры узлов находится панель состояния, содержащая информацию о том, чем в настоящее время занята прикладная программа и требуется ли ввод информации от пользователя.

Примечание: При использовании MacBook с включенной опцией встроенного трекпада **Force Click and haptic feedback** перетаскивание узлов с палитры на холст потока может привести к дублированию добавляемых на холст узлов. Чтобы избежать этой проблемы, мы рекомендуем отключать системное предпочтение трекпада **Force Click and haptic feedback**.

Менеджеры IBM SPSS Modeler

В верхней правой части окна расположена панель менеджеров. Она содержит три вкладки для управления потоками, выводом и моделями.

На вкладке Потоки можно открывать, переименовывать, сохранять и удалять потоки, созданные в сеансе.

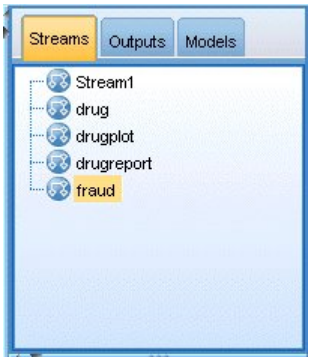


Рисунок 5. Вкладка Потоки

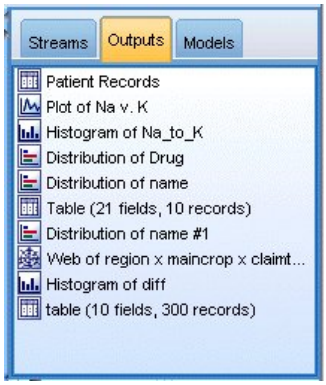


Рисунок 6. Вкладка Вывод

Вкладка Вывод содержит различные файлы, такие как диаграммы и таблицы, созданные операциями потоков в IBM SPSS Modeler. Перечисленные на вкладке таблицы, диаграммы и отчеты можно выводить на экран, сохранять, переименовывать и закрывать.

Рисунок 7. Вкладка моделей, содержащая слепки моделей

Из трех вкладок менеджеров вкладка моделей - самая богатая по арсеналу средств. Эта вкладка содержит все **слепки** моделей по всем моделям, сгенерированным в IBM SPSS Modeler в текущем сеансе. Эти модели можно просматривать непосредственно со вкладки моделей, а можно добавлять в поток на холсте.

Проекты IBM SPSS Modeler

У правого края окна внизу расположена панель проектов, на которой можно создавать **проекты** исследования данных (группы файлов, связанных с некоторой задачей исследования данных) и управлять ими. Есть два способа просматривать проекты, создаваемые в IBM SPSS Modeler — в представлении Классы и в представлении CRISP-DM.

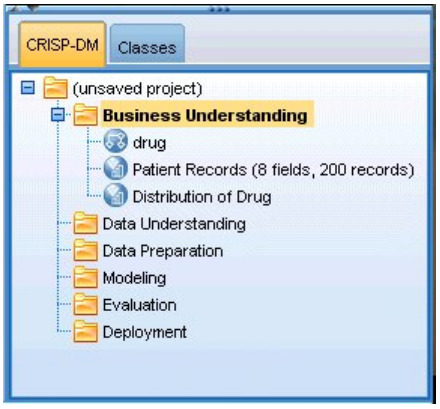


Рисунок 8. Представление CRISP-DM

Вкладка CRISP-DM обеспечивает один из способов организовать проекты - по испытанной на практике, непроприетарной методологии Cross-Industry Standard Process for Data Mining. Как опытным, так и начинающим исследователям данных инструмент CRISP-DM поможет лучше организовать и скоординировать работу.

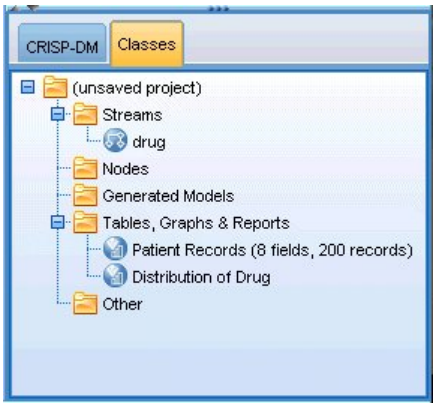


Рисунок 9. Представление классов

Вкладка Классы обеспечивает один из способов организовать работу в IBM SPSS Modeler по категориям — по типам создаваемых объектов. Это представление полезно при инвентаризации данных, потоков и моделей.

Панель инструментов IBM SPSS Modeler

В верхней части окна IBM SPSS Modeler находится панель со значками, выполняющими ряд полезных действий. Ниже описываются значки панели инструментов и их назначение.

Создать новый поток

Открыть поток

Сохранить поток

Вывести на печать текущий поток

Вырезать & переместить в буфер обмена

Копировать в буфер обмена

Вставить выделенное

Отменить последнее действие

Повторить

Искать узлы

Правка свойств потока

Предварительный просмотр сгенерированного SQL

Запустить текущий поток

Запустить выбранную часть потока

Остановить поток (значок активен, только если поток запущен)

Добавить надузел

Увеличить (только на надузлах)

Уменьшить (только на надузлах)

Без разметки в потоке

Вставить комментарий

Скрыть разметку потока (если есть)

Показать скрытую разметку потока

Открыть поток в IBM SPSS Modeler Advantage

Разметка потока состоит из комментариев к потоку, ссылок на модели и оценочных показателей ветвей.

Ссылки на модели описаны в руководстве *Узлы моделирования IBM SPSS*.

Настройка панели инструментов

Можно изменить различные аспекты панели инструментов, такие как:

- Показывать ли панель инструментов
- Доступны ли подсказки при значках
- Крупные значки или мелкие

Чтобы показать и скрыть панель инструментов:

1. Выберите в главном меню:

   Вид > Панель инструментов > Показать

Чтобы изменить настройку подсказок или размер значков:

1. Выберите в главном меню:

   Вид > Панель инструментов > Настроить

Выберите **Показывать подсказки** или **Крупные кнопки**, как это потребуется.

Настройка окна IBM SPSS Modeler

Используя разделители между различными частями интерфейса SPSS Modeler, вы можете изменять размеры и закрывать инструменты, как вам будет удобнее. Например, если вы работаете с большим потоком, вы можете, нажимая на маленькие стрелки, которые есть на каждом разделителе, закрыть палитру узлов, панель менеджеров и панель проектов. Это увеличит размер холста потока, освободив место для работы над большим потоком или несколькими потоками.

Другой вариант - в меню Вид выбрать **Палитра узлов**, **Менеджеры** или **Проект**, чтобы включить или выключить эти элементы.

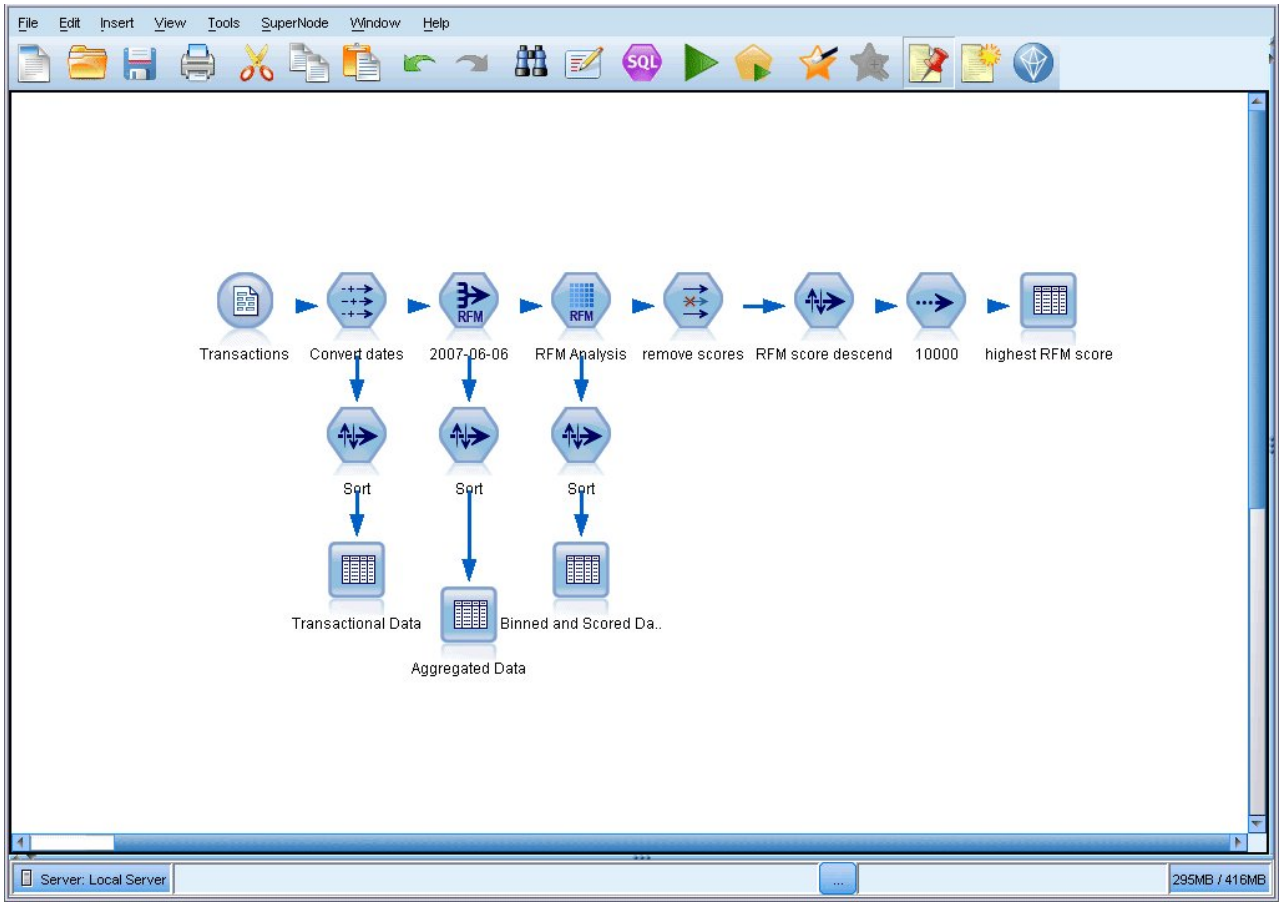


Рисунок 10. Развернуть холст потока

Вместо того, чтобы закрыть палитру узлов и панели менеджеров и проекта, можно пользоваться холстом потока как страницей с прокруткой, перемещаясь по холсту при помощи вертикальной и горизонтальной полос прокрутки сбоку и внизу окна SPSS Modeler.

Кроме того, можно управлять выводом разметки, к которой относятся комментарии к потоку, ссылки на модели и индикаторы ветви оценки. Чтобы показать или скрыть эти подробности, выберите:

Вид > Разметка потока

Изменение размера значка для потока

Размер значков потока можно изменять следующими способами.

- Через настройку свойств потока
- Через всплывающее меню в потоке
- При помощи клавиатуры

Представление всего потока можно масштабировать до ряда размеров от 8% до 200% стандартного размера значка.

Чтобы масштабировать весь поток (используя свойства потока)

1. В главном меню выберите

   Инструменты > Свойства потока > Опции > Макет.
2. Выберите нужный размер в меню Размер значка.
3. Нажмите кнопку **Применить**, чтобы увидеть результат.

4. Нажмите кнопку **OK**, чтобы сохранить изменение.

Чтобы масштабировать весь поток (используя меню)

1. Щелкните правой кнопкой по фоновому участку холста потока.
2. Выберите пункт меню **Размер значка** и выберите нужный размер значка.

Чтобы масштабировать весь поток (используя клавиатуру)

1. Нажмите Ctrl + [-] на основной клавиатуре, чтобы уменьшить размер до предыдущего в ряду допустимых размеров.
2. Нажмите Ctrl + [+] на основной клавиатуре, чтобы увеличить размер до следующего в ряду допустимых размеров.

Эта возможность особенно полезна для получения общего обзора сложного потока. Кроме того, с ее помощью можно минимизировать число необходимых страниц при печати потока.

Использование мыши в IBM SPSS Modeler

Основные способы использования мыши в IBM SPSS Modeler включают следующее:

- **Один щелчок.** Правая или левая кнопка мыши выбирает пункт в меню, открывает всплывающее меню и выбирает обычные элементы контроля и опции. Щелчок и удержание кнопки служит для перетаскивания узлов.
- **Двойной щелчок.** Щелкните дважды левой кнопкой, чтобы поместить узел на холст потока и отредактировать существующий узел.
- **Щелчок средней кнопкой.** Нажмите среднюю кнопку и перетащите указатель, чтобы соединить узлы на холсте потока. Щелкните средней кнопкой дважды, чтобы отсоединить узел. Если у вас нет трехкнопочной мыши, эту возможность можно сымитировать, нажимая клавишу Alt во время щелчка и перетаскивания мышью.

Использование клавиш быстрого вызова

Со многими операциями визуального программирования в IBM SPSS Modeler связаны клавиатурные сокращения. Например, можно удалить узел, щелкнув по этому узлу и нажав клавишу Delete на клавиатуре. Другой пример - можно быстро сохранить поток, если нажать клавишу S, удерживая нажатой клавишу Ctrl. Подобные управляющие команды записываются как сочетание клавиши Ctrl с другой клавишей -- например, как Ctrl+S.

Ряд клавиш быстрого вызова используется при обычных операциях Windows, например, Ctrl+X, чтобы вырезать в буфер обмена. IBM SPSS Modeler поддерживает такие клавиатурные сокращения наряду с приведенными ниже клавишами быстрого вызова специально для прикладной программы.

Примечание: Старые клавиши быстрого вызова, использовавшиеся в IBM SPSS Modeler, в ряде случаев конфликтуют со стандартными клавиатурными сокращениями Windows. Эти старые сочетания клавиш теперь поддерживаются с добавлением клавиши Alt. Например, для включения и выключения кэша теперь можно использовать Ctrl+Alt+C.

Таблица 1. Поддерживаемые клавиши быстрого вызова

Клавиши	Функция
Ctrl+A	Выделить все
Ctrl+X	Вырезать
Ctrl+N	Новый поток
Ctrl+O	Открыть поток
Ctrl+P	Печатать
Ctrl+C	Скопировать

Таблица 1. Поддерживаемые клавиши быстрого вызова (продолжение)

Клавиши	Функция
Ctrl+V	Вставить
Ctrl+Z	Откат
Ctrl+Q	Выделить все узлы вниз по направлению потока от выделенного узла
Ctrl+W	Снять выделение со всех узлов вниз по направлению потока (работает совместно с Ctrl+Q как включение-выключение)
Ctrl+E	Выполнить выделенный узел
Ctrl+S	Сохранить текущий поток
Alt+клавиши со стрелками	Переместить выделенные узлы на холсте потока в направлении стрелки на клавише
Shift+F10	Открыть всплывающее меню для выделенного узла

Таблица 2. Поддерживаемые клавиатурные сокращения вместо старых горячих клавиш

Клавиши	Функция
Ctrl+Alt+D	Дублировать узел
Ctrl+Alt+L	Загрузить узел
Ctrl+Alt+R	Переименовать узел
Ctrl+Alt+U	Создать узел пользовательского ввода
Ctrl+Alt+C	Включить-выключить кэш
Ctrl+Alt+F	Очистить кэш
Ctrl+Alt+X	Развернуть надузел
Ctrl+Alt+Z	Детализировать или свернуть
Удалить	Удалить узел или соединение

Печать

Приведенные ниже объекты можно напечатать в IBM SPSS Modeler:

- Диаграммы потока
- Диаграммы
- Таблицы
- Отчеты (с узла отчетов и из окна Отчеты по проекту)
- Сценарии (из диалоговых окон свойств потока, отдельного сценария или сценария надузла)
- Модели (из браузера моделей, выбранных вкладок диалогового окна, просмотра деревьев)
- Аннотации (используя для вывода вкладку аннотаций)

Чтобы напечатать объект:

- Для печати без предварительного просмотра нажмите кнопку Печать на панели инструментов.
- Чтобы настроить страницу перед печатью, выберите **параметры страницы** в меню Файл.
- Для предварительного просмотра перед печатью выберите **Предварительный просмотр печати** в меню Файл.
- Чтобы открыть обычное диалоговое печати с опциями выбора принтера и настройками стиля выберите **Печать** в меню Файл.

Автоматизация IBM SPSS Modeler

Поскольку исследование данных может быть сложным и подчас продолжительным процессом, IBM SPSS Modeler содержит в себе несколько типов поддержки кодирования и автоматизации.

- Язык **CLEM** (Control Language for Expression Manipulation, управляющий язык для преобразования выражений) - это язык для анализа и обработки данных в потоках IBM SPSS Modeler. Исследователи данных широко используют CLEM в операциях потока для выполнения таких простых задач, как вычисление прибыли по данным затрат и доходов, и таких сложных, как преобразование данных Web-журнала в набор пригодных для использования полей и записей.
- **Сценарии** - мощный инструмент для автоматизации процессов в интерфейсе пользователя. Сценарии выполняют некоторые типы действий, которые пользователи могут выполнять при помощи мыши или клавиатуры. Кроме того, можно задавать вывод и выполнять операции над сгенерированными моделями.

Глава 3. Введение в моделирование

Модель - это набор правил, формул или уравнений, которые можно использовать для предсказания выходных данных на основании набора входных полей или переменных. Например, финансовая компания может использовать модель для предсказания рисков выдачи кредита обратившимся за ней клиентам на основании информации, уже известной о бывших клиентах.

Возможность предсказания выходных данных - это основная цель предсказательной аналитики, а понимание процесса моделирования - это ключ к использованию IBM SPSS Modeler.

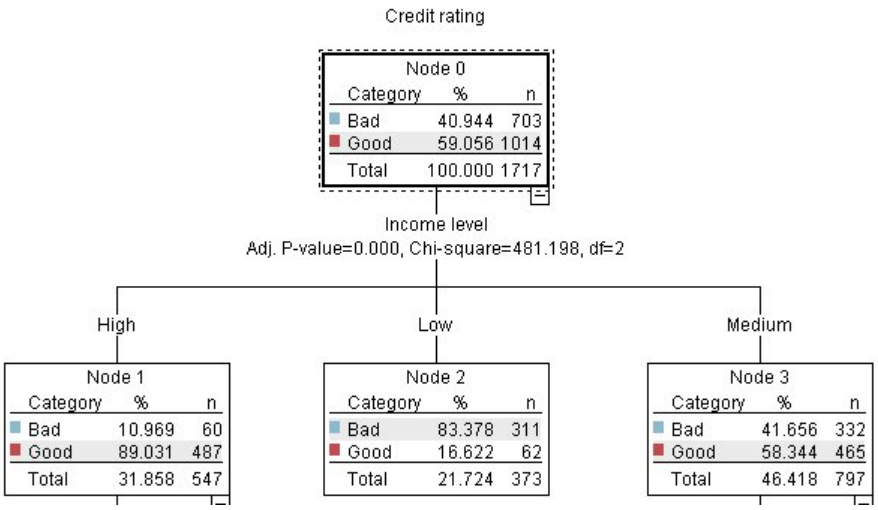


Рисунок 11. Простая модель дерева решений

В этом примере используется модель **дерева решений**, которая классифицирует записи (и предсказывает отклик), используя ряд правил решений, например:

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

Хотя в этом примере используется модель CHAID (Chi-squared Automatic Interaction Detection - автоматическое обнаружение взаимодействия хи-квадрат), он предназначен для общего введения, и большинство понятий широко применяются в других типах моделирования в IBM SPSS Modeler.

Для понимания любой модели сначала нужно понять, какие данные в нее поступают. В этом примере данные содержат информацию о клиентах банка. Используются следующие поля:

Имя поля	**Описание**
Credit_rating	Кредитный рейтинг: 0=Плохой, 1=Хороший, 9=значения отсутствия
Age	Возраст в годах
Income	Уровень дохода: 1=Низкий, 2=Средний, 3=Высокий
Credit_cards	Количество кредитных карт у клиента: 1=меньше пяти, 2=пять или больше
Education	Уровень образования: 1=Высшая школа, 2=Колледж
Car_loans	Количество выданных ссуд на покупку автомобиля: 1=Не было или одна, 2=Две или больше

Банк поддерживает базу данных с хронологической информацией о клиентах, которые брали ссуды в этом банке, в частности, выполняли ли они свои обязательства (Кредитный рейтинг = Хороший) или отказывались от платежей (Кредитный рейтинг = Плохой). Используя эти существующие данные, банк хочет построить модель, которая позволит предсказать, насколько вероятно, что будущие обращения за ссудами приведут к неплатежам.

Используя модель дерева решений, вы можете проанализировать характеристики двух групп клиентов и предсказать вероятность отказа от платежей по ссуде.

В этом примере используется поток с именем *modelingintro.str*, доступный в папке *Demos* в подпапке *streams*. Файл данных - это *tree_credit.sav*. Дополнительную информацию смотрите в разделе “Папка demos” на стр. 4.

Давайте посмотрим на этот поток.

1. В главном меню выберите:
   Файл > Открыть поток
2. Щелкните по золотому значку слепка на панели инструментов диалогового окна Открыть и выберите папку Demos.
3. Дважды щелкните по папке *streams*.
4. Дважды щелкните по файлу с именем *modelingintro.str*.

Построение потока

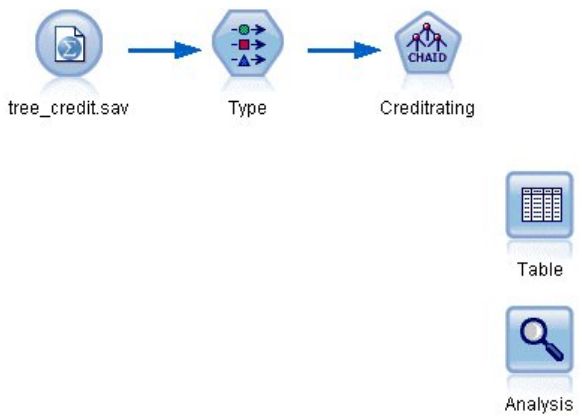


Рисунок 12. Поток моделирования

Для построения потока, который будет создавать модель, нам нужно по крайней мере три элемента:

- Узел источника, читающий данные с некоторого внешнего источника, в данном случае - из файла данных IBM SPSS Statistics.
- Узел источника или дополнительный узел Тип, где задаются свойства полей, такие как уровень измерения (тип данных, содержащихся в поле) и роль каждого поля (входное поле или поле назначения для моделирования).
- Узел моделирования, генерирующий слепок модели, когда запущен поток.

В этом примере мы используем узел моделирования CHAID. CHAID (Chi-squared Automatic Interaction Detection - автоматическое обнаружение взаимодействия хи-квадрат) - это метод классификации для построения деревьев решений с использованием конкретного типа статистики, известного как статистика хи-квадрат, для нахождения оптимальных точек расщепления в дереве решений.

Если уровни измерений заданы на узле источника, отдельный узел Тип можно исключить. Функционально результат будет таким же.

У этого потока есть также узлы Таблица и Анализ, которые будут использоваться для просмотра результатов скоринга после создания слепка модели и добавления его к потоку.

Узел источника Файл статистики читает данные в формате IBM SPSS Statistics из файла данных *tree_credit.sav*, установленного в папке *Demos*. (Специальная переменная с именем *$CLEO_DEMOS* используется для указания на положение этой папки в текущей установке IBM SPSS Modeler. Это обеспечивает правильный путь независимо от папки или версии текущей установки).

Рисунок 13. Чтение данных при помощи узла источника Файл статистики

Узел Тип задает **уровень измерения** для каждого поля. Уровень измерения - это категория, обозначающая тип данных в поле. Наш файл данных использует три разных уровня измерения.

Количественное поле (такое как поле *Возраст*) содержит количественные численные значения, а у **Номинального** поля (такого как поле *Кредитный рейтинг*) есть два или более отдельных значения, например, *Плохой*, *Хороший* или *Нет кредитной истории*. **Порядковое** поле (такое как поле *Уровень дохода*) описывает данные с несколькими отдельными значениями, для которых есть естественный присущий им порядок, в данном случае - *Низкий*, *Средний* и *Высокий*.

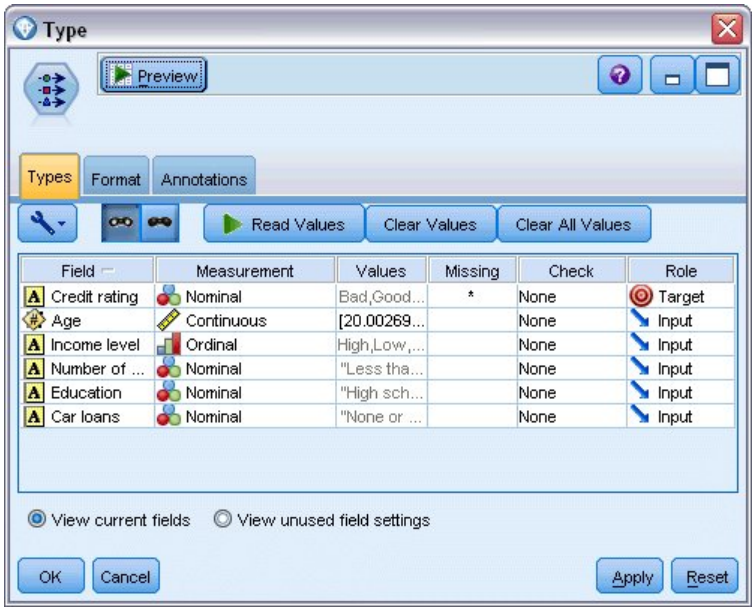


Рисунок 14. Задание полей назначения и входных полей на узле Тип

Для каждого поля узел Тип задает также **роль**, которую каждое поле играет при моделировании. Роль *Назначение* задается для поля *Кредитный рейтинг*, которое обозначает, выполнит ли конкретный клиент свои обязательства по кредиту. Это **назначение** моделирования, поле, значение в котором мы хотим предсказать.

Для других полей задается роль *Входные* поля. Входные поля иногда называют **предикторами**, то есть полями, значения которых используются для алгоритма моделирования, чтобы предсказать значение в поле назначения.

Узел моделирования CHAID генерирует модель.

На вкладке Поля узла моделирования выбирается опция **Использовать предварительно определенные роли**, что означает использование поля назначения и входных полей, заданных на узле Тип. В этом месте мы могли бы изменить роли полей, но для этого примера будем использовать их, как есть.

1. Откройте вкладку Опции построения.

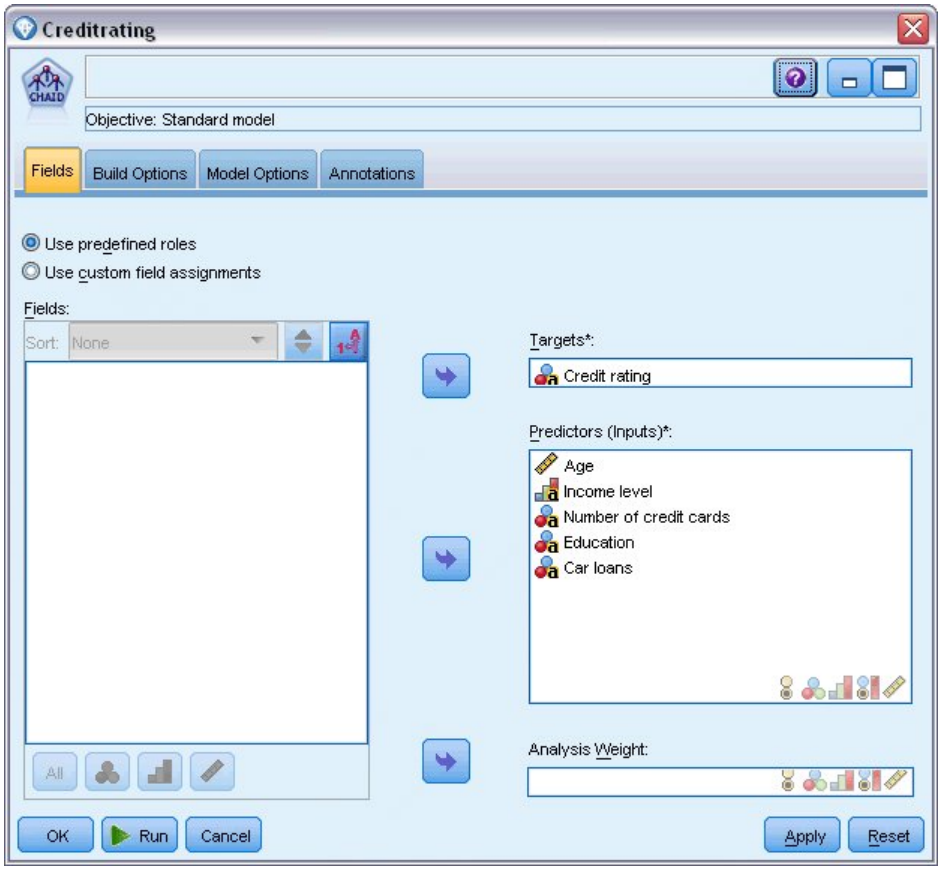


Рисунок 15. Вкладка Поля узла Моделирование CHAID

Здесь есть несколько опций, с помощью которых можно задать тип нужной модели для построения.

Мы хотим получить еще не применявшуюся модель, поэтому будем использовать опцию по умолчанию **Построить новую модель**.

Также мы хотим получить просто одну стандартную модель дерева решений без каких-то усовершенствований, поэтому оставим опцию цели по умолчанию - **Построить одно дерево**.

Хотя дополнительно мы можем запустить сеанс интерактивного моделирования, позволяющий уточнить модель, в этом примере генерируется модель с использованием режима по умолчанию - **Построить модель**.

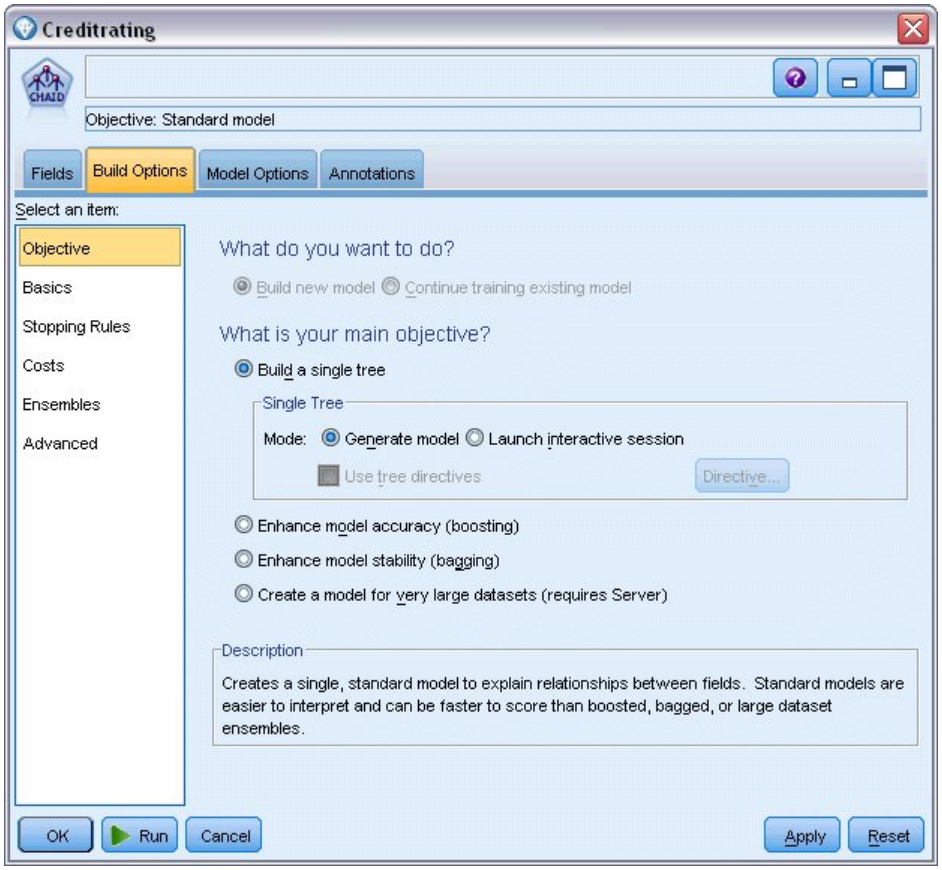


Рисунок 16. Вкладка Опции построения узла Моделирование CHAID

В этом примере мы хотим построить понятное и простое дерево, поэтому ограничим рост дерева минимальным числом наблюдений для родительских и дочерних узлов.

2. На вкладке Опции построения выберите **Правила остановки** слева на панели навигатора.
3. Выберите опцию **Использовать абсолютное значение**.
4. Задайте для параметра **Минимальное число записей в родительской ветви** значение 400.
5. Задайте для параметра **Минимальное число записей в дочерней ветви** значение 200.

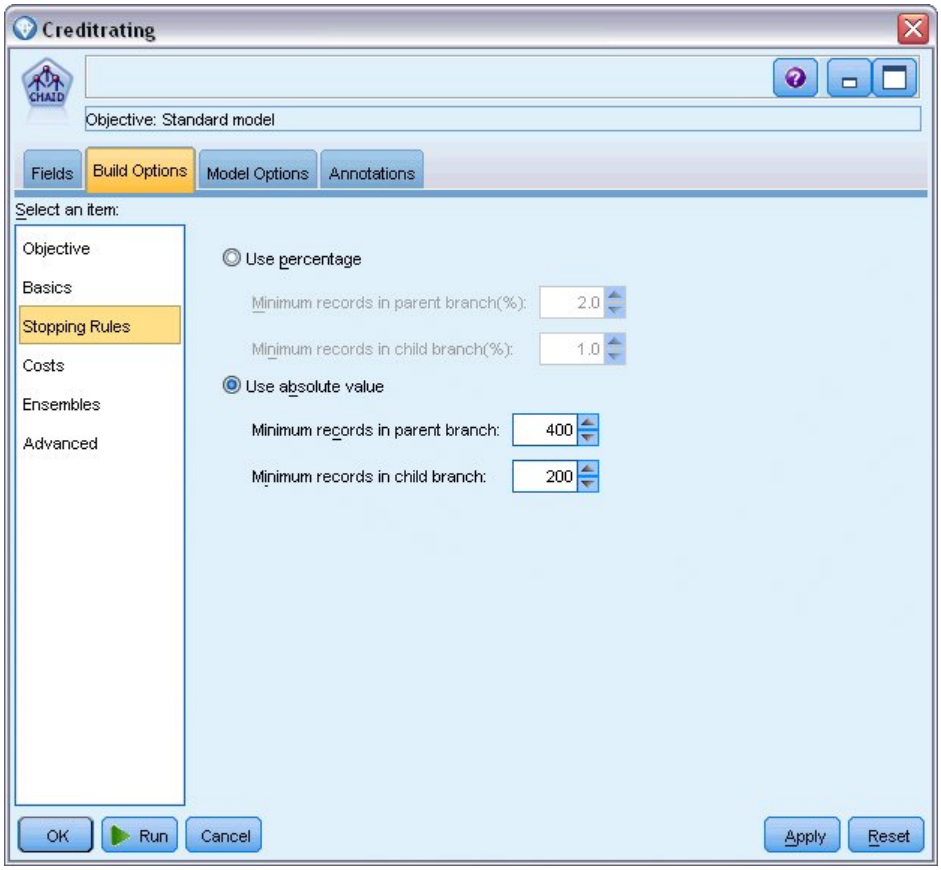


Рисунок 17. Задание критериев остановки для построения дерева решений

Для этого примера значения всех других опций мы можем использовать по умолчанию, поэтому нажмите кнопку **Выполнить**, чтобы создать модель. (Возможные варианты: щелкните правой кнопкой мыши по узлу и в контекстном меню выберите **Выполнить** или выберите узел и перейдите в меню Инструменты к пункту **Выполнить**).

Просмотр модели

После завершения выполнения слепок модели добавляется на палитру Модели в верхнем правом углу окна прикладных программ, а также размещается на холсте потока со ссылкой на узел моделирования, где он был создан. Для просмотра подробностей модели щелкните правой кнопкой мыши по слепку модели и выберите **Просмотр** (на палитре моделей) или **Изменить** (на холсте).

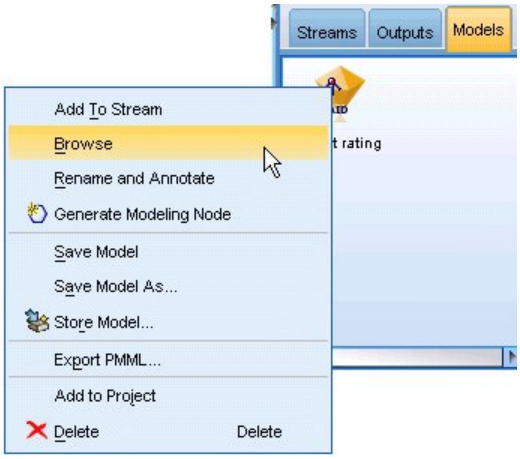


Рисунок 18. Палитра моделей

В случае слепка CHAID на вкладке Модель подробности выводятся в виде набора правил, и важно, что приводится ряд правил, которые можно использовать для назначения индивидуальных записей дочерним узлам на основе значений различных входных полей.

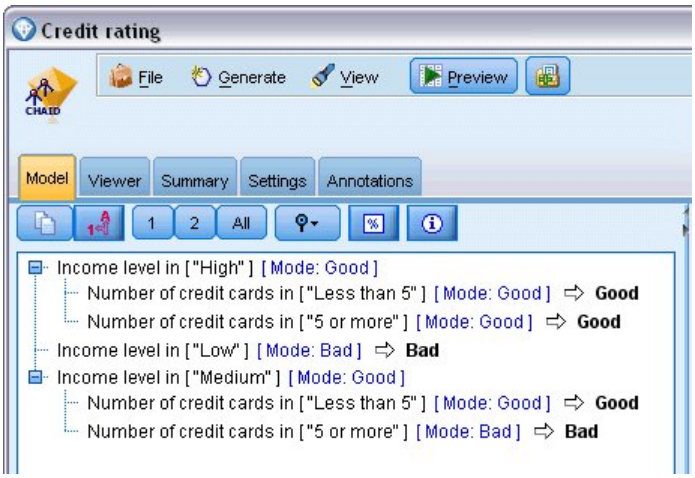


Рисунок 19. Слепок модели CHAID, набор правил

Для каждого конечного узла дерева решений, то есть для такого узла дерева, который нельзя расщепить дальше, возвращается предсказание *Хороший* или *Плохой*. В каждом случае предсказание определяется **модой**, или наиболее частым ответом, для записей, попавших на этот узел.

Справа от набора правил на вкладке Модель выводится диаграмма Важности, которая показывает относительную важность каждого предиктора при оценке модели. Отсюда легко видеть, что *Уровень дохода* - наиболее важный предиктор в данном случае, и есть еще только один важный фактор *Количество кредитных карт*.

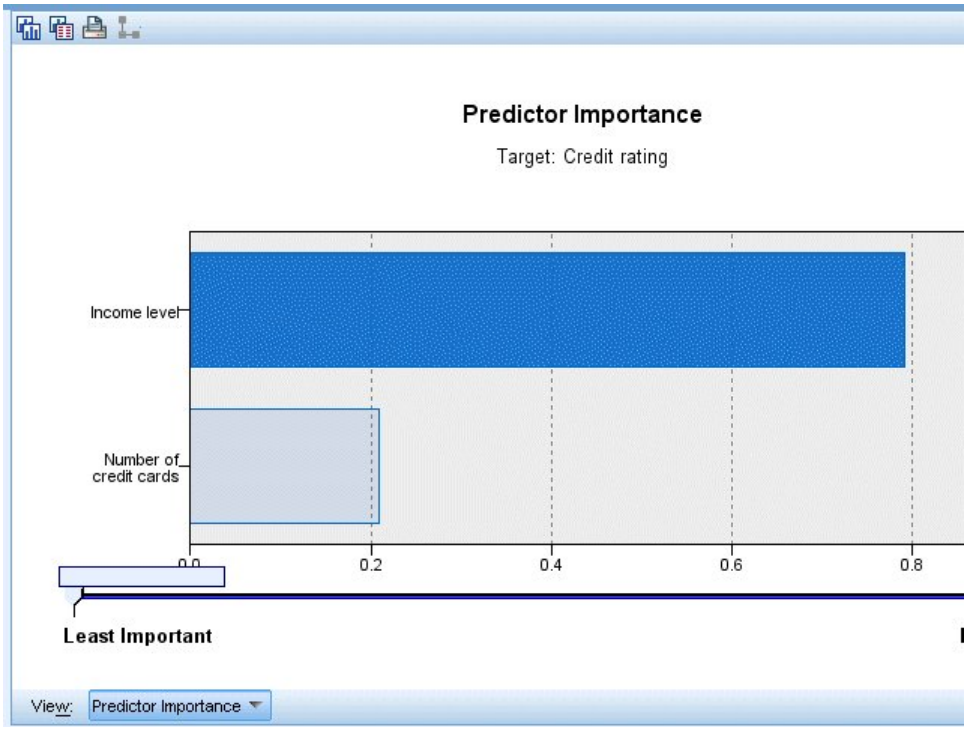


Рисунок 20. Диаграмма важности предикторов

На вкладке Программа просмотра слепка модели та же модель выводится в форме дерева с узлом в каждой точке решения. Используйте управляющие элементы масштабирования на панели инструментов, чтобы приблизить конкретный узел или уменьшить детализацию и увидеть большую часть дерева.

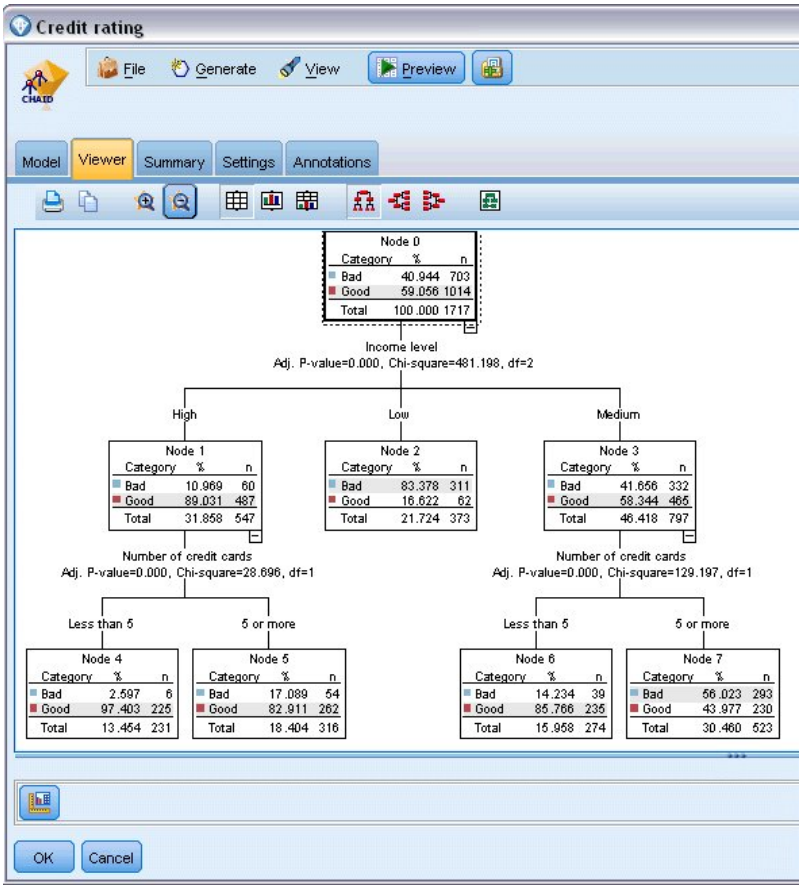


Рисунок 21. Вкладка Программа просмотра в слепке модели, выбор уменьшенной детализации

В верхней части дерева первый узел (Узел номер 0) дает сводку по всем записям в наборе данных. Более 40% наблюдений в наборе данных классифицируются как плохие (рискованные). Это очень большая часть, поэтому посмотрим, может ли дерево подсказать, какие факторы могут быть за это ответственны.

Видно, что первое расщепление производится по показателю *Уровень дохода*. Записи, для которых уровень дохода принадлежит категории *Низкий*, назначаются узлу 2, и неудивительно, что эта категория содержит максимальную процентную долю лиц, не выполняющих своих обязательств по кредитам. Это с очевидностью приводит к выводу, что клиенты в этой категории связываются с наибольшим риском.

Однако на самом деле 16% клиентов в этой категории *не* относятся к неплательщикам, то есть предсказание не всегда будет правильным. Никакая модель не может предсказать каждый отклик, но хорошая модель должна позволить вам предсказать *наиболее вероятный* отклик для каждой записи на основе доступных данных.

Аналогично, если посмотреть на клиентов с самым высоким доходом (Узел номер 1), мы увидим, что абсолютное большинство клиентов связаны с наименьшим риском (89%). Но более десятой части из них тоже не выполняли обязательств. Можно ли уточнить критерии, чтобы минимизировать для них риск?

Посмотрим, как модель разделила этих клиентов на две подкатегории (Узлы 4 и 5) на основании количества имеющихся у них кредитных карт. Если давать взаймы только клиентам с высоким доходом, у которых меньше пяти кредитных карт, успешность операции повысится с 89% до 97% и даже больше.

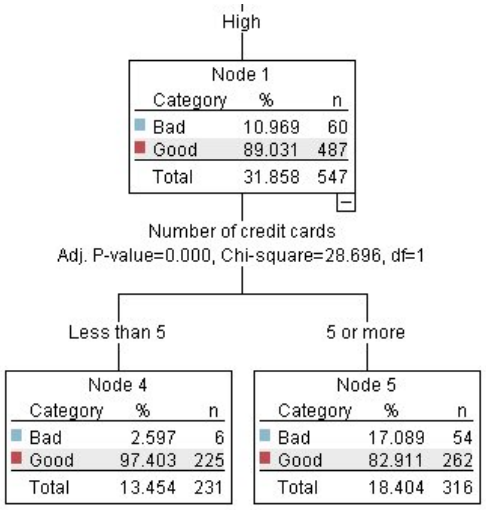


Рисунок 22. Представление дерева для клиентов с высоким доходом

Но что можно сказать о клиентах в категории дохода Средний (Узел 3)? Они более явно разделены между рейтингами Хороший и Плохой.

Нам и здесь могут помочь подкатегории (Узлы 6 и 7). На этот раз, ограничив займы только клиентами, у которых меньше пяти кредитных карт, мы увеличим показатель Хороших операций с 58% до 85%, существенное повышение.

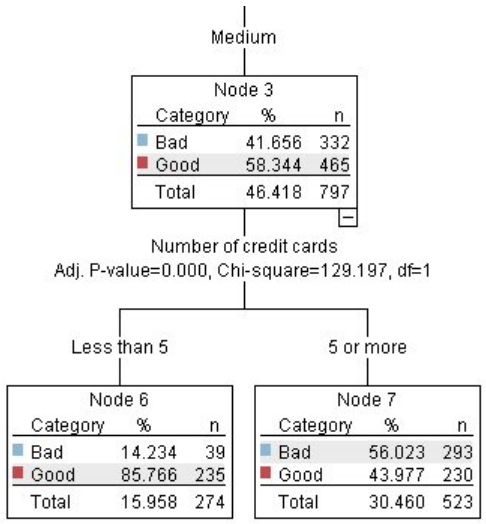


Рисунок 23. Представление дерева для клиентов со средним доходом

Итак, мы выяснили, что каждая запись, представляющая входную информацию модели, будет назначена конкретному узлу, и ей будет назначено предсказание *Хороший* или *Плохой* на основании наиболее частого отклика для этого узла.

Этот процесс назначения предсказаний индивидуальным записям известен как **скоринг**. Проводя скоринг для тех же записей, по которым оценивалась модель, мы можем выяснить, насколько точно он выполняется на данных обучения, то есть данных, для которых известен результат. Посмотрим, как это делается.

Оценка модели

Мы просмотрели модель, чтобы понять, как работает скоринг. Но для оценки, *насколько точно* от работает, нужно оценить некоторые записи и сравнить предсказанные моделью отклики с действительными результатами. Мы собираемся исследовать те же записи, которые использовались для оценки модели, что позволит сравнить наблюдаемые и предсказанные отклики.

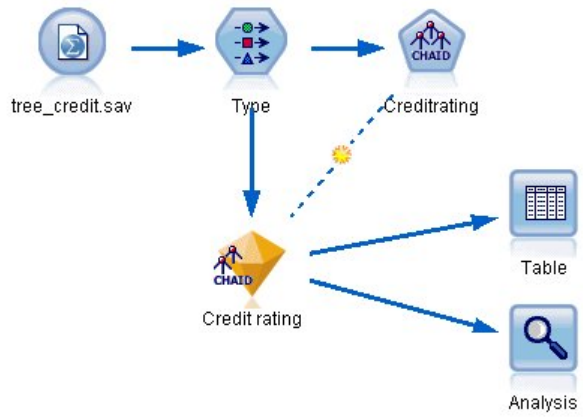
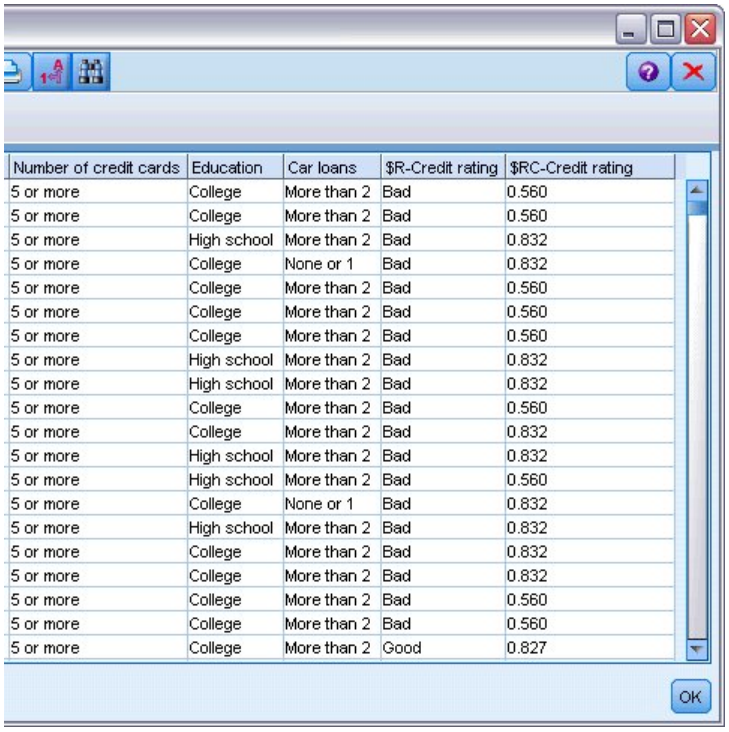


Рисунок 24. Присоединение слепка модели к узлу выходных данных для оценки модели

1. Чтобы увидеть оценки или предсказания, присоедините узел Таблица к слепку модели, дважды щелкните по узлу Таблица и нажмите кнопку **Выполнить**.

   Таблица покажет предсказанные оценки в поле с именем *$R-Кредитный рейтинг*, которое было создано моделью. Эти значения можно сравнить с исходным полем *Кредитный рейтинг*, которое содержит настоящие отклики.

   По соглашению имена сгенерированных при скоринге полей состоят из имени поля назначения со стандартным префиксом. Префиксы *$G* и *$GE* генерируются обобщенной линейной моделью; *$R* - это префикс, используемый для предсказания, генерируемого в данном случае моделью CHAID; *$RC* - для значения конфиденциальности, *$X* обычно генерируется при помощи ансамбля, а *$XR*, *$XS* и *$XF* используются в качестве префиксов в случаях, где поле назначения - непрерывное, категориальное, поле набора, или флаговое поле соответственно. Модели разных типов используют разные наборы префиксов. **Значение достоверности** - это собственная оценка модели в диапазоне от 0,0 до 1,0, насколько точно предсказано каждое значение.



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

OK

Рисунок 25. Таблица, показывающая сгенерированные оценки и значения достоверности

Как и ожидалось, предсказанное значение совпадает с фактическими откликами для многих записей, но не для всех. Причина этого в том, что у каждого конечного узла CHAID есть смесь откликов. Предсказание совпадает с *самым общим* откликом, но оно неправильно для всех остальных откликов этого узла. (Вспомним о 16%-ном меньшинстве клиентов с низким доходом, которые не отказывались выполнять обязательства по кредиту).

Для исключения этой ситуации мы можем продолжить расщепление дерева на всё меньшие и меньшие ветви, пока на каждом узле не окажется по 100% абсолютно *Хороших* или абсолютно *Плохих* клиентов без примеси других откликов. Но такая модель может быть чрезвычайно усложненной и скорее всего не будет хорошо обобщаться на другие наборы данных.

Чтобы точно подсчитать, сколько есть правильных предсказаний, мы можем пройти по таблице и учесть количество записей, для которых значение в предсказанном поле *$R-Кредитный рейтинг* совпадает со значением в поле *Кредитный рейтинг*. К счастью, есть гораздо более простой способ - использовать узел Анализ, который делает это автоматически.

2. Соедините слепок модели с узлом Анализ.
3. Дважды щелкните по узлу Анализ и нажмите кнопку **Выполнить**.

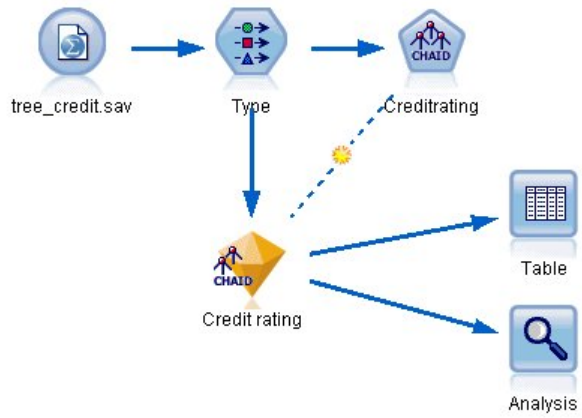


Рисунок 26. Присоединение узла Анализ

Анализ показывает, что для 1899 из 2464 записей (более 77%) предсказанное моделью значение совпадает с действительным откликом.

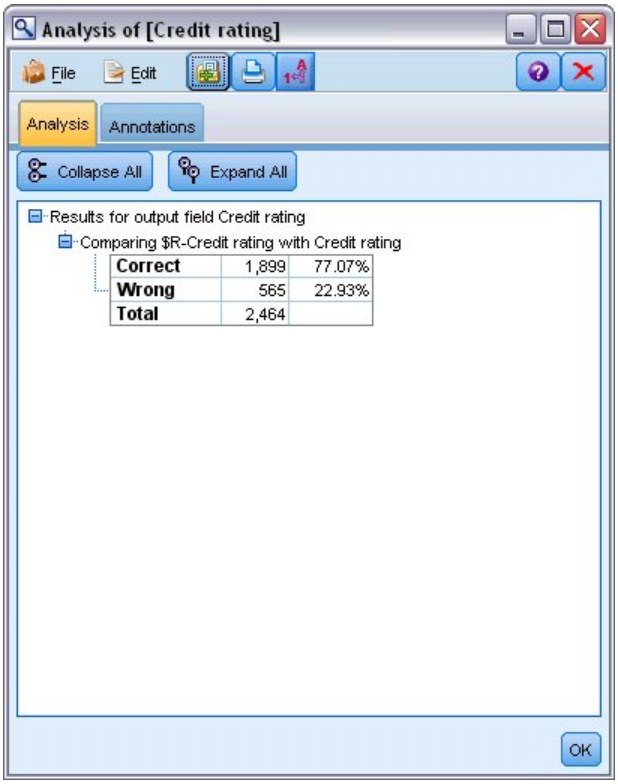


Рисунок 27. Результаты анализа, сравнивающие наблюдаемые и предсказанные отклики

Этот результат ограничен тем фактом, что оцениваемые записи были теми же, что использовались для оценки самой модели. В реальной ситуации можно было использовать узел Разделение, чтобы разбить данные на две отдельные выборки, для обучения и оценки.

Используя один раздел выборки для генерирования модели и другой - для ее испытания, можно получить гораздо более точный показатель, насколько хорошо модель обобщается на другие наборы данных.

Узел Анализ помогает испытать модель на записях, для которых мы уже знаем фактический результат. Следующая стадия иллюстрирует, как можно использовать модель, чтобы оценить записи, для которых мы

не знаем выходных значений. Например, сюда могут быть включены потенциальные клиенты, которые еще не работают с банком, но представляют из себя будущих получателей рекламной рассылки.

Скоринг записей

Ранее мы проводили скоринг тех же записей, которые использовались для оценки модели, чтобы понять, насколько точно построена модель. Теперь мы собираемся оценить другой набор записей, отличный от использованного для создания модели. Это цель моделирования с использованием поля назначения: изучить записи, для которых известны выходные данные, чтобы идентифицировать структуры, которые позволят предсказать еще не известные выходные данные.

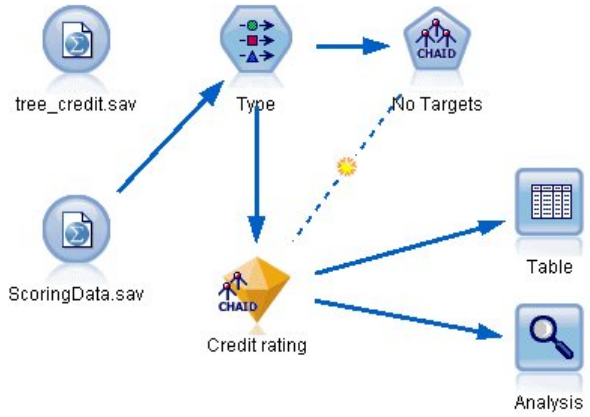


Рисунок 28. Присоединение новых данных для скоринга

Можно изменить узел источника Файл статистики для указания на другой файл данных или добавить новый узел источника, читающий данные, которые вы хотите оценить. В любом случае новый набор данных должен содержать те же входные поля, что использовала модель (*Возраст*, *Уровень дохода*, *Образование* и так далее), но не поле назначения *Кредитный рейтинг*.

Другой вариант - добавить слепок модели к любому потоку, включающему в себя ожидаемые входные поля. Тип источника (чтение из файла или базы данных) не имеет значения, если имена и типы полей совпадают с используемыми в модели.

Можно сохранить также слепок модели как отдельный файл, экспортировать модель в формате PMML для использования с другими прикладными программами, использующими этот формат, или сохранить эту модель в репозитории IBM SPSS Collaboration and Deployment Services, который на уровне предприятия обеспечивает внедрение, скоринг и управление моделями.

Сама модель независимо от используемой инфраструктуры работает одинаково.

Итог

Этот пример демонстрирует основные шаги по созданию, исследованию качества и скорингу модели.

- Узел моделирования оценивает модель, изучая записи, для которых известны выходные данные, и создает слепок модели. Иногда это называется обучением модели.
- Слепок модели можно добавить к любому потоку с ожидаемыми полями для скоринга записей. По скорингу записей, для которых известны выходные данные (например, для существующих клиентов), можно оценить, насколько хорошо все работает.
- После того, как вы будете удовлетворены качеством модели, можно оценивать новые данные (например, возможных клиентов) для предсказания их отклика.

- Данные, которые использовались для обучения или оценки модели, можно назвать аналитическими или хронологическими данными; данные скоринга можно назвать также операционными.

Глава 4. Автоматическое моделирование для флагового поля назначения

Моделирование ответа покупателя (автоматический классификатор)

Узел Автоклассификация позволяет автоматически создавать и сравнивать несколько различных моделей с флаговыми (например, вероятно ли, что данный клиент решит взять кредит или откликнется на конкретное предложение) или номинальными (из набора) назначениями. В данном примере нас будет интересовать флаговый результат (да или нет). В относительно простом потоке узел генерирует и оценивает ряд моделей-кандидатов, выбирает те, который работают лучше, и комбинирует их в единую объединенную модель. Такой подход сочетает простоту автоматизации с выгодами использования нескольких моделей для получения более точных предсказаний по сравнению с получаемыми от любой одной модели.

В этом примере используется вымышленная компания, которая хочет достичь более выгодных результатов, подбирая правильные предложения для каждого из покупателей.

Этот подход подчеркивает выгоды автоматизации. Подобный пример с непрерывным (в диапазоне чисел) назначением смотрите в разделе Значения свойств (автонумерация).

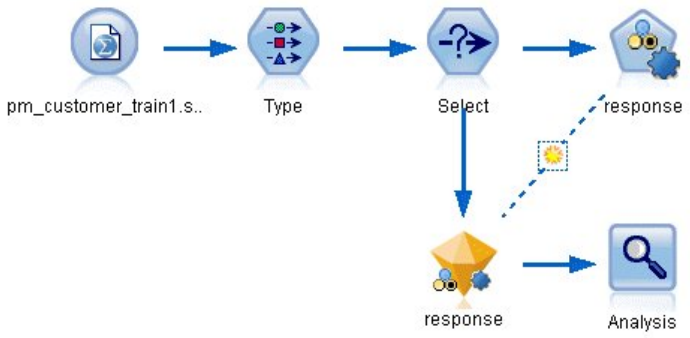


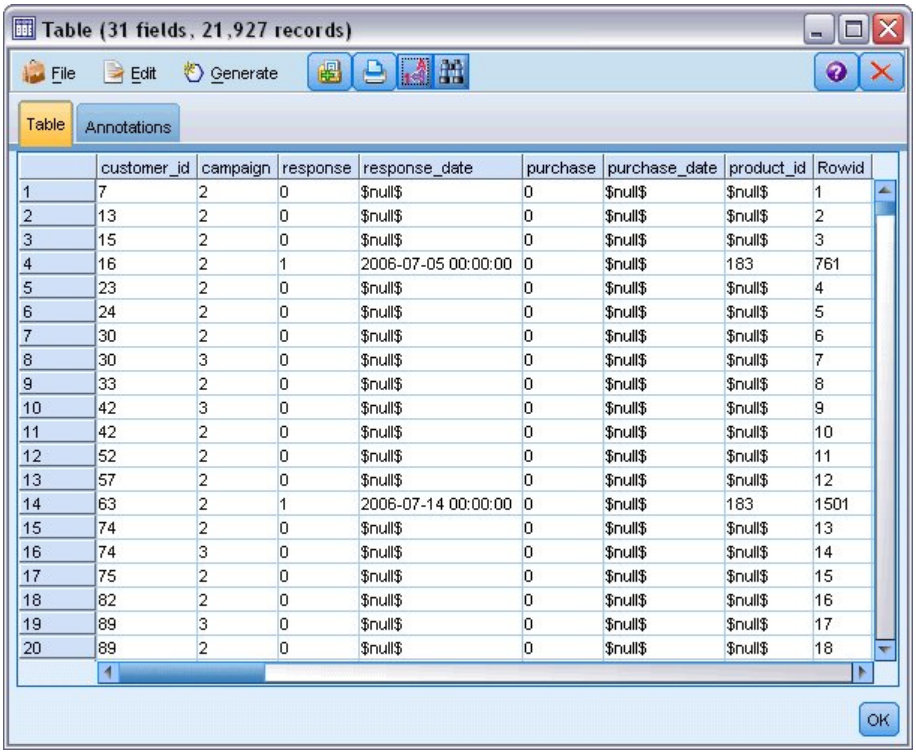
Рисунок 29. Поток примера автоклассификации

Этот пример использует поток *pm_binaryclassifier.str*, установленный в подпапке Demo папки *streams*. Используется файл данных *pm_customer_train1.sav*. Дополнительную информацию смотрите в разделе “Хронологические данные”.

Хронологические данные

В файле *pm_customer_train1.sav* есть хронологические данные, отслеживающие предложения, сделанные для конкретных покупателей в прошлых кампаниях, на что указывает значение в поле *кампания*. Наибольшее число записей подпадает под кампанию *Премиальная учетная запись*.

Значения поля *campaign* фактически кодируются целыми числами в данных (например, *2 = Премиальная учетная запись*). Позже вы определите метки для этих значений, которые можно будет использовать для получения более осмысленного вывода.



	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

Рисунок 30. Данные о предыдущих спецпредложениях

Этот файл содержит также поле *response*, указывающее, было ли предложение принято (0 = *нет*, 1 = *да*). Это будет **полем назначения**, то есть значением, которое мы хотим предсказать. Включены также много полей, содержащих демографическую и финансовую информацию о каждом клиенте. Их можно использовать, чтобы построить или "обучить" модель, предсказывающую показатели отклика для отдельных людей или групп на основе таких характеристик, как доход, возраст или число транзакций в месяц.

Построение потока

1. Добавьте узел источников файлов Statistics, указав файл *pm_customer_train1.sav*, расположенный в папке *Demos* вашего каталога установки IBM SPSS Modeler. (Можно задать `$CLEO_DEMOS/` в пути файла как быстрый вызов ссылки на эту папку. Обратите внимание на то, что нужно использовать обычную, а не обратную дробную черту, как показано в примере.)

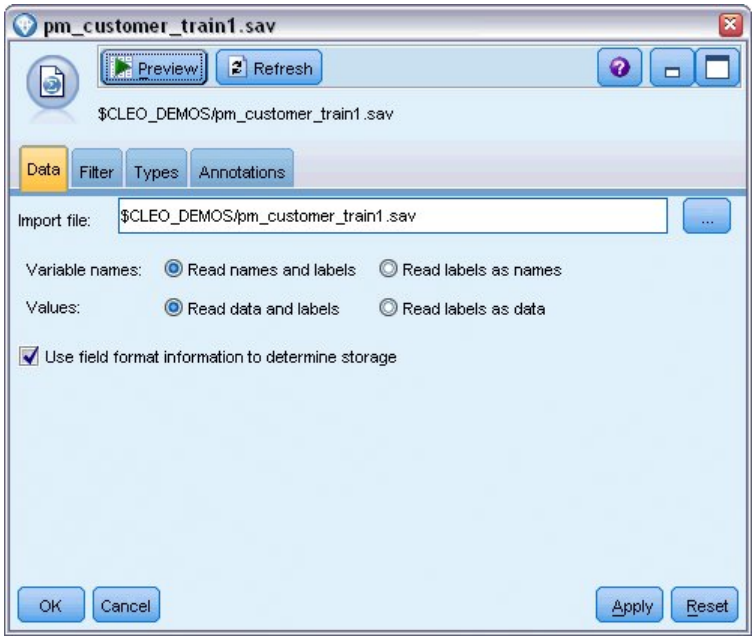


Рисунок 31. Чтение в данных

2. Добавьте узел Тип и выберите *отклик* как целевое поле (Роль = **Назначение**). Задайте для измерения этого поля значение **Флаг**.

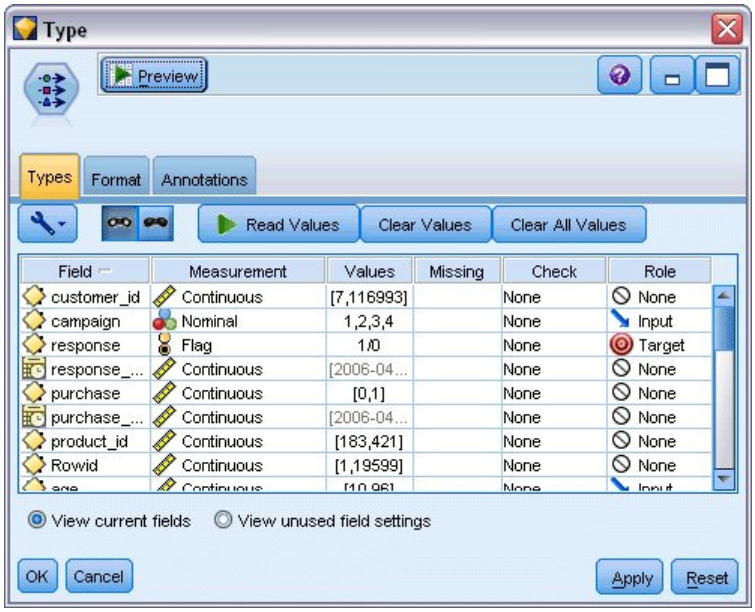


Рисунок 32. Задание уровня измерения и роли

3. Для следующих полей задайте роль **Нет**: *customer_id*, *campaign*, *response_date*, *purchase*, *purchase_date*, *product_id*, *Rowid* и *X_random* (ID заказчика, кампания, дата отклика, закупка, дата закупки, ID товара, ID строки и X_random). При построении модели эти поля будут игнорироваться.
4. Нажмите кнопку **Прочитать значения** в узле Тип, чтобы гарантировать инициирование всех значений.

   Как было показано ранее, наши данные источника включают в себя информацию о разных кампаниях, каждая из которых была нацелена на свой тип учетной записи клиента. В данных эти кампании кодируются целыми числами, поэтому для простоты запоминания, какой тип учетной записи какое

целое число представляет, определим метки для каждого типа учетной записи.

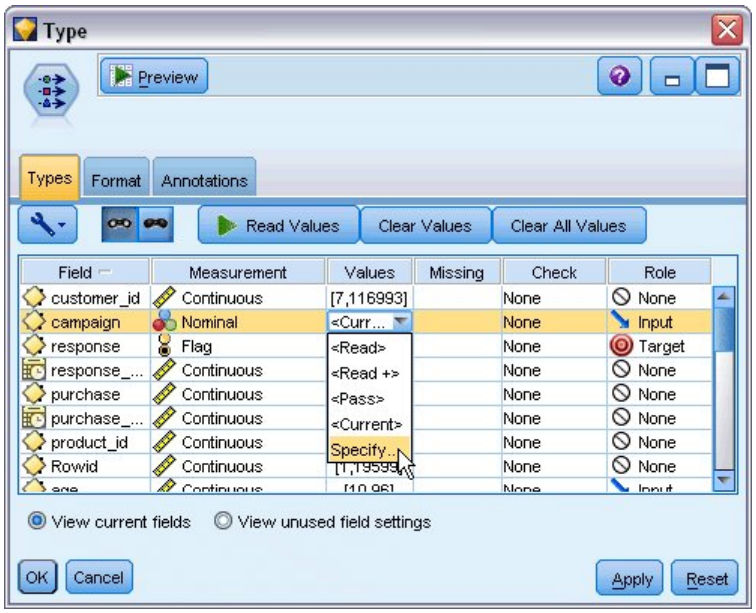


Рисунок 33. Выбор задаваемых в поле значений

5. В строке для поля **campaign** щелкните по записи в столбце **Значения**.
6. В выпадающем списке выберите опцию **Задать**.

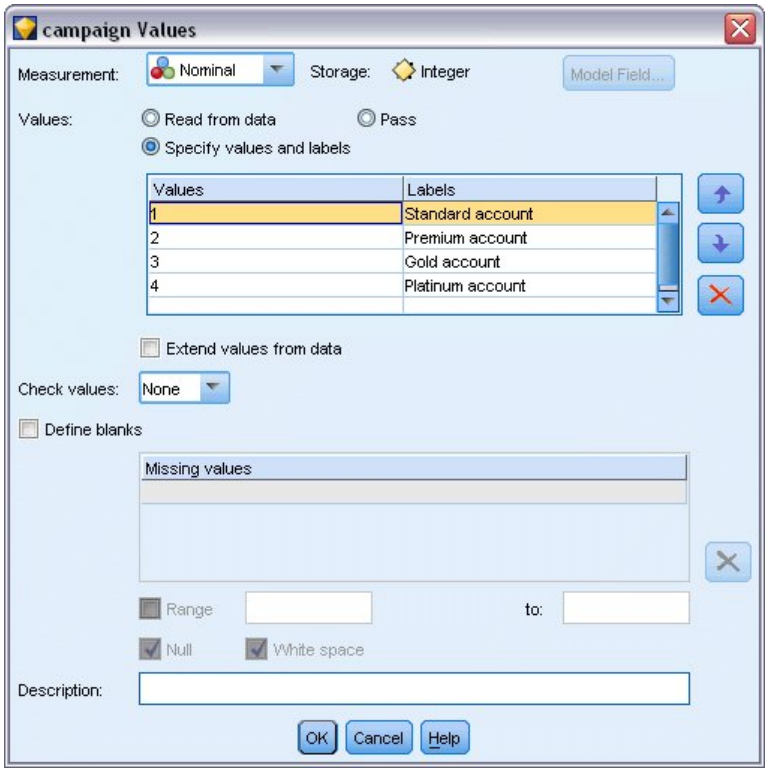
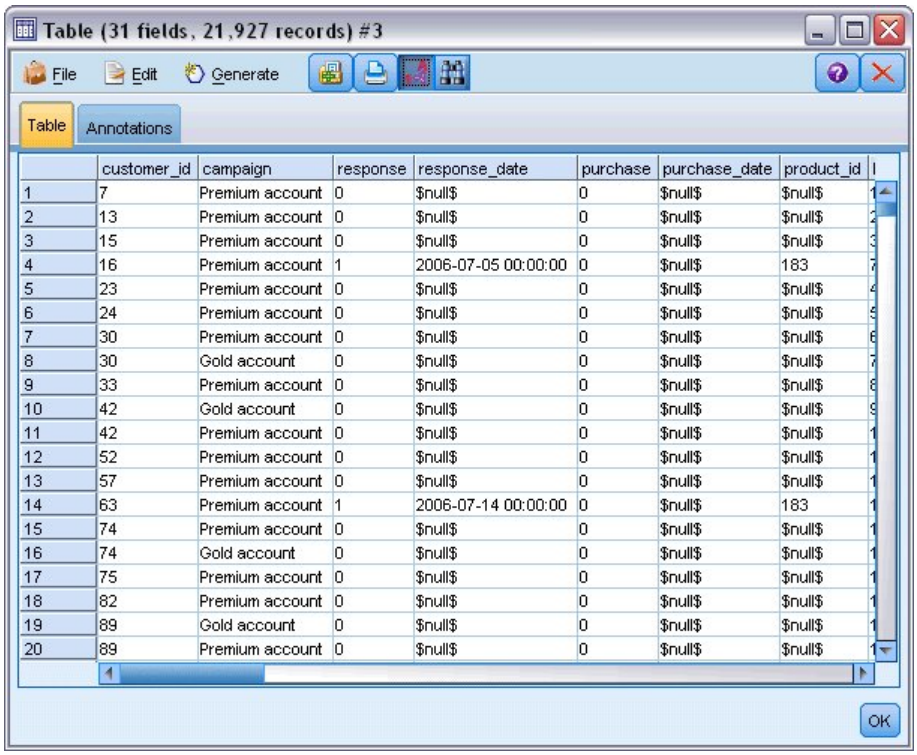


Рисунок 34. Определение меток для значений полей

7. В столбце **Метки** введите метки, как это показано для каждого из четырех значений поля **campaign**.

8. Щелкните по **OK**.

   Теперь в окнах вывода вместо целых чисел можно показывать метки.



	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

Рисунок 35. Вывод меток значений полей

9. Присоедините узел Таблица к узлу Тип.
10. Откройте узел Таблица и щелкните по **Выполнить**.
11. В окне вывода нажмите кнопку панели инструментов **Показать поле и метки значений**, чтобы вывести метки.
12. Нажмите кнопку **OK**, чтобы закрыть окно вывода.

Хотя эти данные включают в себя информацию о четырех разных кампаниях, вы будете всякий раз фокусироваться на одной кампании. Так как наибольшее число записей относится к кампании для учетных записей Премиум (в данных закодированы как *campaign=2*), можно использовать узел выбора для включения в поток только таких записей.

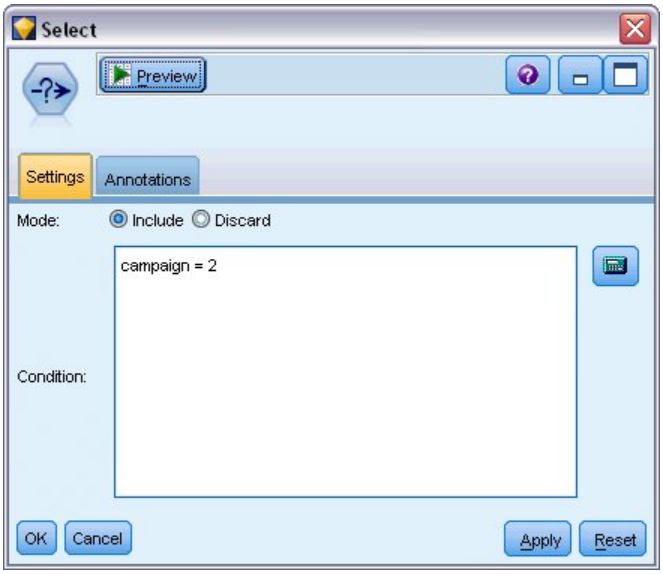


Рисунок 36. Выбор записей для одной кампании

Создание и сравнение моделей

1. Присоедините узел автоклассификации и выберите **Общую точность** как используемый для ранжирования моделей показатель.
2. Задайте для **Числа используемых моделей** значение 3. Это означает, что при выполнении узла будет построено три лучших модели.

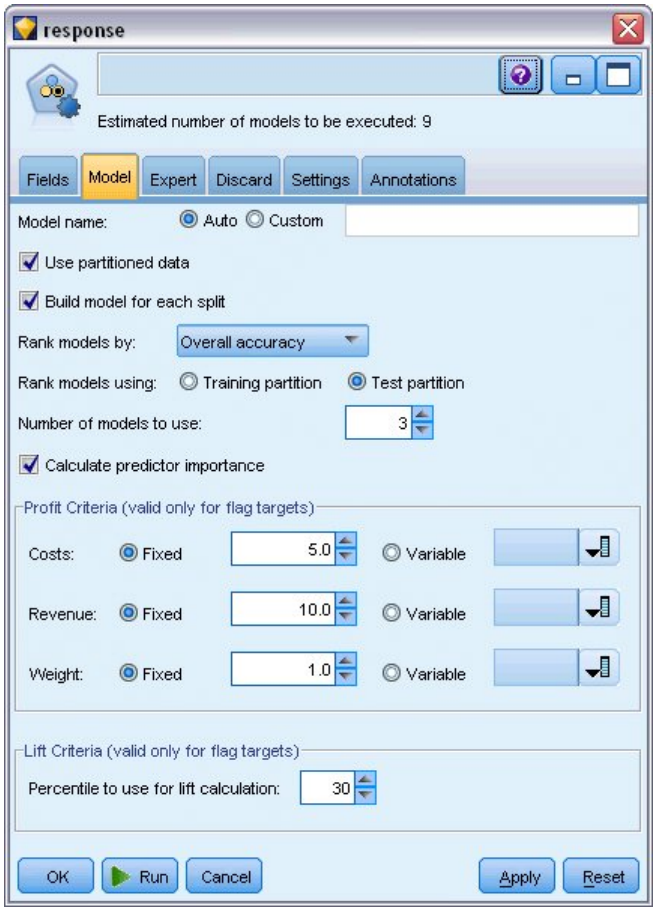


Рисунок 37. Вкладка Модели узла автоклассификации

На вкладке Эксперт можно выбрать от одного до 11 разных алгоритмов модели.

3. Отмените выбор типов моделей **Дискриминант** и **SVM**. (Для этих моделей требуется большее время на обучение с этими данными, поэтому отмена их выбора ускорит выполнение примера. Если время выполнения не важно, конечно, эти модели можно оставить выбранными.)

   Так как на вкладке Модель для **Числа используемых моделей** задано значение 3, этот узел может вычислить точность остальных девяти алгоритмов и построить один слепок модели, содержащий три наиболее точных модели.

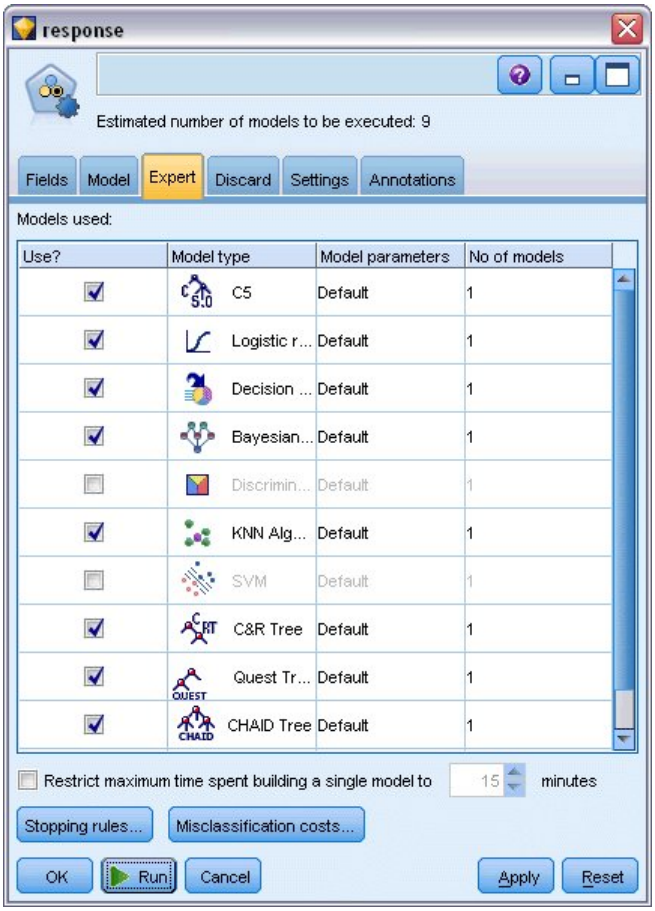


Рисунок 38. Узел автоклассификации: Вкладка Эксперт

4. На вкладке Параметры выберите для метода ансамбля **Голосование со взвешенными доверительными вероятностями**. Этим определяется, как для каждой записи будет производиться одна агрегированная оценка.

   При простом голосовании, если две из трех моделей предсказывают *да*, *да* побеждает со счетом 2 к 1. В случае голосования со взвешенными доверительными вероятностями голоса взвешиваются на основе значения доверительной вероятности для каждого предсказания. Таким образом, если одна модель предсказывает *нет* с большей доверительной вероятностью, чем два предсказания *да* вместе, победит *нет*.

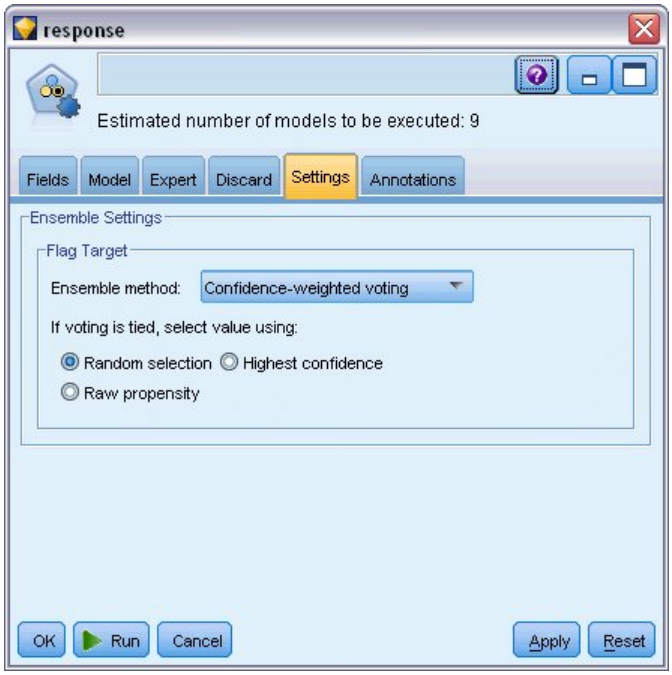


Рисунок 39. Узел автоклассификации: Вкладка Параметры

5. Нажмите кнопку **Выполнить**.

Через несколько минут сгенерированный слепок будет построен и размещен на холсте и на палитре Модели в правом верхнем углу окна. Этот слепок модели можно просмотреть, сохранить или внедрить каким-то еще способом.

Откройте слепок модели; здесь перечислены подробности о каждой из моделей, созданных во время выполнения. (В реальной ситуации, когда для большого набора данных создаются сотни моделей, это может занять много часов.) Смотрите рис. 29 на стр. 39.

Если вы хотите исследовать какую-то из моделей более подробно, можно дважды щелкнуть по значку слепка модели в столбце **Модель** для расширении детализации и просмотра результатов отдельной модели; здесь можно сгенерировать узлы моделирования, слепки моделей или диаграммы оценки. В столбце **Диаграмма** можно дважды щелкнуть по миниизображению, чтобы сгенерировать полноразмерную диаграмму.

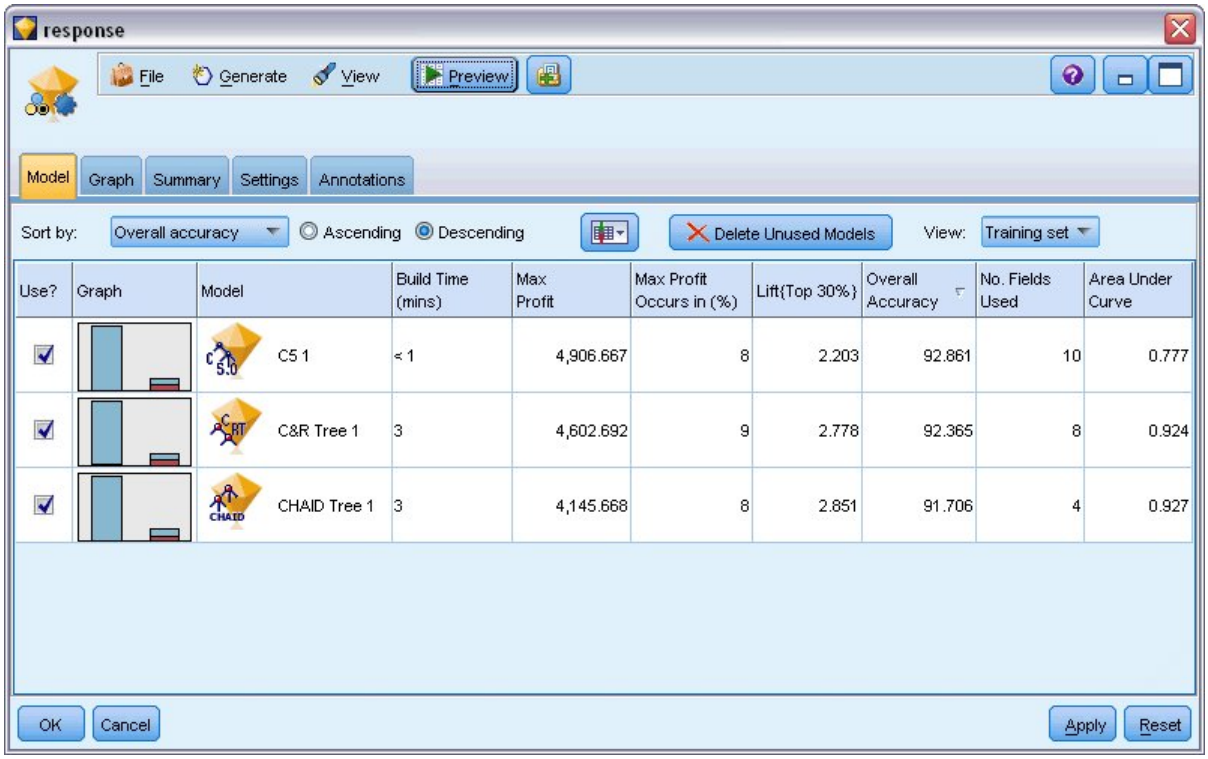


Рисунок 40. Результаты автоклассификации

По умолчанию модели сортируются на основании общей точности, так как это был показатель, выбранный на вкладке Модель узла автоклассификации. Модель C51 ранжируется по этому показателю как лучшая, но модели Дерево C&R и CHAID дают примерно такую же точность.

Сортировку в другом столбце можно выполнить, щелкнув по его заголовку или выбрав нужный показатель в в выпадающем списке **Сортировать по** на панели инструментов.

На основе этих результатов вы решаете использовать все три из этих самых точных моделей. Объединяя в сочетание предсказания из нескольких моделей, можно избежать ограничений в отдельных моделях, что приведет к более высокой общей точности.

В столбце **Использовать?** выберите модели C51, Дерево C&R и CHAID.

Присоедините узел Анализ (палитра Вывод) после слепка модели. Щелкните правой кнопкой по узлу Анализ и выберите **Запуск**, чтобы выполнить поток.

Агрегированная оценка, сгенерированная объединенной моделью, показана в поле *$XF-response*. При измерениях на данных обучения предсказанные значения совпадают с действительным ответом (как он записан в исходном поле *ответ*) с общей точностью 92,82%.

Хотя это значение не точно совпадает с наилучшим из трех отдельных моделей (92,86% для C51), это различие очень мало, чтобы быть значимым. Вообще говоря, обычно объединенная модель скорее всего будет хорошо работать при применении к наборам данных, отличным от данных обучения.

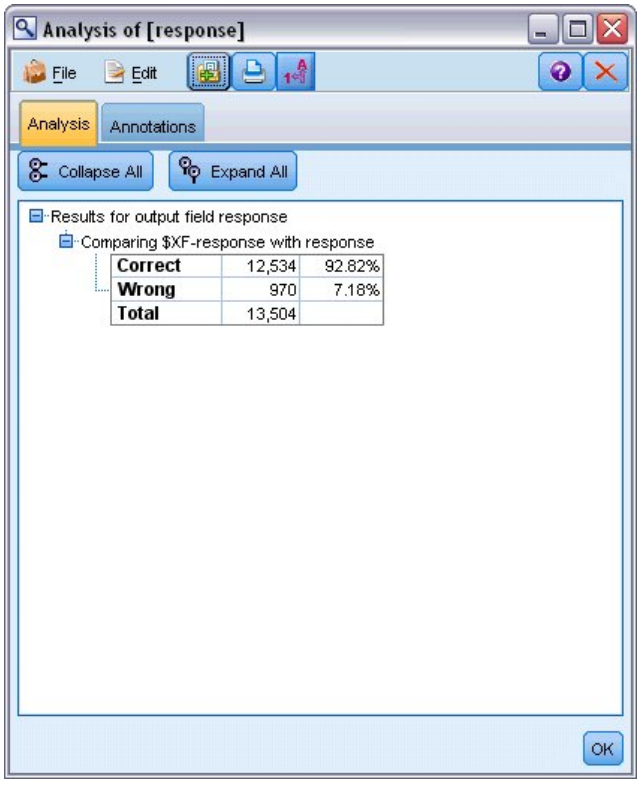


Рисунок 41. Анализ трех объединенных моделей

Итог

Итак, вы использовали узел Автоклассификатор, чтобы сравнить ряд различных моделей, использовали три самые точные и добавили их в поток в слепок модели ансамблей Автоклассификатора.

- С учетом общей точности для обучающих данных лучше всего себя показали модели C51, Дерево C&R и CHAID.
- Комбинированная модель работала почти также хорошо, как лучшие из отдельных моделей и, возможно, превзойдет их при использовании для других наборов данных. Если ваша цель - это возможно большая автоматизация процесса, этот подход позволяет получить устойчивую модель при многих обстоятельствах, не углубляясь в особенности каждой конкретной модели.

Глава 5. Автоматическое моделирование для количественного целевого поля

Стоимость имущества (автонумерация)

Узел Автонумерация позволяет автоматически создавать и сравнивать различные модели для непрерывных (в числовом диапазоне) выходных данных, например, предсказывать облагаемую налогом стоимость имущества. При помощи одного узла вы можете оценить и сравнить набор моделей-кандидатов и сгенерировать подмножество моделей для будущего анализа. Этот узел работает так же, как узел Автоклассификация, но с числовыми, а не номинальными полями назначения.

Этот узел объединяет лучшие из моделей-кандидатов в один слепок агрегированной (собранной в ансамбль) модели. Такой подход комбинирует простоту автоматизации с преимуществами объединения нескольких моделей, что часто приводит к более точным оценкам, чем получается из любой одной модели.

Предмет данного примера - работа вымышленного муниципалитета, ответственного за корректировку и оценку налогов на недвижимое имущество. Чтобы сделать это с большей точностью, они строят модель, предсказывающую значения стоимости имущества на основании типа зданий, района города, размера и других известных факторов.

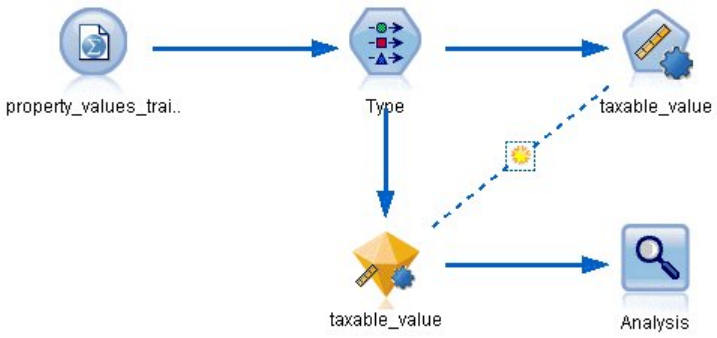


Рисунок 42. Поток примера автонумерации

В этом примере используется поток *property_values_numericpredictor.str*, установленный в подпапке *streams* папки Demos. Используется файл данных *property_values_train.sav*. Дополнительную информацию смотрите в разделе “Папка demos” на стр. 4.

Данные обучения

Файл данных содержит поле *taxable_value*, представляющее собой **поле назначения**, то есть значение, которое вы хотите предсказать. Остальные поля содержат такую информацию как район города, тип здания и внутренний объем и могут использоваться как предикторы.

Имя поля	**Метка**
property_id	ID владения
neighborhood	Район города
building_type	Тип здания
year_built	Год постройки
volume_interior	Внутренний объем
volume_other	Объем гаража и дополнительных построек

Имя поля	Метка
lot_size	Размер лота
taxable_value	Налогооблагаемая ценность

Файл данных скоринга *property_values_score.sav* также содержится в папке Demos. Он содержит те же поля, кроме поля *taxable_value*. После обучения моделей с помощью набора данных, в котором налогооблагаемая ценность известна, можно выполнить скоринг записей, для которых это значение еще неизвестно.

Построение потока

1. Добавьте узел источников файлов Statistics, указывающий на файл *property_values_train.sav* в папке *Demos* вашего каталога установки IBM SPSS Modeler. (Можно задать `$CLEO_DEMOS/` в пути файла как быстрый вызов ссылки на эту папку. Обратите внимание на то, что нужно использовать обычную, а не обратную дробную черту, как показано в примере.)

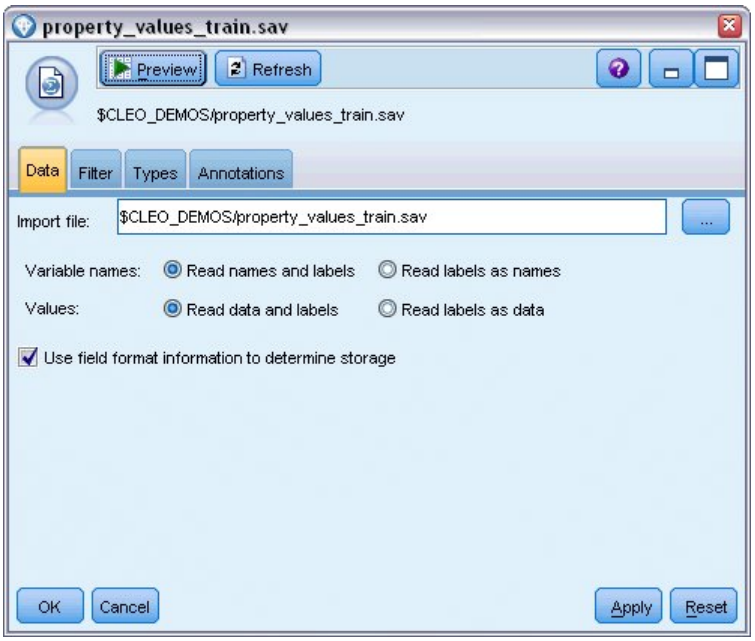


Рисунок 43. Чтение в данных

2. Добавьте узел Тип и выберите *taxable_value* (налогооблагаемая ценность) как целевое поле (Роль = **Назначение**). Для роли всех остальных полей должно быть задано значение **Входное**, указывающее на их использование в качестве предикторов.

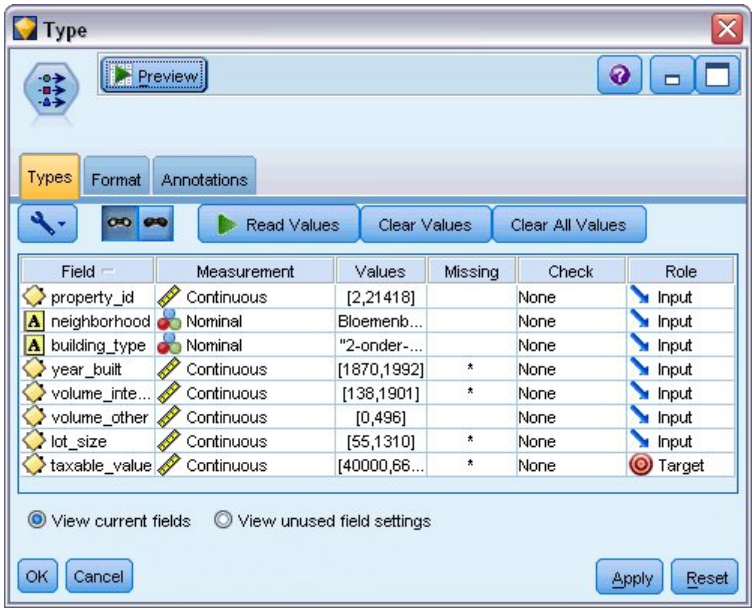


Рисунок 44. Задание поля назначения

3. Присоединитесь к узлу Автонумерация и выберите **Корреляцию** как показатель, используемый для ранжирования моделей.
4. Задайте для **Числа используемых моделей** значение 3. Это означает, что при выполнении узла будет построено три лучших модели.

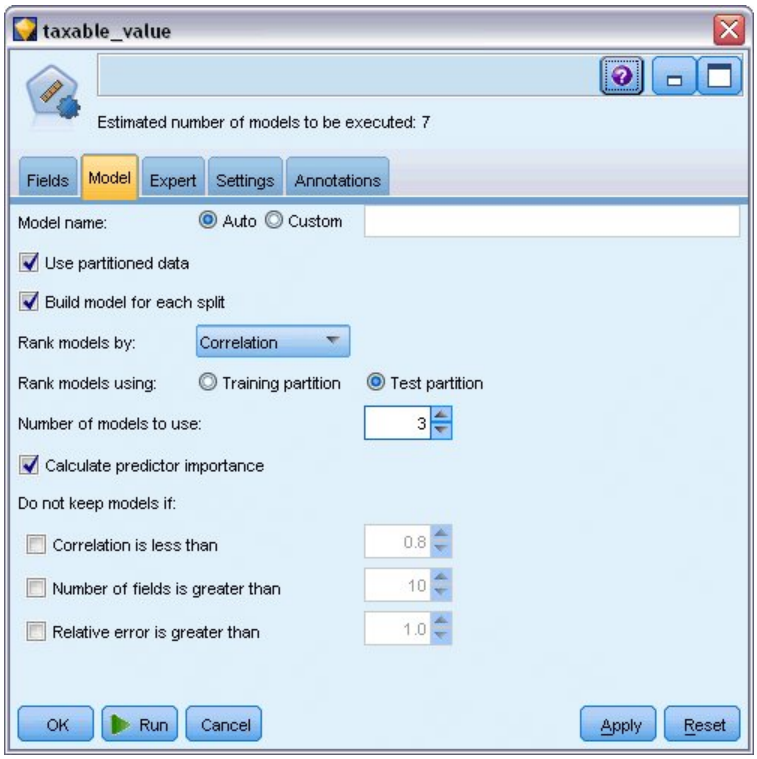


Рисунок 45. Вкладка Модель узла автонумерации

5. На вкладке Эксперт оставьте параметры по умолчанию на месте; узел будет оценивать одну модель для каждого алгоритма, всего до семи моделей. (Вместо этого можно изменить данные параметры, чтобы сравнить несколько вариантов для каждого типа модели.)

   Так как на вкладке Модель для **Числа используемых моделей** задано значение 3, этот узел может вычислить точность семи алгоритмов и построить один слепок модели, содержащий три наиболее точные модели.

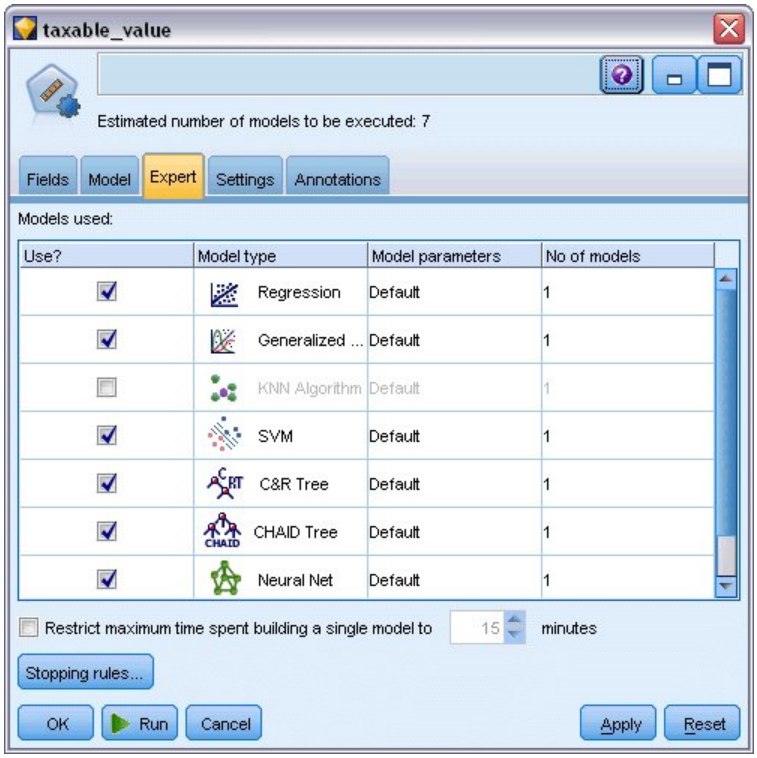


Рисунок 46. Вкладка Эксперт узла автонумерации

6. На вкладке Параметры оставьте на месте параметры по умолчанию. Так как это числовое поле назначения, оценка ансамбля генерируется при усреднении оценок для индивидуальных моделей.

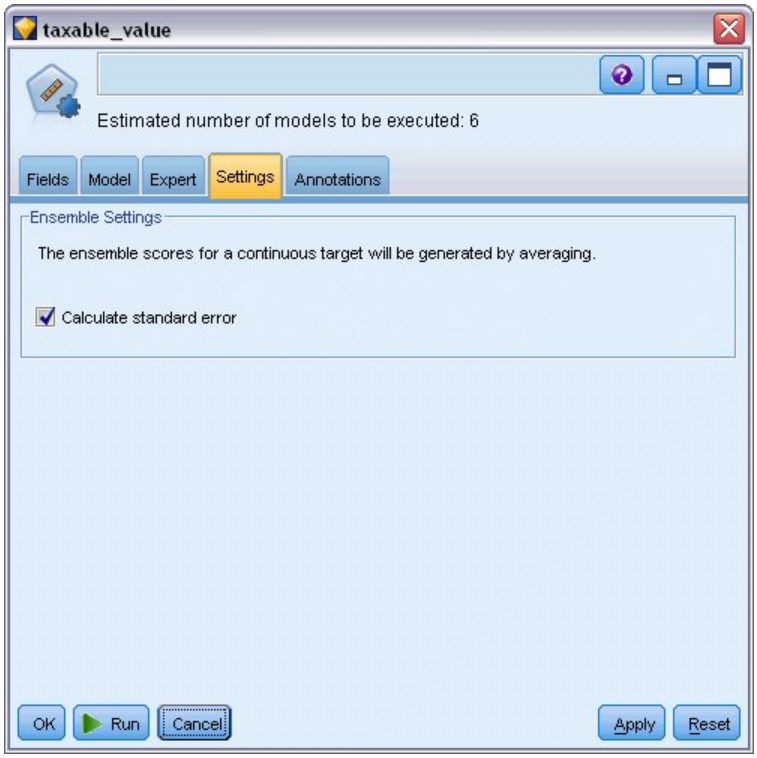


Рисунок 47. Вкладка Параметры узла автонумерации

Сравнение моделей

1. Нажмите кнопку Выполнить.

Слепок модели будет построен и размещен на холсте и на палитре Модели в правом верхнем углу окна. Этот слепок модели можно просмотреть, сохранить или внедрить каким-либо еще способом.

Откройте слепок модели; здесь перечислены подробности о каждой из моделей, созданных во время выполнения. (В реальной ситуации, когда для большого набора данных создаются сотни моделей, это может занять много часов.) Смотрите рис. 42 на стр. 51.

Если вы хотите исследовать какую-то из моделей более подробно, можно дважды щелкнуть по значку слепка модели в столбце **Модель** для расширении детализации и просмотра результатов отдельной модели; здесь можно сгенерировать узлы моделирования, слепки моделей или диаграммы оценки.

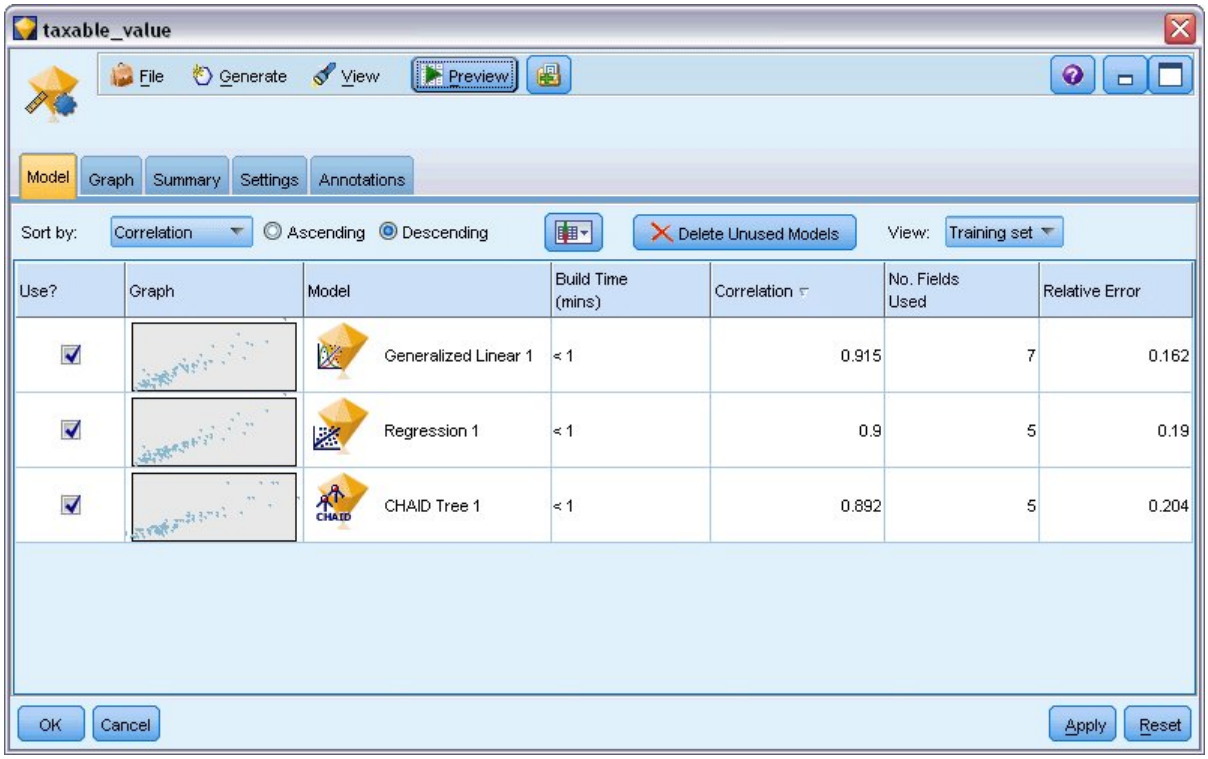


Рисунок 48. Результаты автонумерации

По умолчанию модели сортируются по корреляции, так как это был показатель, выбранный в узле Автонумерация. Для целей ранжирования используется абсолютное значение коэффициента корреляции, при этом более близкие к единице значения соответствуют более сильной взаимосвязи. Обобщенная линейная модель ранжируется как лучшая по этому показателю, но некоторые другие модели почти столь же точны. Кроме этого, у обобщенной линейной модели самая маленькая относительная ошибка.

Сортировку по другому столбцу можно выполнить, щелкнув по его заголовку или выбрав нужный показатель в выпадающем списке **Сортировать по** на панели инструментов.

На каждом графике отложены зависимости наблюдаемых значений от предсказанных для модели, что обеспечивает быстрое наглядное представление корреляций между ними. Для хорошей модели точки графика должны группироваться вдоль диагонали, что справедливо для всех моделей в этом примере.

В столбце **Диаграмма** можно дважды щелкнуть по миниизображению, чтобы сгенерировать полноразмерную диаграмму.

На основе этих результатов вы решаете использовать все три из этих самых точных моделей. Объединяя в сочетание предсказания из нескольких моделей, можно избежать ограничений в отдельных моделях, что приведет к более высокой общей точности.

Убедитесь, что в столбце **Использовать?** выбраны все три модели.

Присоедините узел Анализ (палитра Вывод) после слепка модели. Щелкните правой кнопкой по узлу Анализ и выберите **Запуск**, чтобы выполнить поток.

Сгенерированная объединенной моделью средняя оценка будет добавлена в поле *$XR-taxable_value* с коэффициентом корреляции 0,922, превышающим этот коэффициент для трех отдельных моделей. Оценки ансамбля показывают также небольшую среднюю абсолютную ошибку и могут дать лучшие значения при

применении к другим наборам данных, чем любая из моделей по отдельности.

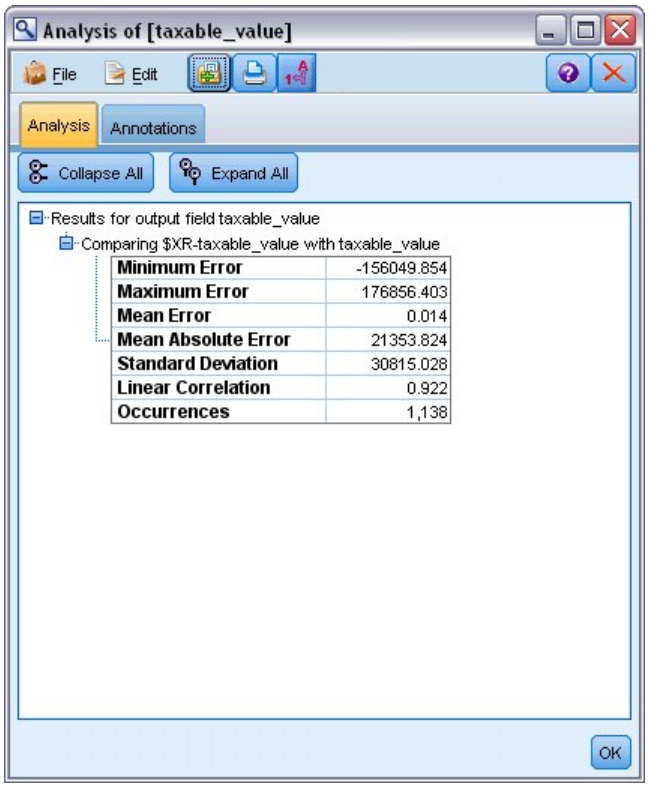


Рисунок 49. Поток примера автонумерации

Итог

Итак, вы использовали узел Автонумерация для сравнения нескольких разных моделей, выбрали три наиболее точные модели и добавили их в поток в составе слепка объединенной модели Автонумерация.

- Судя по общей точности, модели Обобщенная линейная, Регрессия и CHAID лучше всего выполняются на данных обучения.
- Комбинированная модель показала результаты лучшие, чем у двух отдельных моделей, и может лучше работать при применении к другим наборам данных. Если ваша цель - это возможно большая автоматизация процесса, этот подход позволяет получить устойчивую модель при многих обстоятельствах, не углубляясь в особенности каждой конкретной модели.

Глава 6. Автоматическая подготовка данных (АПД)

В любом проекте подготовка данных для анализа - один из важнейших шагов в любом проекте исследования данных; именно этот шаг традиционно требовал наибольших затрат времени. Узел Инструмент Автоматическая подготовка данных (АПД) решает эту задачу, для чего анализирует данные и находит решения выявленных проблем, выявляет проблемные и малополезные поля, создает при необходимости производные атрибуты и повышает производительность, применяя интеллектуальные методы скрининга. Этот узел можно использовать в полностью автоматическом режиме, позволив ему выбирать и применять исправления или предварительно просматривать изменения перед тем, как они сделаны и приняты, а при желании применять или исправлять их.

Благодаря узлу автоматической подготовки данных вы сможете легко и быстро подготовить данные для исследования, даже если не были раньше знакомы с используемыми при этом понятиями статистики. Если запустить этот узел с параметрами по умолчанию, модели скорее всего будут построены быстрее и быстрее выполнится их скоринг.

В этом примере используется поток *ADP_basic_demo.str*, содержащий ссылки на файл данных *telco.sav*, чтобы продемонстрировать повышение точности, свойственное использованию параметров узла АПД по умолчанию при построении моделей. Эти файлы находятся в каталоге *Demos* любой установки IBM SPSS Modeler. Доступ к ним можно получить из группы программ IBM SPSS Modeler в меню Пуск Windows. Файл *ADP_basic_demo.str* находится в каталоге *streams*.

Построение потока

1. Чтобы построить поток, добавьте узел источников файла статистики, указывающий на файл *telco.sav*, расположенный в каталоге *Demos* вашей установки IBM SPSS Modeler.

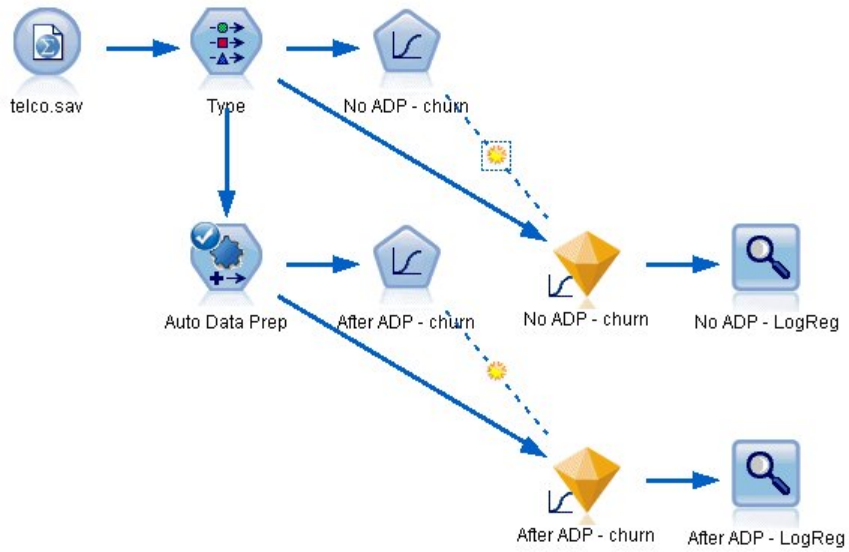


Рисунок 50. Построение потока

2. Присоедините узел Тип к узлу источника и задайте для поля *Отток клиентов* уровень измерения **Флаг** и роль **Назначение**. Для всех остальных полей нужно задать роль **Ввод**.

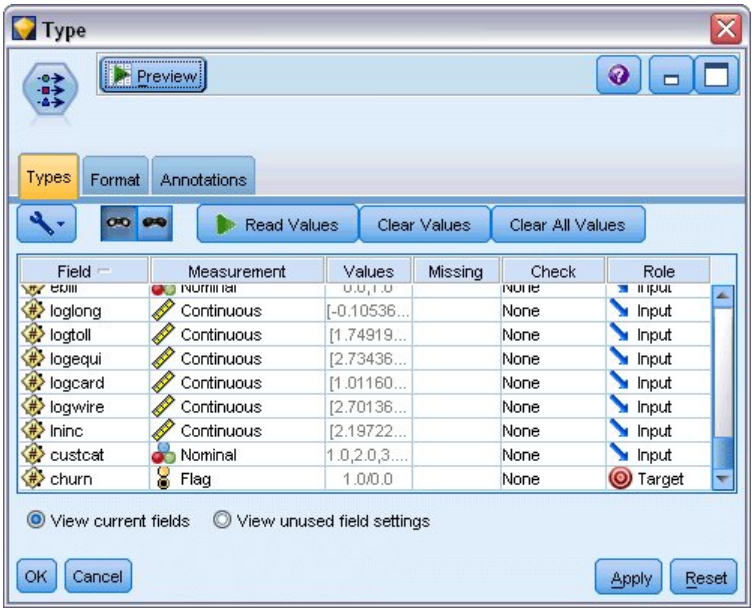


Рисунок 51. Выбор назначения

3. Присоедините Логистический узел к узлу Тип.
4. В Логистическом узле щелкните по вкладке Модель и выберите процедуру **Биномиальная**. В поле *Имя модели* выберите опцию **Пользовательское** и введите значение `Без АПД - отток клиентов`.

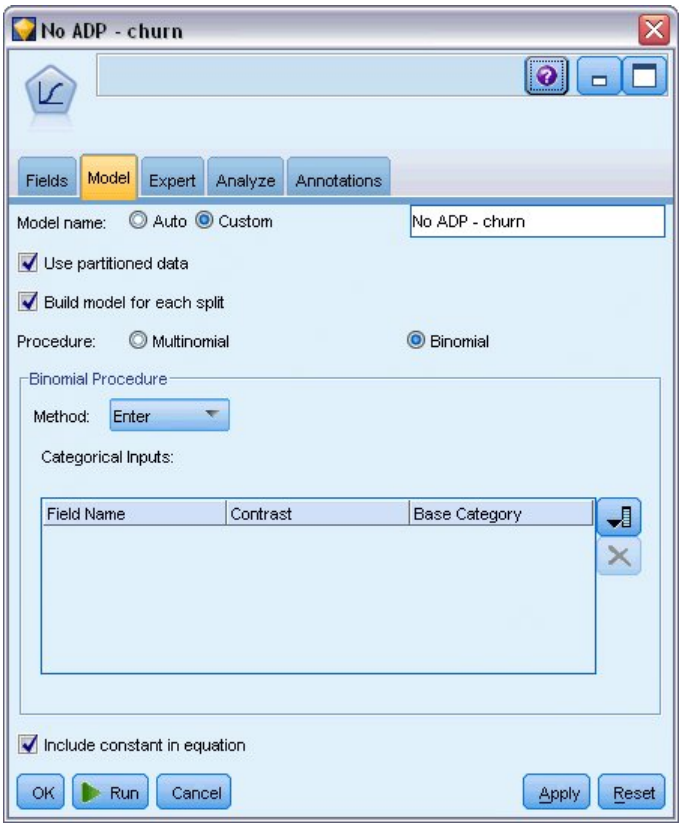


Рисунок 52. Выбор опций модели

5. Присоедините узел АПД к узлу Тип. На вкладке Цели оставьте на месте параметры по умолчанию, чтобы проанализировать и подготовить ваши данные, балансируя точность и скорость.
6. Наверху вкладки Цели щелкните по **Анализировать данные**, чтобы проанализировать и обработать ваши данные.

   Другие опции на вкладке АПД позволяют указать, что нужно в большей степени сконцентрироваться на точности или на скорости обработки, или точнее настроить шаги обработки при подготовке данных.

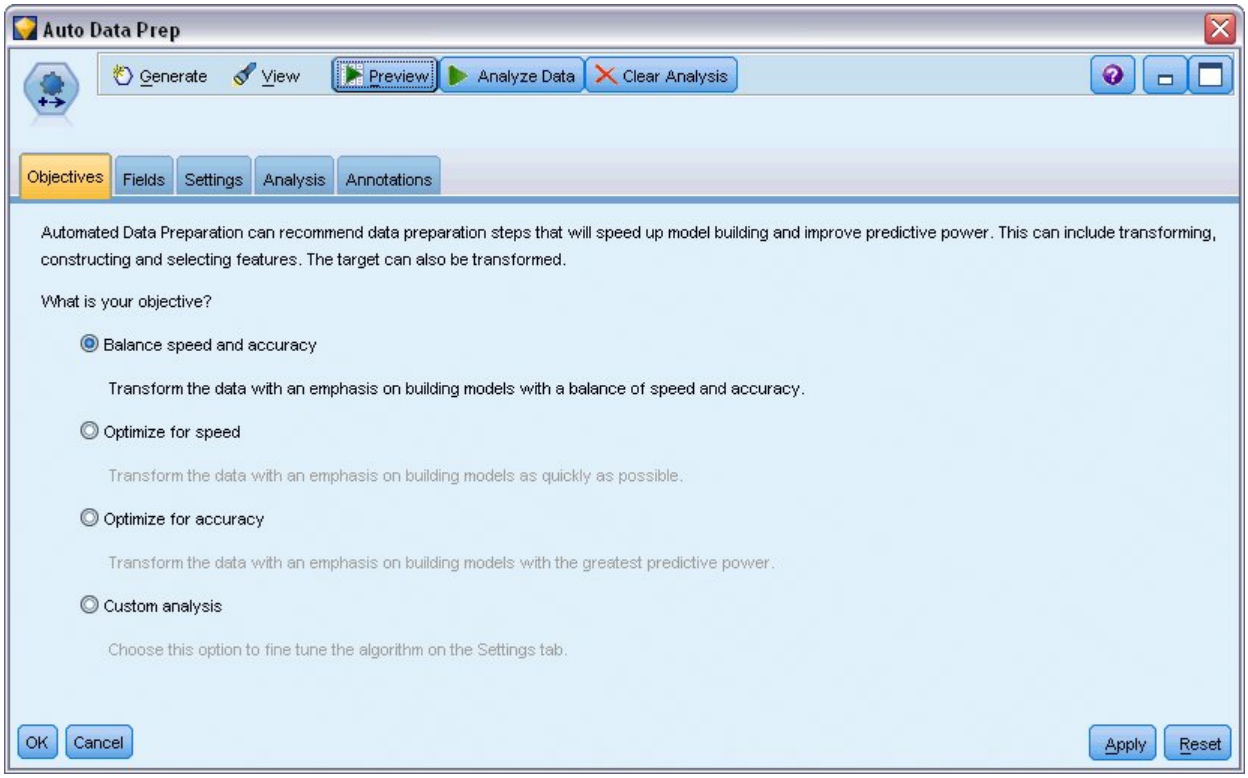


Рисунок 53. Цели АПД по умолчанию

Результаты обработки данных выводятся на вкладке Анализ. **Сводка обработки полей** показывает, что из 41 элемента данных, переданных на узел АПД, 19 были преобразованы для улучшения обработки, а 3 отброшены как неиспользуемые.

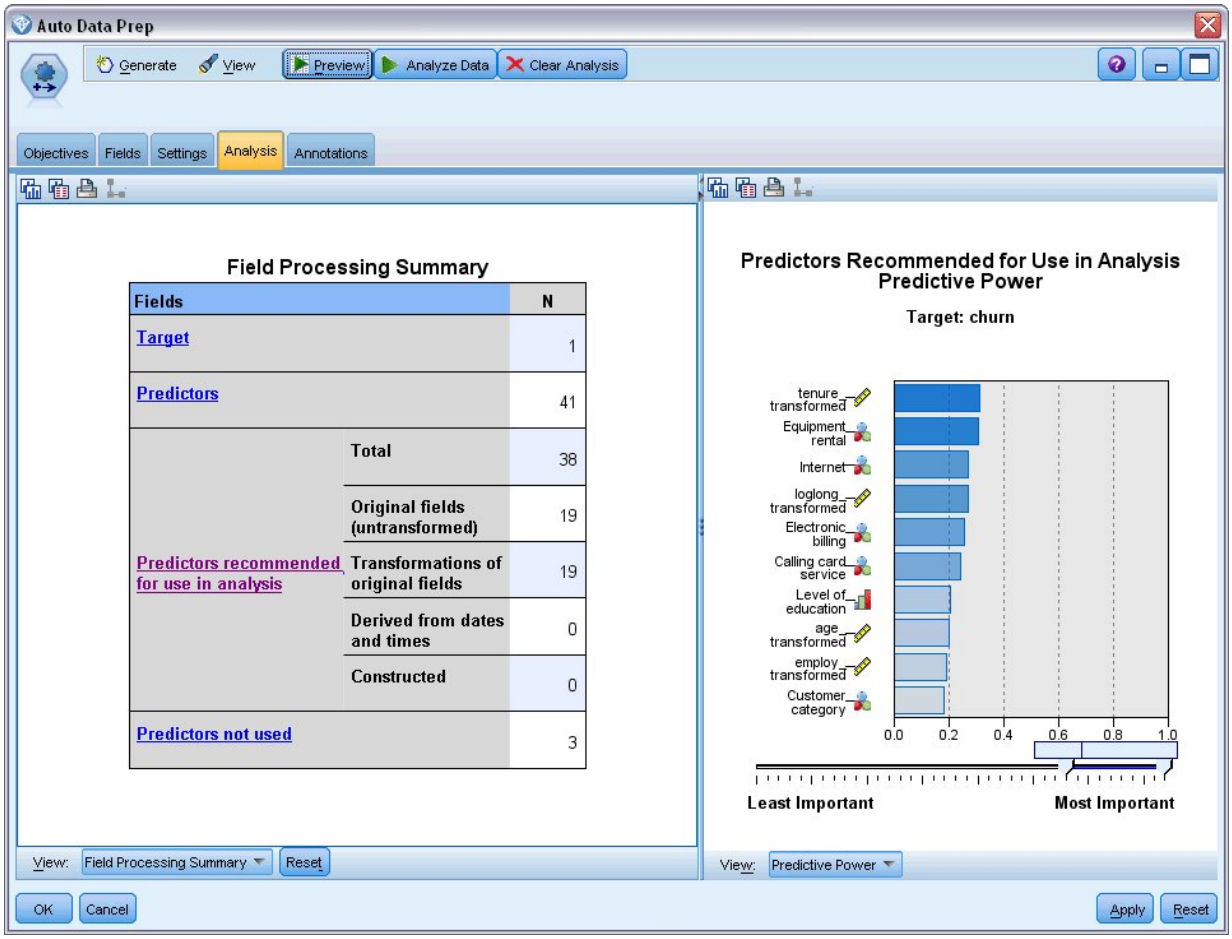


Рисунок 54. Сводка обработки данных

7. Присоедините Логистический узел к узлу ADP.
8. В Логистическом узле щелкните по вкладке Модель и выберите процедуру **Биномиальная**. В поле *Имя модели* выберите опцию **Пользовательское** и введите значение `Без АПД - отток клиентов`.

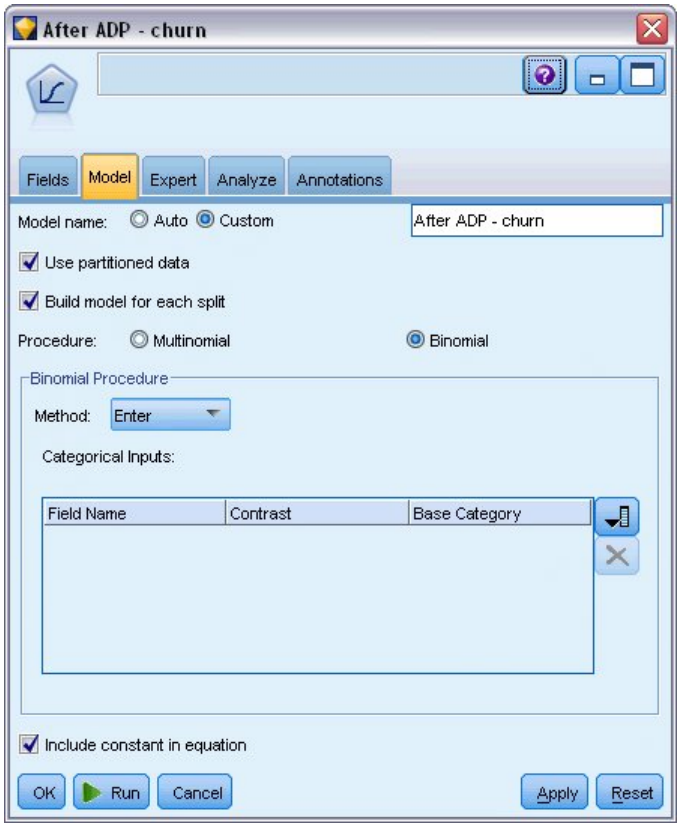


Рисунок 55. Выбор опций модели

Сравнение точности моделей

1. Запустите оба логистических узла для создания слепков моделей, которые будут добавлены в поток и на палитру Модели в правом верхнем углу.

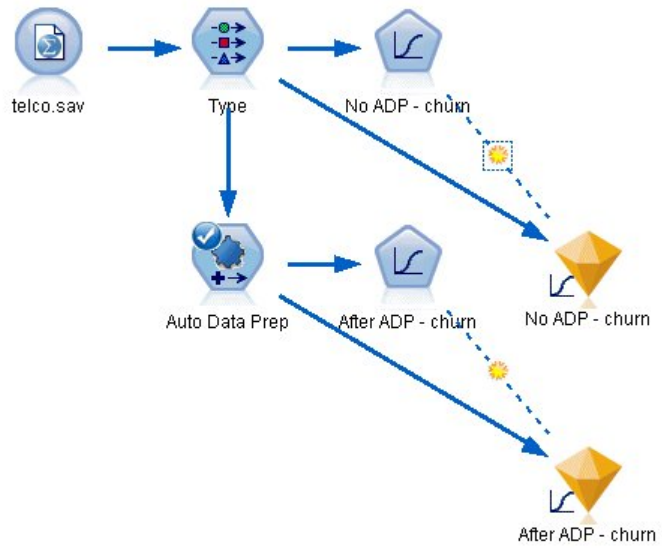


Рисунок 56. Присоединение слепков моделей

2. Присоединить узлы Анализ к слепкам моделей и запустить эти узлы с их параметрами по умолчанию.

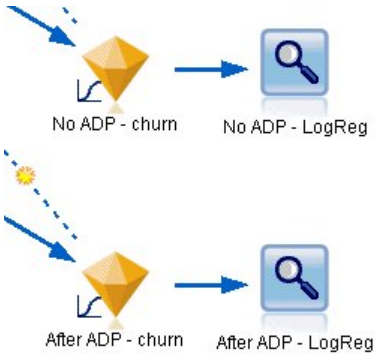


Рисунок 57. Присоединение узлов Анализ

Анализ модели, полученной без АПД, показывает, что просто обработка данных на узле Логистическая регрессия с его параметрами по умолчанию дает модель с низкой точностью - всего 10,6%.

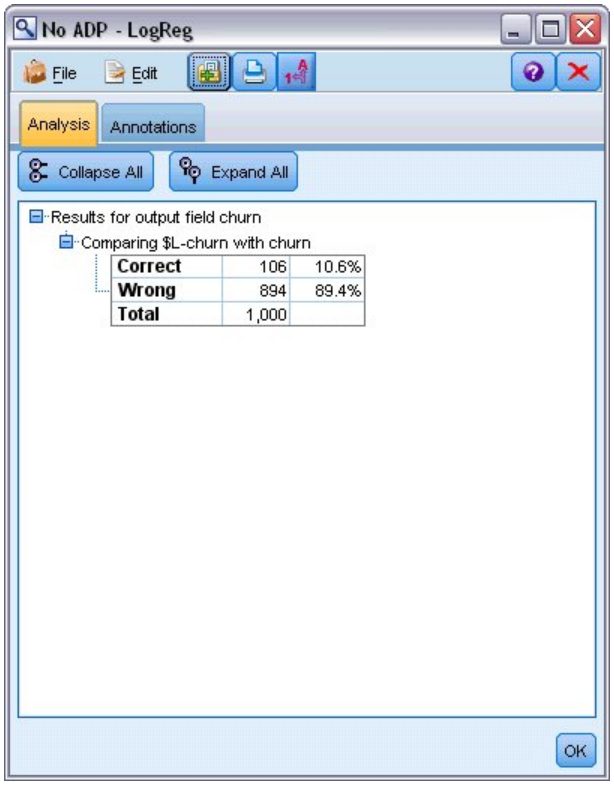


Рисунок 58. Результаты модели, полученной без АПД

Анализ полученной с АПД модели показывает, что обработка данных с параметрами АПД по умолчанию приводит к построению гораздо более точной модели (точность 78,8%).

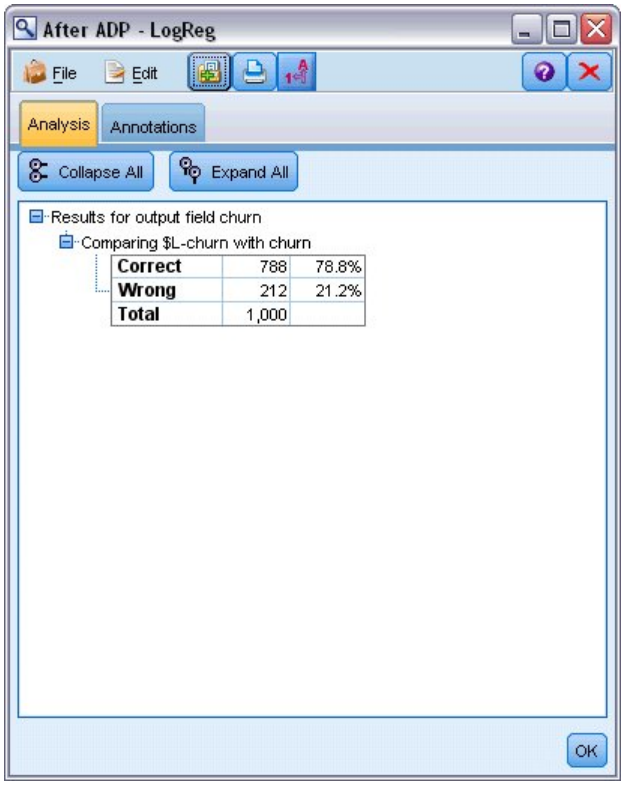


Рисунок 59. Результаты модели, полученной с использованием АПД

Таким образом, просто запустив узел АПД для тонкой настройки обработки ваших данных, удалось построить более точную модель, совсем немного непосредственно манипулируя с данными.

Конечно, если вы заинтересованы в доказательстве или опровержении некоторой теории или хотите построить специфические модели, может потребоваться работать непосредственно с параметрами модели; однако при недостатке времени или при большом объеме данных для подготовки узел АПД может обеспечить существенные преимущества.

Объяснение математических основ методов моделирования, используемых в IBM SPSS Modeler, смотрите в публикации *IBM SPSS Modeler: Руководство по алгоритмам* в каталоге *\Documentation* на установочном диске.

Обратите внимание на то, что результаты из этого примера основаны только на данных обучения. Чтобы оценить, насколько хорошо модели обобщаются на другие данные реального мира, рекомендуется применить узел Разбиение, который будет содержать поднабор записей в целях проверки.

Глава 7. Подготовка данных для анализа (Аудит данных)

Узел Аудит данных обеспечивает первое всестороннее представление данных, перемещаемых в IBM SPSS Modeler. Часто применяемый при начальном исследовании данных, отчет аудита данных показывает сводную статистику, а также гистограммы и диаграммы распределения для каждого поля данных, и позволяет задавать обработку пропущенных значений, выбросов и экстремальных значений.

Этот пример использует поток *telco_dataaudit.str*, в котором используется файл данных *telco.sav*. Эти файлы находятся в каталоге *Demos* любой установки IBM SPSS Modeler. Доступ к ним можно получить из группы программ IBM SPSS Modeler в меню Пуск Windows. Файл *telco_dataaudit.str* находится в каталоге *streams*.

Построение потока

1. Чтобы построить поток, добавьте узел источников файла статистики, указывающий на файл *telco.sav*, расположенный в каталоге *Demos* вашей установки IBM SPSS Modeler.

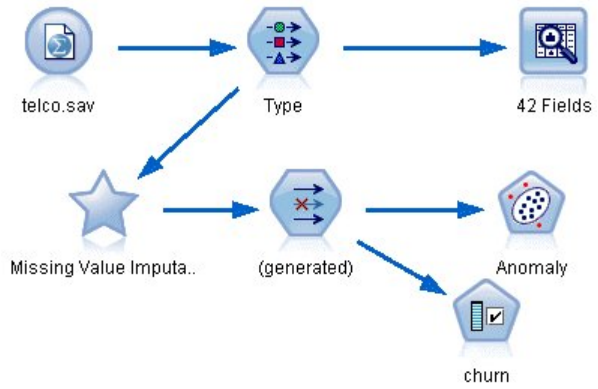


Рисунок 60. Построение потока

2. Добавьте узел Тип, чтобы определить поля, и укажите *churn* (отток клиентов) как поле назначения (Роль = **Назначение**). Для всех остальных полей должна быть задана роль **Входное**, так что это единственное поле назначения.

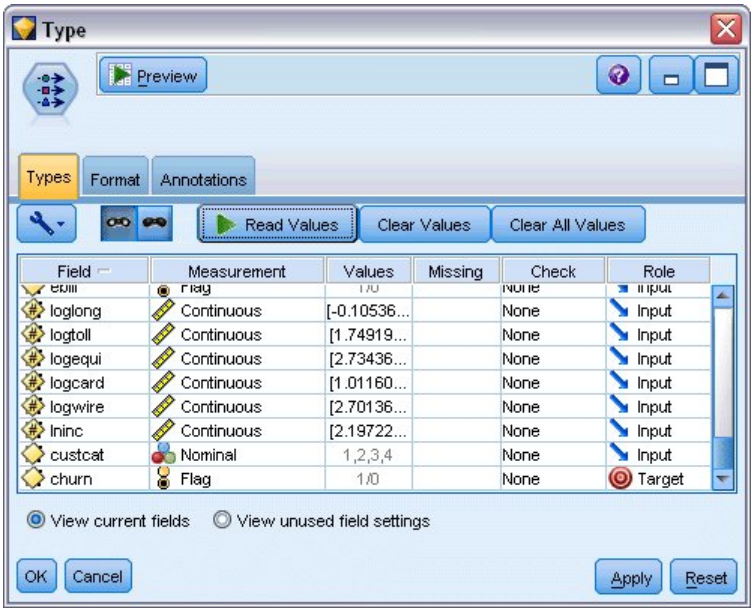


Рисунок 61. Задание назначения

3. Убедитесь, что для полей были правильно определены уровни измерений. Например, большинство полей со значениями 0 и 1 можно рассматривать как флаги, но некоторые поля, такие как пол, точнее считать номинальными полями с двумя значениями.

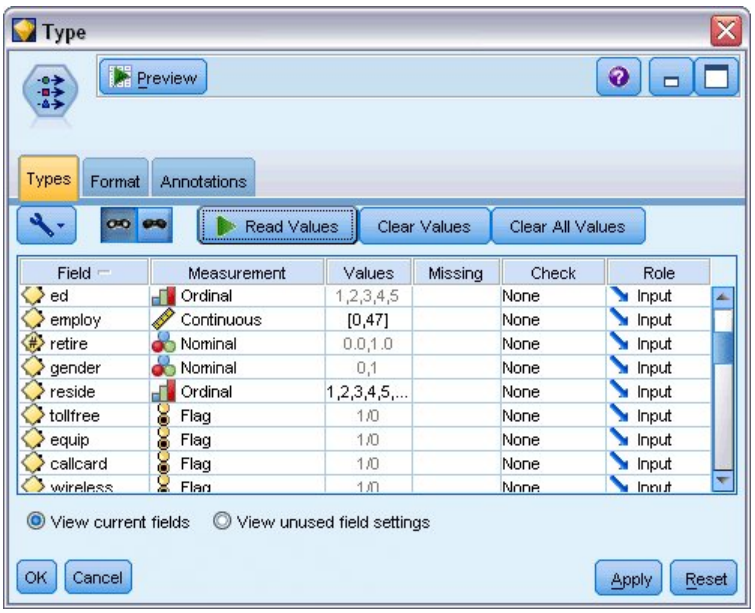


Рисунок 62. Задание уровней измерений

Подсказка: Чтобы изменить свойства для нескольких полей с аналогичными значениями (такими как 0/1), щелкните по заголовку столбца *Значения*, чтобы отсортировать поля в этом столбце, а затем удерживайте нажатой кнопку Shift, используя мышь или клавишу со стрелкой для выбора всех полей, которые вы хотите изменить. Затем можно щелкнуть правой кнопкой мыши по выбранному, чтобы изменить уровень измерения или другие атрибуты всех выбранных полей.

4. Присоедините узел Аудит данных к потоку. На вкладке Параметры оставьте на месте параметры по умолчанию, чтобы включить в отчет все поля. Так как *churn* - это единственное поле назначения, определенное в узле Тип, оно автоматически используется как наложение.

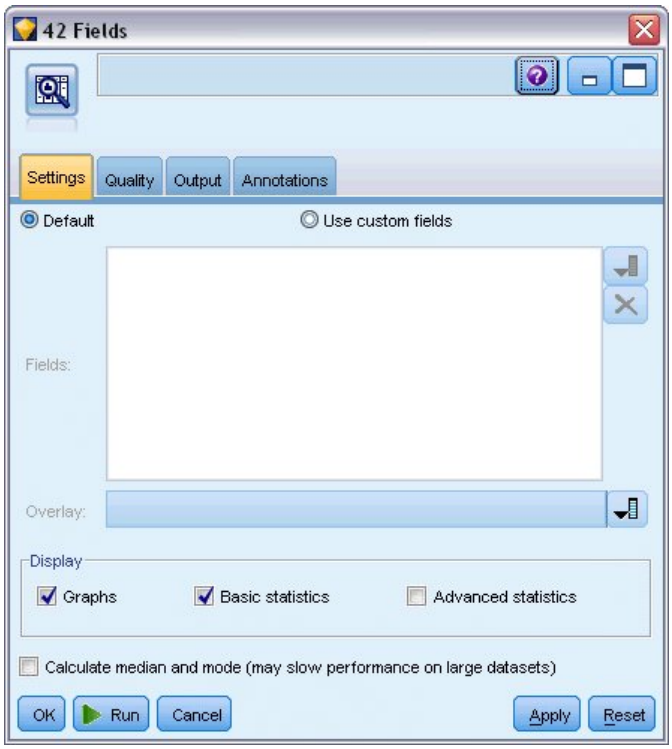


Рисунок 63. Узел аудита данных: вкладка Параметры

На вкладке Качество оставьте на месте параметры по умолчанию для детектирования пропущенных значений, выбросов и экстремальных значений и нажмите кнопку **Выполнить**.

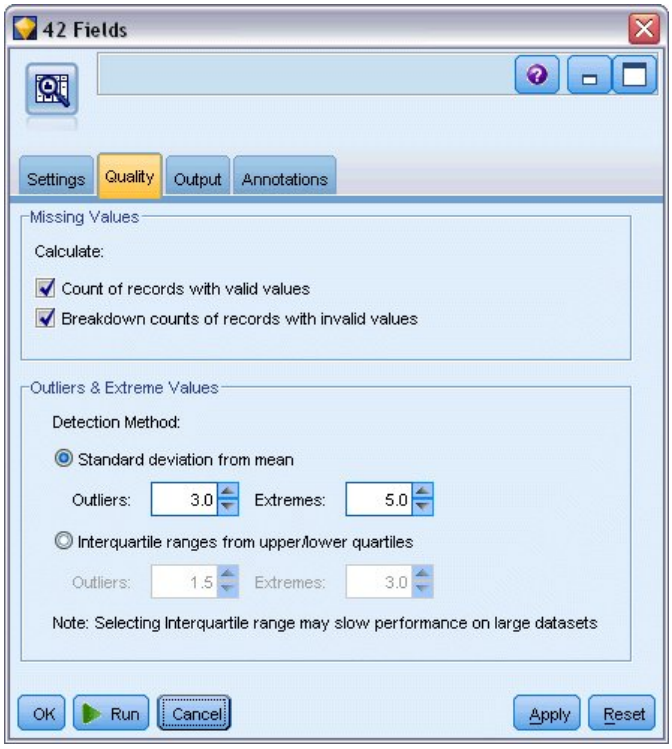


Рисунок 64. Узел аудита данных: вкладка Качество

Просмотр статистики и диаграмм

Браузер аудита данных открывается с миниизображениями диаграмм и описательной статистикой для каждого поля.

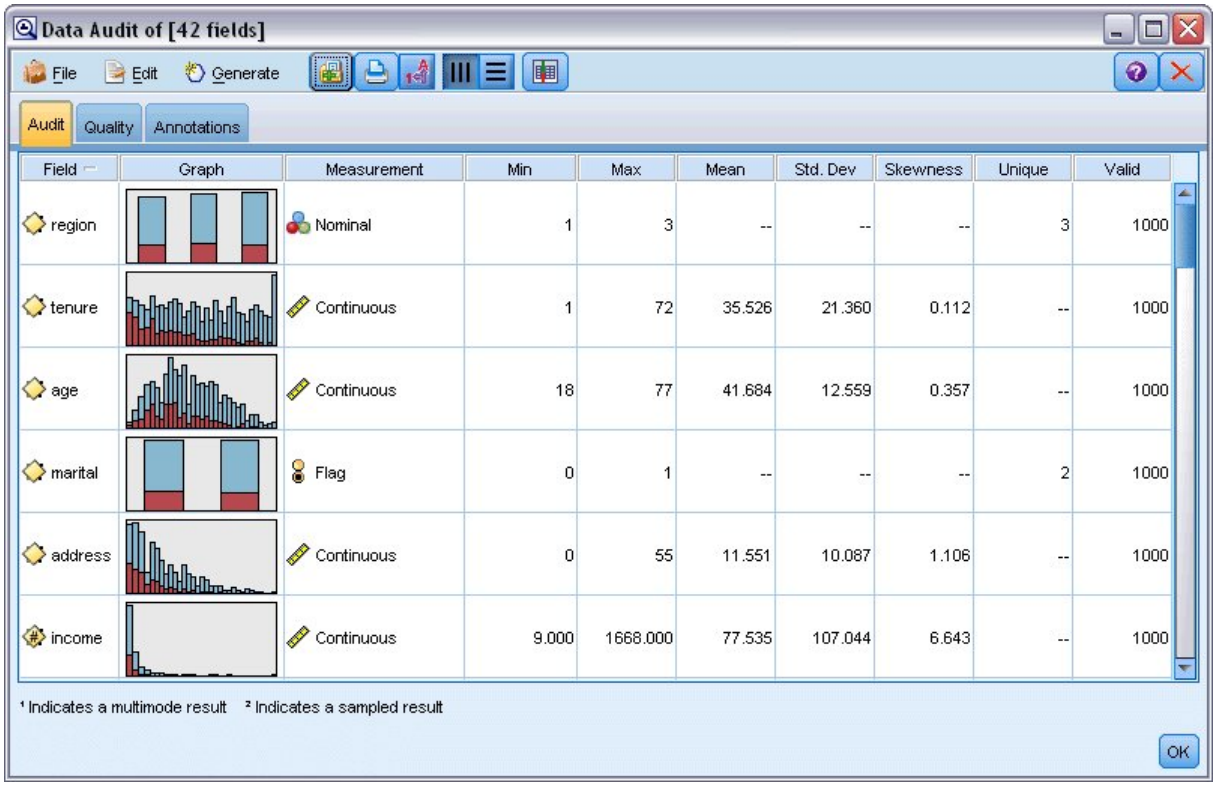


Рисунок 65. Браузер аудита данных

Используйте панель инструментов для вывода меток значений и полей и для переключения выравнивания диаграмм между горизонтальным и вертикальным (только для категориальных полей).

1. Панель инструментов или меню Изменить можно использовать также для выбора выводимой статистики.

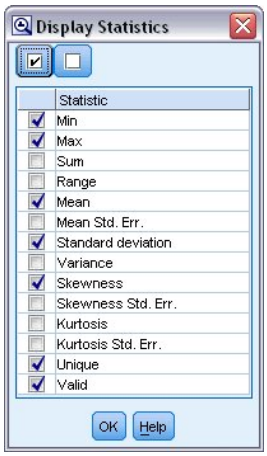


Рисунок 66. Вывести статистики

Дважды щелкните по любому миниизображению диаграммы в отчете аудита, чтобы просмотреть полноразмерную версию этой диаграммы. Так как *churn* (отток клиентов) - это единственное поле назначения в потоке, оно автоматически используется как наложение. Можно переключить вывод меток значений и полей, используя панель инструментов окна диаграммы, или нажать кнопку Режим редактирования для дальнейшей настройки диаграммы.

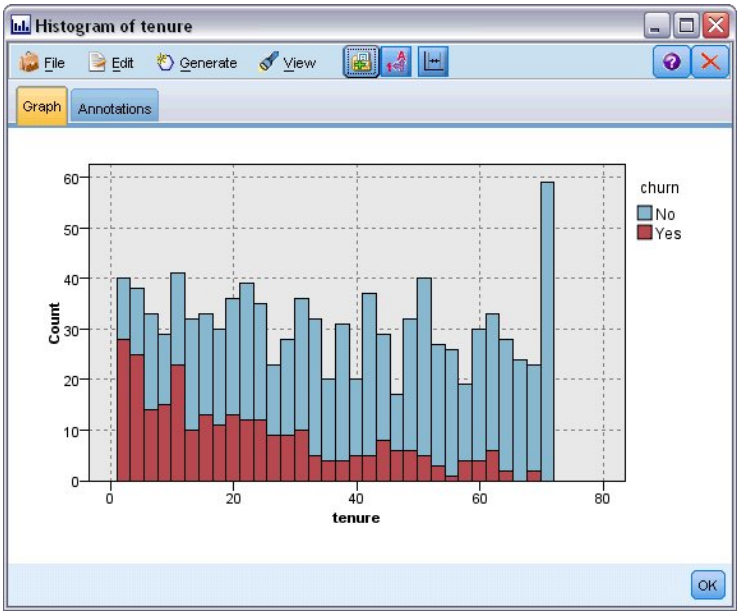


Рисунок 67. Гистограмма срока пребывания

Вместо этого можно выбрать одно или несколько миниизображений и сгенерировать узел Диаграмма для каждого из них. Сгенерированные узлы будут размещены на холсте потока, и их можно добавить в поток для повторного создания этой конкретной диаграммы.

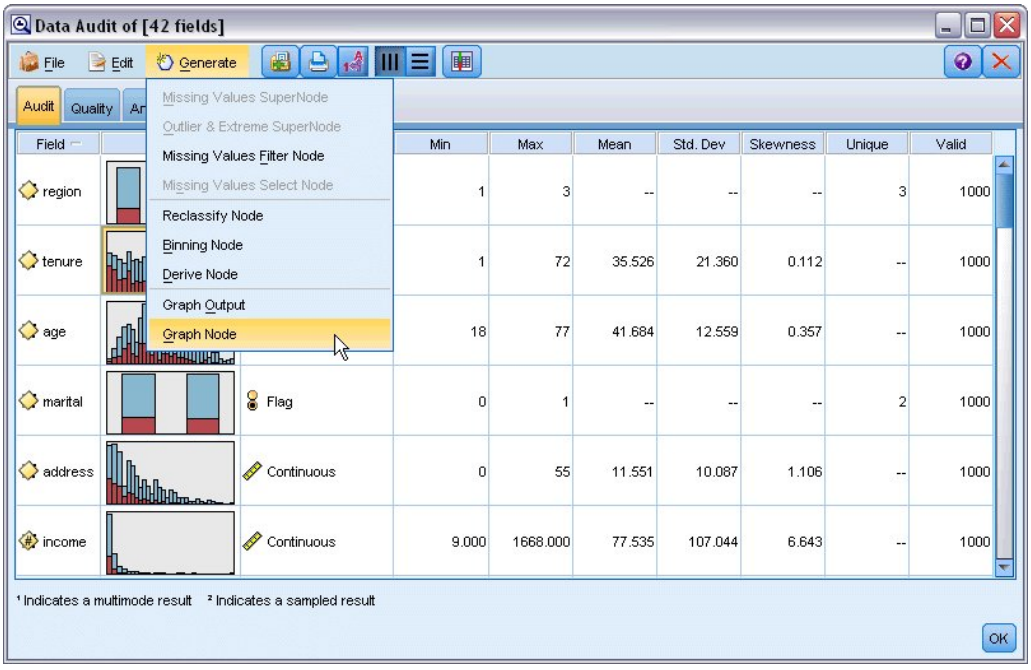


Рисунок 68. Узел Генерирование диаграммы

Обработка значений выбросов и пропущенных значений

На вкладке Качество в отчете аудита выводится информация о выбросах, экстремумах и пропущенных значениях.

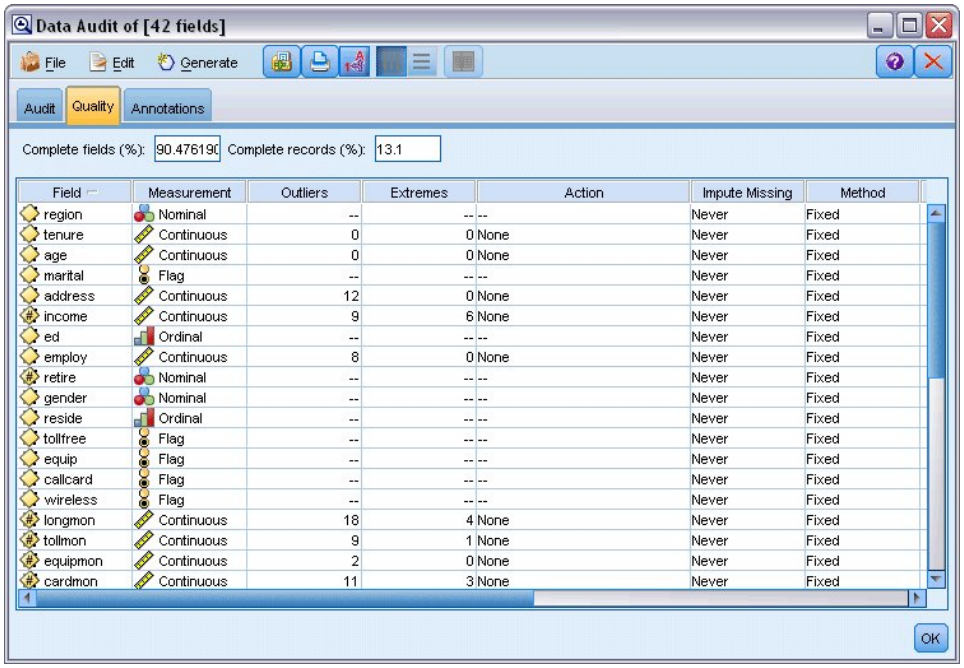


Рисунок 69. Вкладка Качество браузера аудита данных

Можно указать также способы для обработки этих значений и генерирования Надузлов для автоматического применения преобразований. Например, можно выбрать одно или несколько полей и выбрать опцию импутации или замены пропущенных значений для этих полей, используя один из нескольких способов, например, алгоритм C&RT.

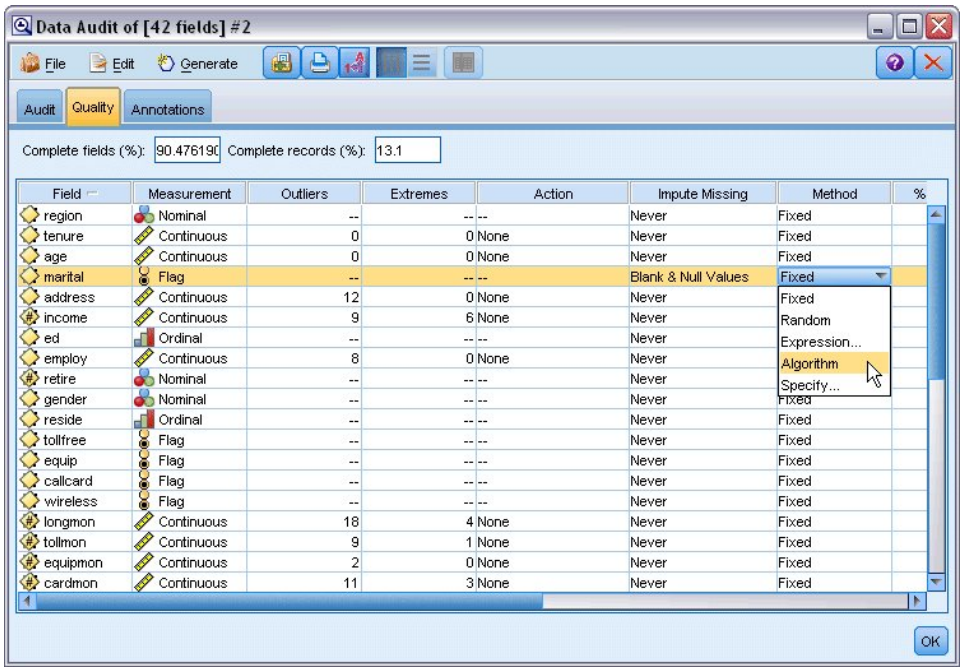


Рисунок 70. Выбор способа импутации

После указания способа импутации для одного или нескольких полей, чтобы сгенерировать надузел Пропущенные значения, выберите в меню:

Создать > Надузел пропущенных значений

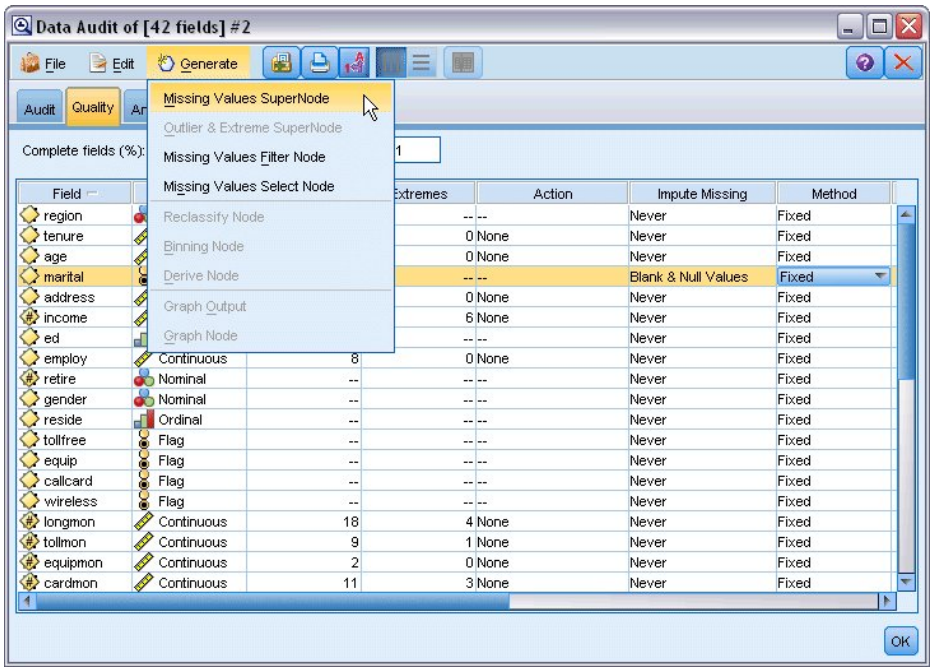


Рисунок 71. Генерирование надузла

Сгенерированный надузел добавляется на холст потока, где его можно присоединить к потоку, чтобы применить преобразования.

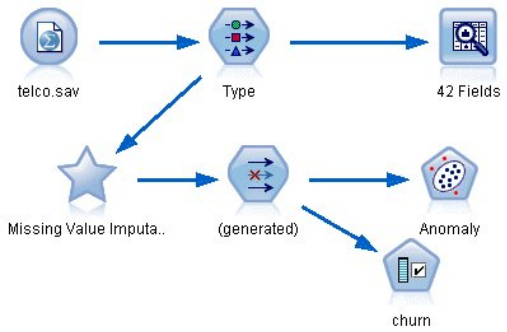


Рисунок 72. Поток с надузлом пропущенных значений

Фактически надузел содержит ряд узлов, выполняющих требуемые преобразования. Чтобы понять, как это работает, можно изменить надузел и щелкнуть по **Увеличить масштаб**.

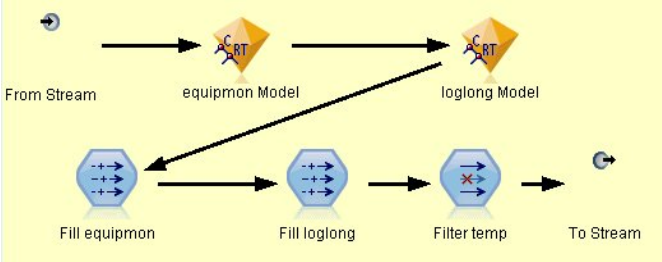


Рисунок 73. Надузел крупным планом

Для каждого поля, импутированного, например, методом алгоритма, будет отдельная модель C&RT наряду с узлом заполнения, заменяющим пробелы и пустые значение на значение, предсказанное моделью. Можно добавлять, изменять или удалять отдельные узлы в составе надузла, чтобы выполнить дальнейшую настройку поведения.

Вместо этого можно сгенерировать узел Выбор или Фильтр, чтобы удалить поля или записи с пропущенными значениями. Например, можно отфильтровать все поля с процентной долей качества ниже заданного порога.

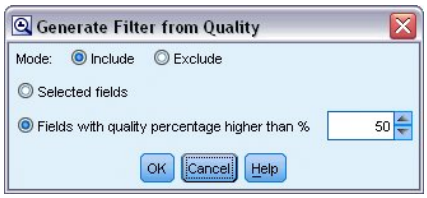


Рисунок 74. Генерирование узла фильтра

Аналогичным образом можно обрабатывать выбросы и экстремальные значения. Задайте действие, которое вы хотите применить для каждого из полей - подавление, отбрасывание или аннулирование, а затем сгенерируйте надузел для применения выбранных преобразований.

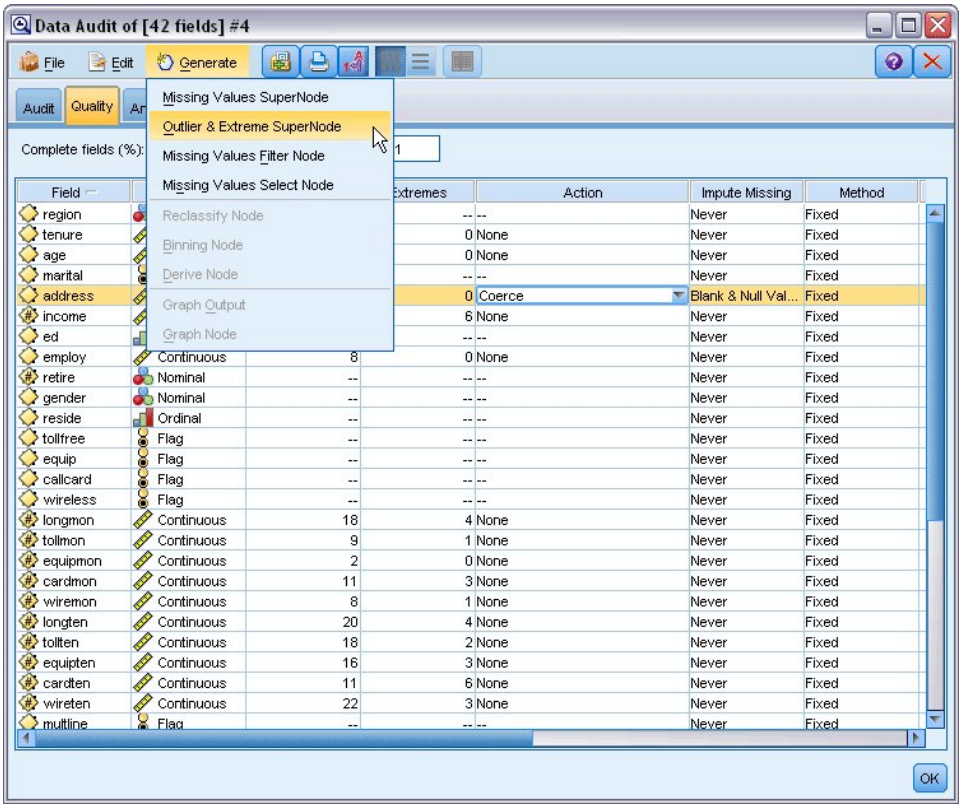


Рисунок 75. Генерирование узла фильтра

После завершения аудита и добавления сгенерированных узлов в поток можно продолжить анализ. При необходимости вы можете и дальше отбирать свои данные, используя Детектирование аномалий, Выбор характеристик и некоторые другие способы.

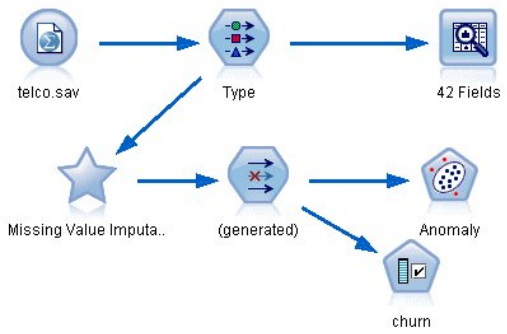


Рисунок 76. Поток с надузлом пропущенных значений

Глава 8. Лечение препаратами (Исследовательские диаграммы/C5.0)

В этом разделе представьте себе, что вы - врач-исследователь, компилирующий данные для некоторого исследования. Вы собрали данные о некотором множестве пациентов с одним и тем же заболеванием. Во время курса лечения каждый пациент принимал одно из пяти лекарств. Часть вашей работы - использовать технику исследования данных, чтобы найти, какой препарат оптимален для будущих пациентов с таким же заболеванием.

Этот пример использует поток *druglearn.str*, в котором используется файл данных *DRUG1n*. Эти файлы находятся в каталоге *Demos* любой установки IBM SPSS Modeler. Доступ к ним можно получить из группы программ IBM SPSS Modeler в меню Пуск Windows. Файл *druglearn.str* находится в каталоге *streams*.

В демонстрационном примере использованы такие поля данных:

Поле данных	**Описание**
Age (Возраст)	(Число)
Sex (Пол)	*M* или *F*
BP (артериальное давление)	Давление крови: *HIGH* (высокое), *NORMAL* (нормальное) или *LOW* (низкое)
Cholesterol (Холестерин)	Содержание холестерина в крови: *NORMAL* (нормальное) или *HIGH* (высокое)
Na	Концентрация натрия в крови
K	Концентрация калия в крови
Drug (Препарат)	Рецептурный препарат, для которого наблюдалась реакция пациента

Чтение в текстовых данных

Рисунок 77. Добавление узла файла переменных

Можно считывать данные в формате текст с разделителями, используя узел **узел файла переменных**. Узел файла переменных можно добавить с палитры - либо щелкните по вкладке **Источники** и найдите узел, либо перейдите на вкладку **Избранное**, содержащую этот узел по умолчанию. Далее щелкните дважды по новому узлу, чтобы открыть его диалоговое окно.

Нажмите кнопку с символом многоточия (...) справа от поля Файл, чтобы перейти в каталог, в котором в вашей системе установлен модуль IBM SPSS Modeler. Откройте каталог демо-версий *Demos* и выберите файл с именем *DRUG1n*.

Убедитесь, что включен переключатель **Считывать имена полей из файла**, и обратите внимание на поля и значения, загруженные теперь в диалоговое окно.

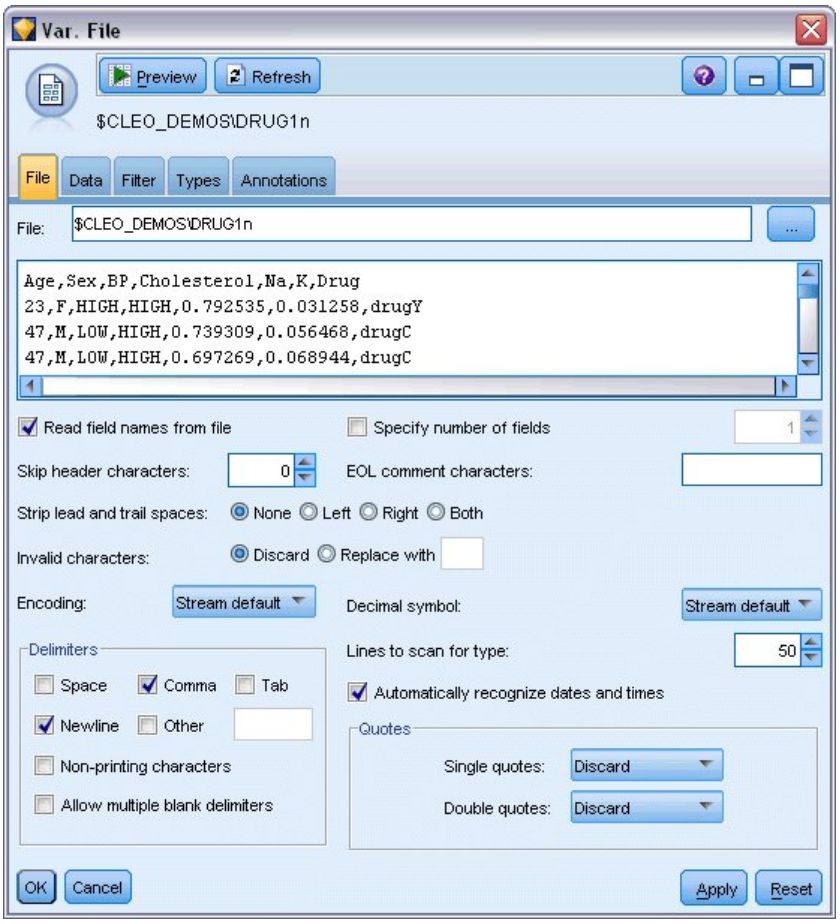


Рисунок 78. Диалоговое окно Файл переменных

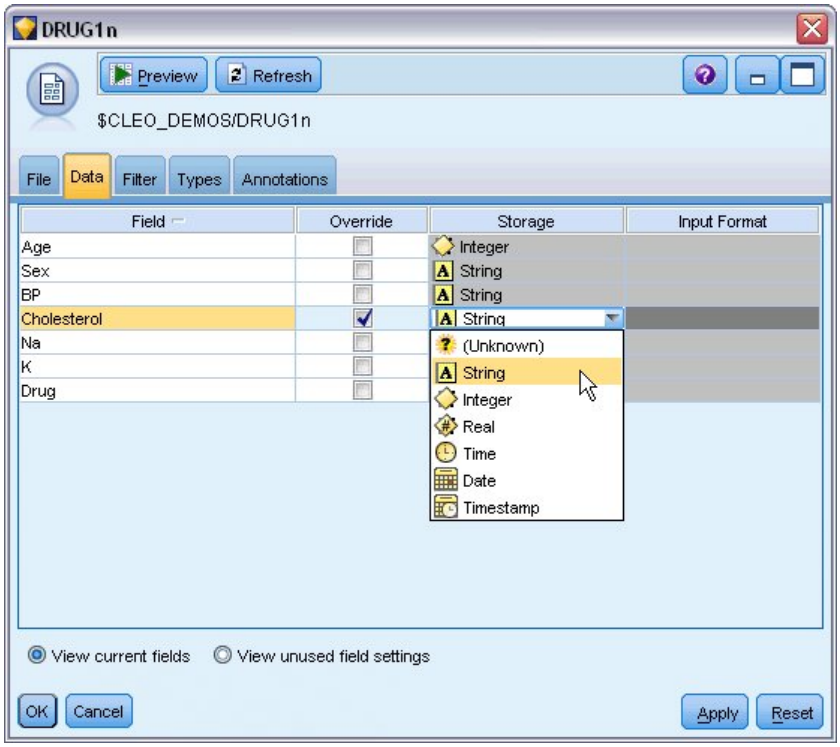


Рисунок 79. Изменение типа хранения для поля

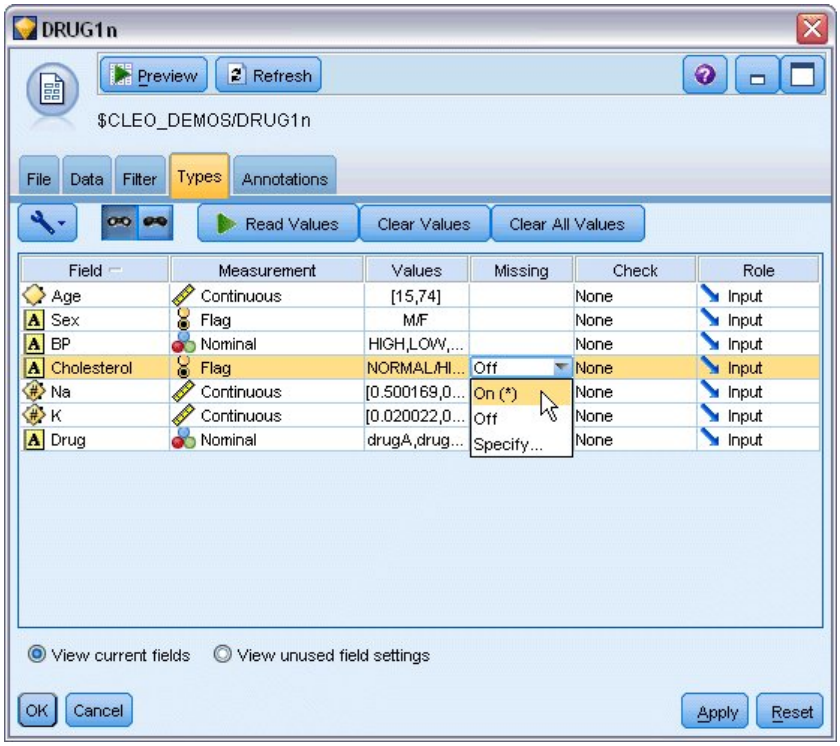


Рисунок 80. Выбор опций Значение на вкладке Типы

Щелкните по вкладке **Данные**, чтобы перезадать поле и изменить для него **Систему хранения**. Обратите внимание на то, что система хранения отличается от параметра **Измерение**, то есть типа измерений (типа использования) поля данных. Дополнительную информацию о типах полей для ваших данных можно найти

на вкладке **Типы**. Кроме того, можно выбрать **Прочитать значения**, чтобы просмотреть фактические значения для каждого поля с учетом выбранного в столбце *Значения*. Этот процесс называется **инстанциирование**.

Добавление таблицы

Теперь, когда вы загрузили файл данных, можете посмотреть на значения некоторых записей. Один из способов сделать это - построить поток, содержащий узел Таблица. Чтобы поместить в поток узел Таблица, либо дважды щелкните по значку на палитре, либо перетащите его на холст.

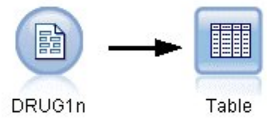
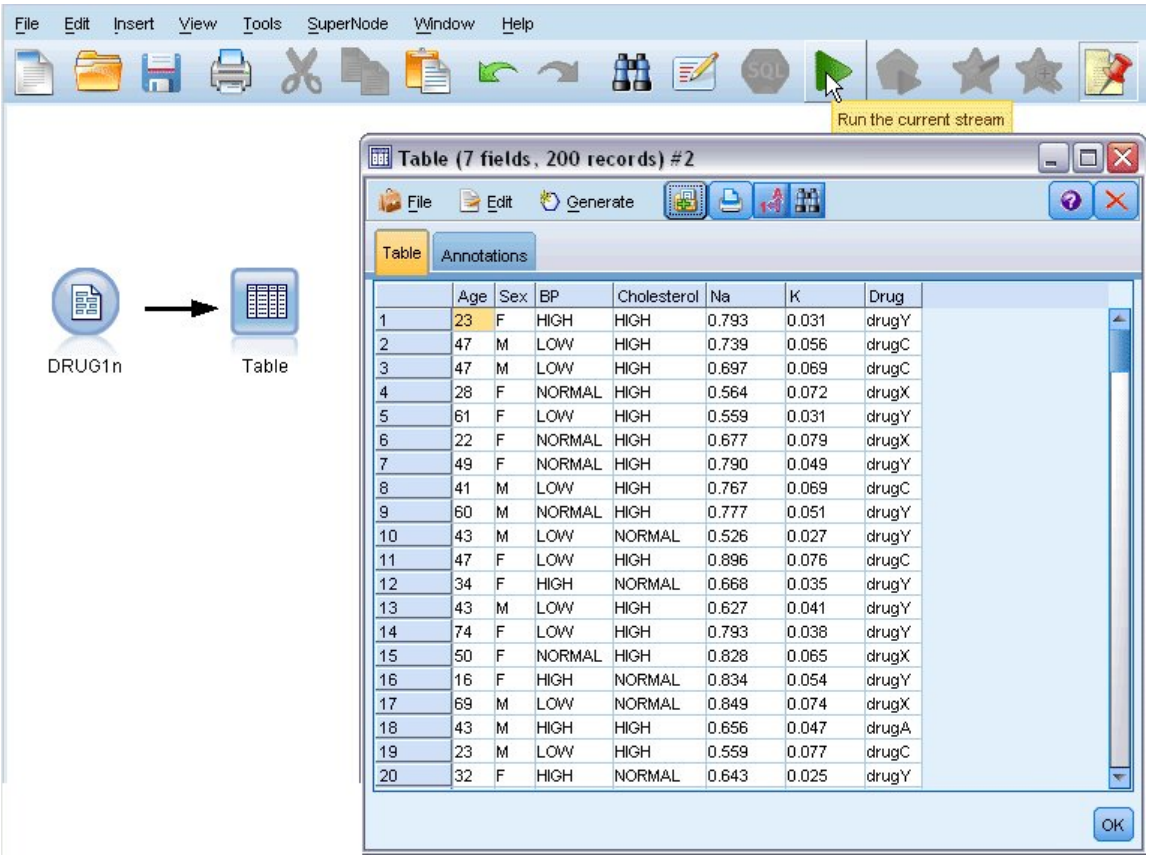


Рисунок 81. Узел Таблица, подсоединенный к узлу источника данных



	Age	Sex	BP	Cholesterol	Na	K	Drug
1	23	F	HIGH	HIGH	0.793	0.031	drugY
2	47	M	LOW	HIGH	0.739	0.056	drugC
3	47	M	LOW	HIGH	0.697	0.069	drugC
4	28	F	NORMAL	HIGH	0.564	0.072	drugX
5	61	F	LOW	HIGH	0.559	0.031	drugY
6	22	F	NORMAL	HIGH	0.677	0.079	drugX
7	49	F	NORMAL	HIGH	0.790	0.049	drugY
8	41	M	LOW	HIGH	0.767	0.069	drugC
9	60	M	NORMAL	HIGH	0.777	0.051	drugY
10	43	M	LOW	NORMAL	0.526	0.027	drugY
11	47	F	LOW	HIGH	0.896	0.076	drugC
12	34	F	HIGH	NORMAL	0.668	0.035	drugY
13	43	M	LOW	HIGH	0.627	0.041	drugY
14	74	F	LOW	HIGH	0.793	0.038	drugY
15	50	F	NORMAL	HIGH	0.828	0.065	drugX
16	16	F	HIGH	NORMAL	0.834	0.054	drugY
17	69	M	LOW	NORMAL	0.849	0.074	drugX
18	43	M	HIGH	HIGH	0.656	0.047	drugA
19	23	M	LOW	HIGH	0.559	0.077	drugC
20	32	F	HIGH	NORMAL	0.643	0.025	drugY

Рисунок 82. Запуск потока с панели инструментов

При двойном щелчке по узлу на палитре он автоматически подсоединяется к выбранному узлу на холсте потока. Другой вариант - если узлы еще не соединены, можно подсоединить узел Источник к узлу Таблица при помощи средней кнопки мыши. Вместо средней кнопки мыши можно использовать обычную, удерживая нажатой клавишу Alt. Чтобы просмотреть таблицу, на панели инструментов нажмите зеленую кнопку со стрелкой, запускающую поток, или щелкните правой кнопкой по узлу Таблица и выберите **Запуск**.

Создание графа распределения

Во время исследования данных часто полезно изучать данные, создавая визуальные сводки. IBM SPSS Modeler предлагает на выбор ряд различных типов диаграмм для различных типов сводимых данных. Например, чтобы найти, какая часть пациентов оказалась чувствительна к тому или иному препарату, используйте узел Распределение.

Добавьте в поток узел Распределение, подключите добавленный узел к узлу Источник и затем дважды щелкните по узлу, чтобы отредактировать опции вывода.

Выберите *Drug* (препарат) как поле назначения, для которого нужно вывести распределение. Затем щелкните по **Запуск** в диалоговом окне.

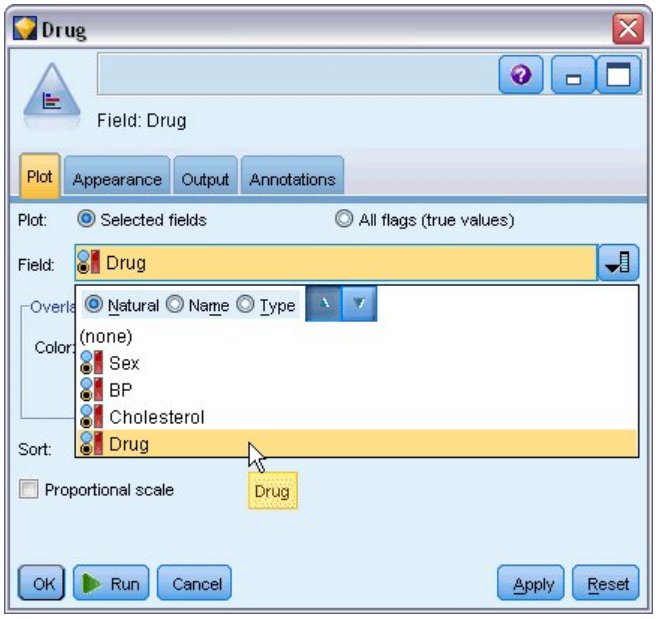


Рисунок 83. Выбор препарата как поля назначения

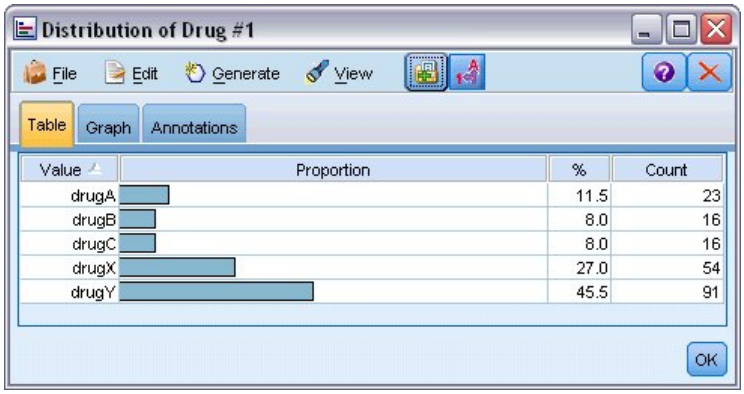


Рисунок 84. Распределение отклика на тип препарата

Полученная диаграмма помогает увидеть "форму" данных. На ней видно, что что пациенты чаще всего реагировали на препарат *Y* и реже всего на препараты *B* и *C*.

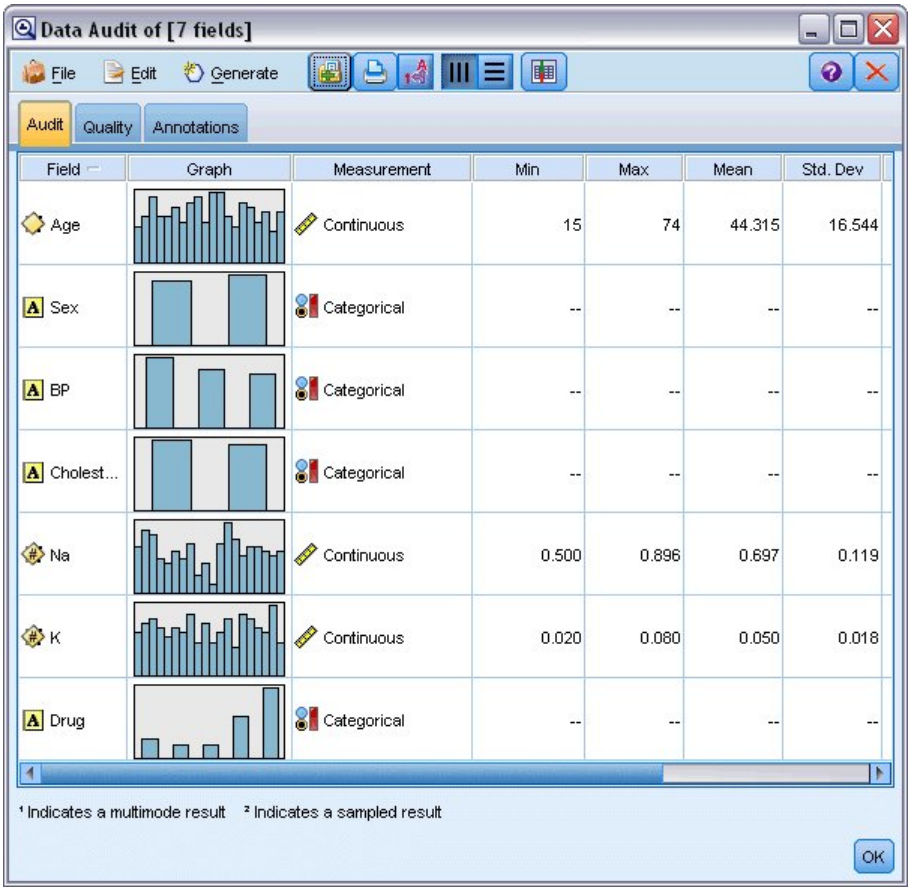


Рисунок 85. Результаты аудита данных

Другой вариант - подсоединить и выполнить узел Аудит данных, чтобы увидеть распределения и диаграммы сразу по всем полям. Узел Аудит данных доступен на вкладке Вывод.

Создание диаграммы рассеяния

Теперь посмотрим, какие факторы могли повлиять на целевую переменную - *Drug* (препарат). Как исследователь, вы знаете о таких важных факторах, как концентрация натрия и калия в крови. Поскольку это числовые показатели, можно создать диаграмму рассеяния натрия в зависимости от калия, использовав категории препарата как наложение цветов.

Поместите узел График в рабочее пространство и соедините его с узлом Источник, затем дважды щелкните, чтобы отредактировать.

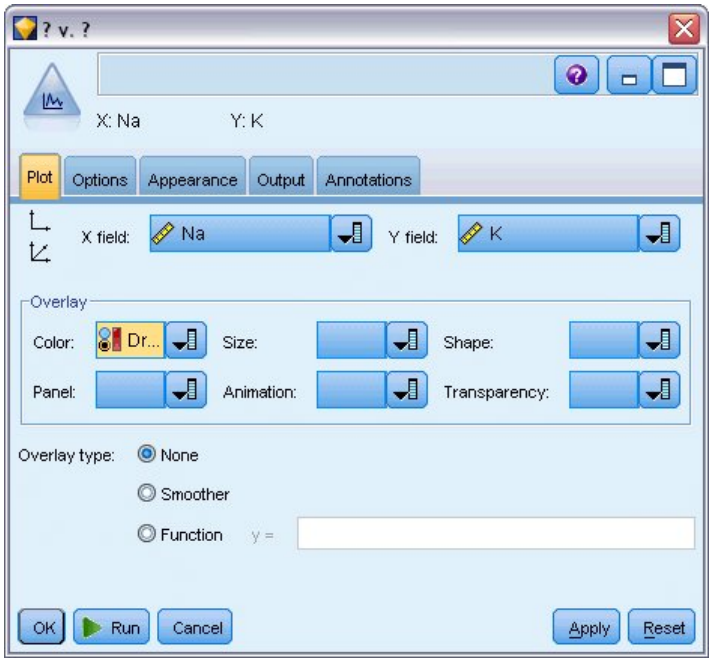


Рисунок 86. Создание диаграммы рассеяния

На вкладке График выберите *Na* как поле X, *K* как поле Y и *Drug* как поле наложения. Затем щелкните по **Запуск**.

На графике ясно видно порог, выше которого препарат *Y* всегда предпочтителен и ниже которого препарат *Y* никогда не предпочтителен. Этот порог представляет определенное отношение натрия (*Na*) к калию (*K*).

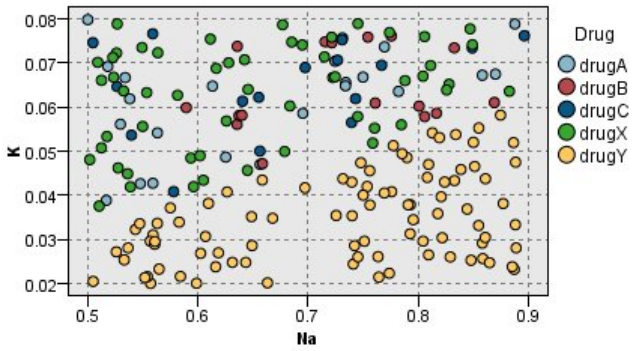


Рисунок 87. Диаграмма рассеяния для распределения препарата

Создание веб-диаграммы

Поскольку многие поля данных - категориальные, можно также попробовать построить веб-граф, изображающий связи между между различными категориями. Для начала в рабочем пространстве к узлу Источник подсоедините узел Web. В диалоговом окне Веб-узла выберите *BP* (артериальное давление) и *Drug* (препарат). Затем щелкните по **Запуск**.

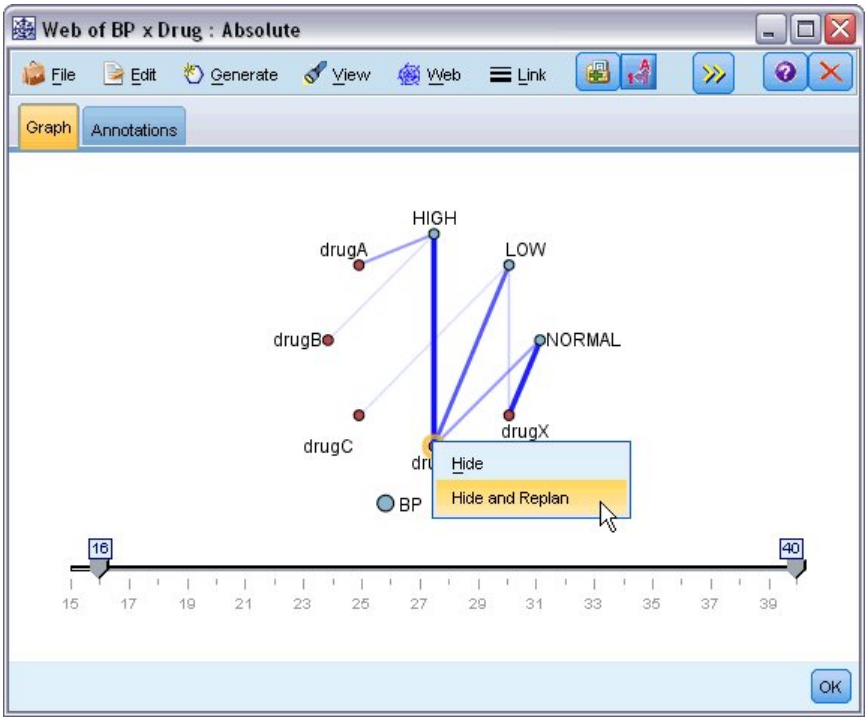


Рисунок 88. Веб-граф препаратов в зависимости от артериального давления

Глядя на диаграмму, можно видеть, что препарат *Y* связан со всеми тремя уровнями артериального давления. Это не удивительно - вы уже узнали, в какой ситуации в который препарат *Y* оптимален. Чтобы сосредоточиться на остальных препаратах, препарат *Y* можно скрыть. В меню **Вид** выберите **Режим редактирования**, затем щелкните правой кнопкой по точке препарата *Y* и выберите **Скрыть и перепланировать**.

На упростившейся диаграмме препарат *Y* и все его связи скрыты. Теперь ясно видно, что с высоким артериальным давлением связаны только препараты *A* и *B*. Только препараты *C* и *X* связаны с пониженным артериальным давлением. А нормальное артериальное давление связано только с препаратом *X*. Однако пока неясно, как для конкретного пациента выбрать между препаратами *A* и *B* или *C* и *X*. Здесь на помощь приходит моделирование.

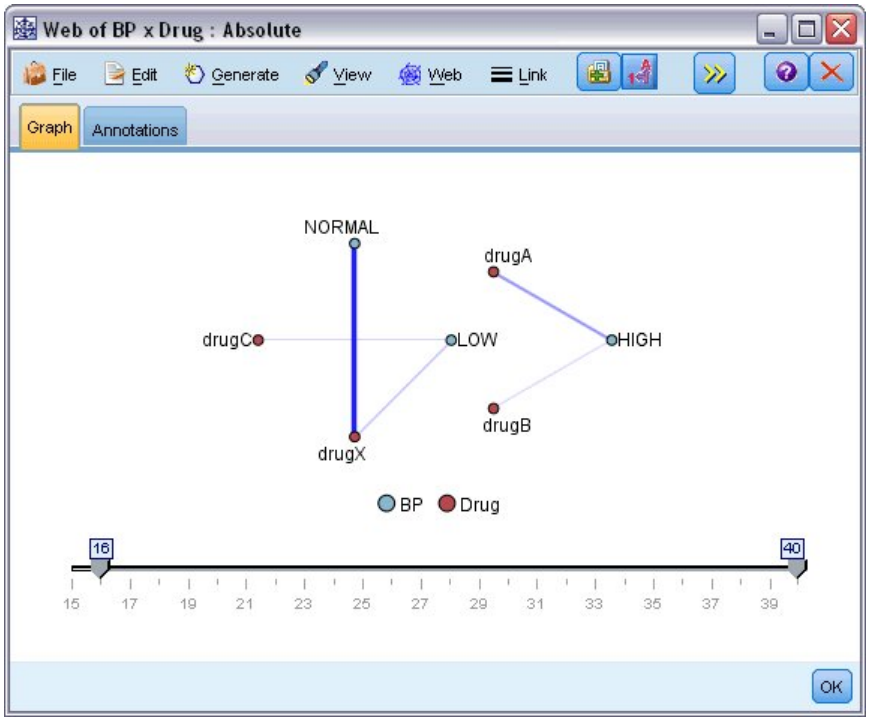


Рисунок 89. Веб-граф, на котором скрыт препарат Y и его связи

Вычисление нового поля

Поскольку отношение натрия к калию, по-видимому, предсказывает, когда следует использовать препарат *Y*, можно вычислить поле, содержащее значение этого отношения для каждой записи. Это поле может оказаться полезным в дальнейшем, когда вы построите модель, прогнозирующую применение каждого из пяти препаратов. Для начала, чтобы упростить структуру потока, удалите все узлы, кроме узла источника DRUG1n. Присоедините к DRUG1n узел вычислений (вкладка Опции поля), затем щелкните дважды по узлу вычислений, чтобы отредактировать его.

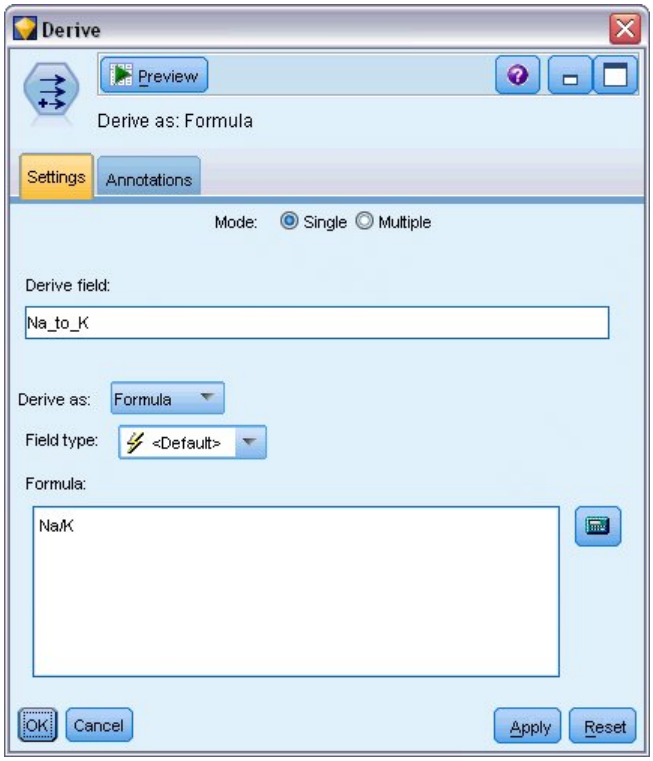


Рисунок 90. Редактирование узла производных данных

Назовите новое поле *Na_to_K* (натрий к калию). Поскольку новое поле нужно вычислить, деля значение натрия на значение калия, в качестве формулы введите `Na/K`. Кроме того, можно создать формулу, щелкнув по значку справа от поля. Откроется Построитель выражений, в котором можно создавать выражения, пользуясь встроенным списком функций, операндов и полей и их значений.

Можно проверить распределение нового поля, присоединив к узлу вычислений узел гистограмм. В диалоговом окне узла гистограмм укажите *Na_to_K* как поле, для которого строится диаграмма, *Drug* как поле наложения.

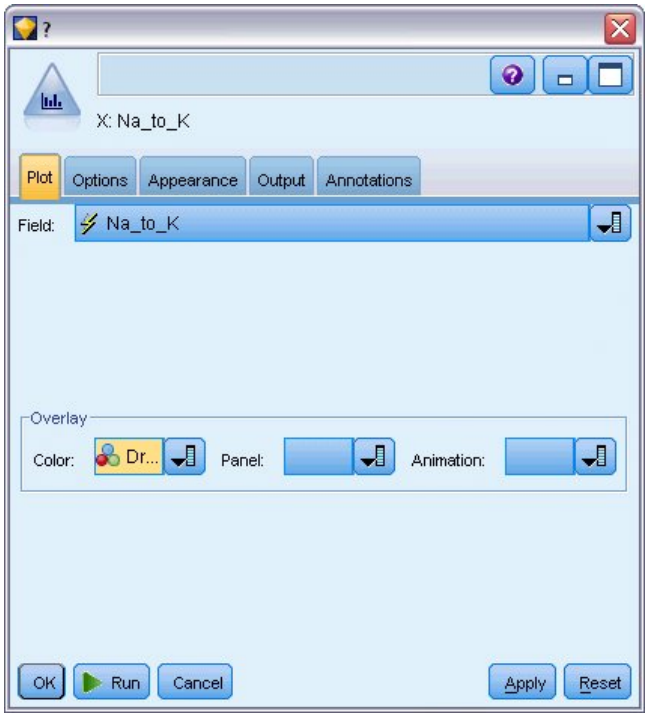


Рисунок 91. Редактирование узла гистограммы

Запустив поток, вы получите приведенную здесь диаграмму. По этому экрану вы можете заключить, что при значении *Na_to_K* около 15 и выше оптимален препарат *Y*.

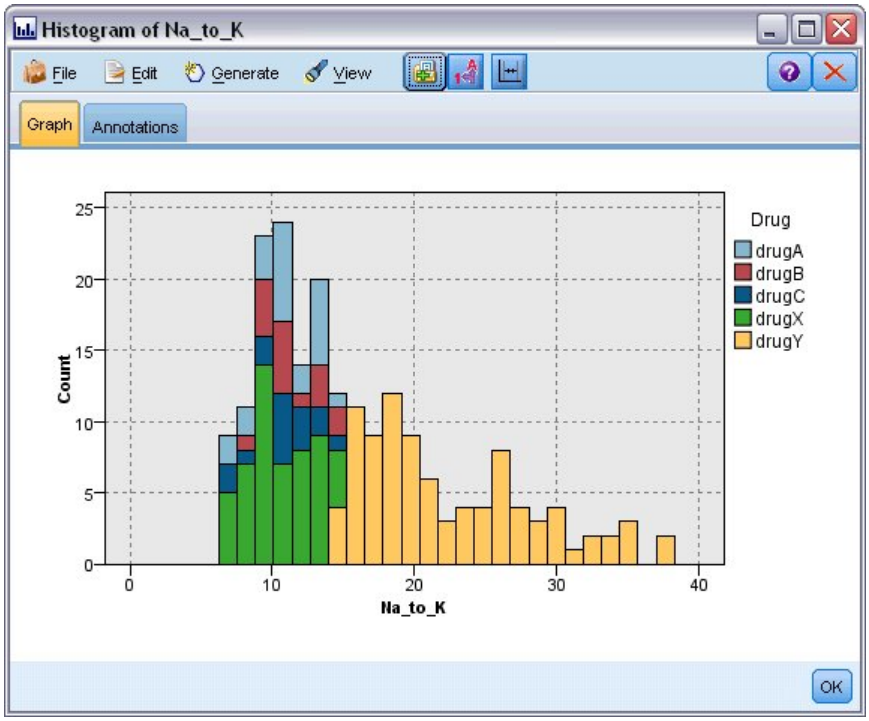


Рисунок 92. Вывод гистограммы

Построение модели

Исследование и обработка данных помогают сформировать некоторые гипотезы. На выбор препарата, по-видимому, влияет отношение натрия к калию в крови, а также артериальное давление. Но пока вы не можете вполне объяснить все эти взаимосвязи. В этой ситуации найти ответы поможет моделирование. В данном случае мы попробуем подогнать под данные модель построения правил C5.0.

Поскольку используется производное поле, *Na_to_K*, можно отфильтровать исходные поля, *Na* и *K*, чтобы не использовать их дважды в алгоритме моделирования. Это можно сделать при помощи узла Фильтр.

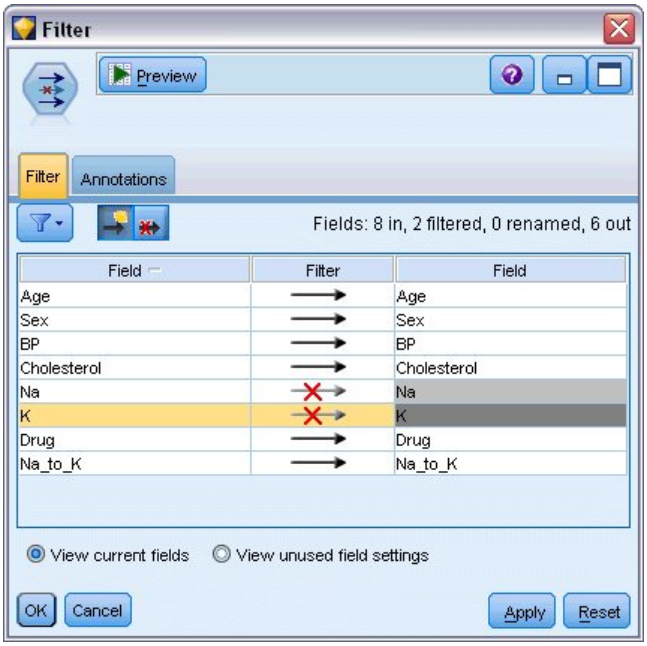


Рисунок 93. Редактирование узла Фильтр

На вкладке Фильтр щелкните по стрелкам рядом с *Na* и *K*. На стрелках появятся красные буквы X, показывая, что эти поля теперь отфильтрованы.

Затем подсоедините узел Тип, подключенный к узлу Фильтр. При помощи узла Тип можно указать, какие типы полей используются и какую роль играют для прогнозирования результатов.

На вкладке Типы задайте для поля *Drug* (препарат) роль **Назначение**, указав, что *Drug* - это то поле, которое нужно прогнозировать. Оставьте для остальных полей роль **Источник**, чтобы использовать их как предикторы.

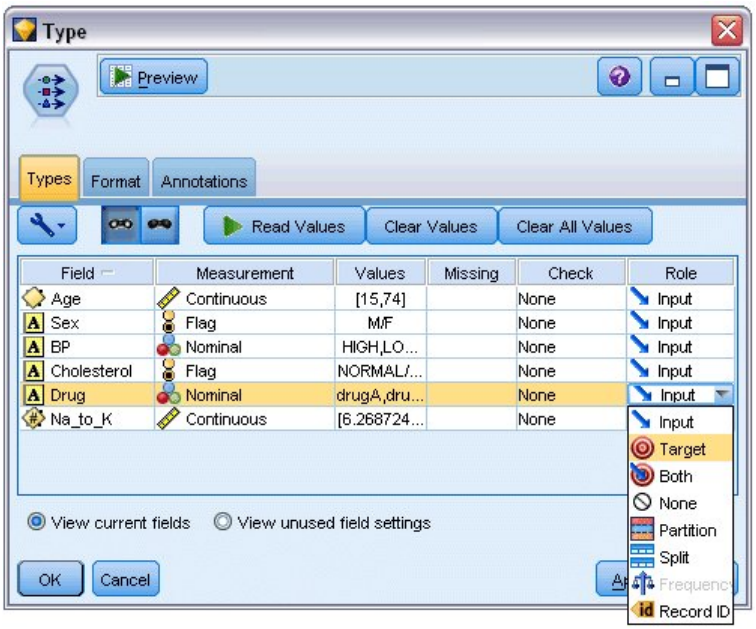


Рисунок 94. Редактирование узла Тип

Чтобы оценить модель, поместите узел C5.0 в рабочую область и подсоедините его к концу потока, как показано. Затем нажмите зеленую кнопку **Запуск** на панели инструментов, чтобы запустить поток.

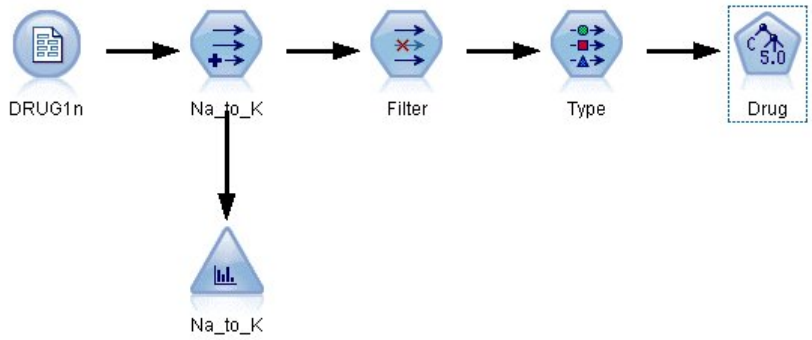


Рисунок 95. Добавление узла C5.0

Просмотр модели

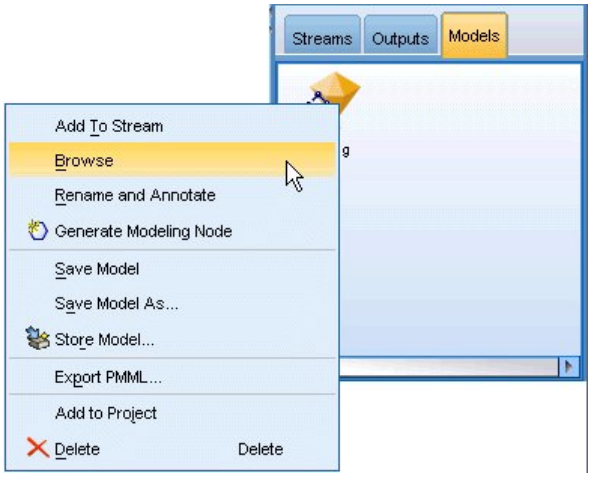


Рисунок 96. Просмотр модели

После выполнения узла C5.0 слепок модели добавляется в поток, а также на палитру Модели в верхнем правом углу окна. Для просмотра модели щелкните правой кнопкой по любому из значков и в контекстном меню выберите **Изменить** или **Обзор**.

В браузере правил выводится набор правил, сгенерированных узлом C5.0, в формате дерева решений. Вначале дерево свернуто. Чтобы развернуть его, щелкните по кнопке **Все**, которая выводит все уровни.

Рисунок 97. Браузер правил

Теперь вы видите недостающие части картины. Для пациентов с отношением *Na* к *K* ниже 14,64 и высоким кровяным давлением выбор препарата зависит от возраста. Для пациентов с низким кровяным давлением оптимальным предиктором, по-видимому, будет уровень холестерина.

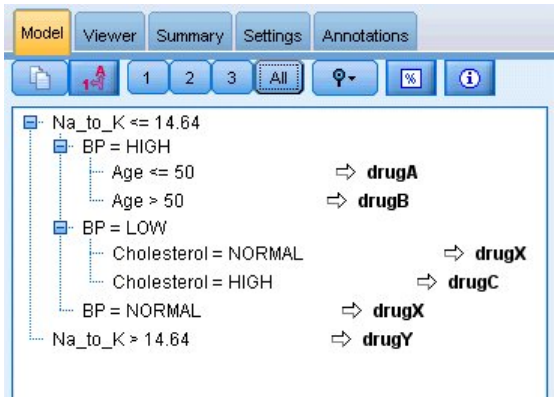


Рисунок 98. Браузер правил в полностью раскрытом виде

То же самое дерево решений можно просмотреть в более сложном графическом формате, щелкнув по вкладке **Средство просмотра**. Здесь удобно просмотреть число и процент наблюдений для каждой категории кровяного давления.

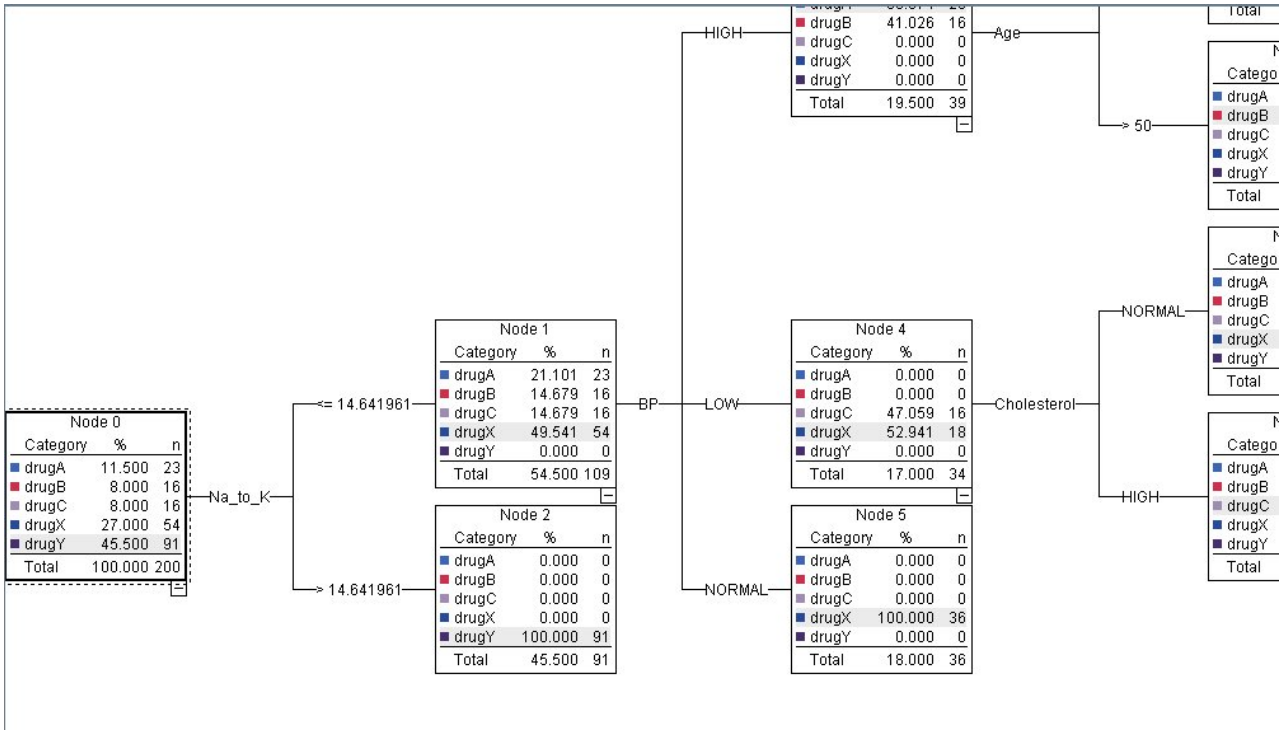


Рисунок 99. Дерево решений в графическом формате

Использование узла Анализ

Оценить точность модели можно при помощи узла Анализ. Подсоедините узел Анализ (с палитры узла вывода) к слепку модели, откройте узел Анализ и выберите **Запуск**.

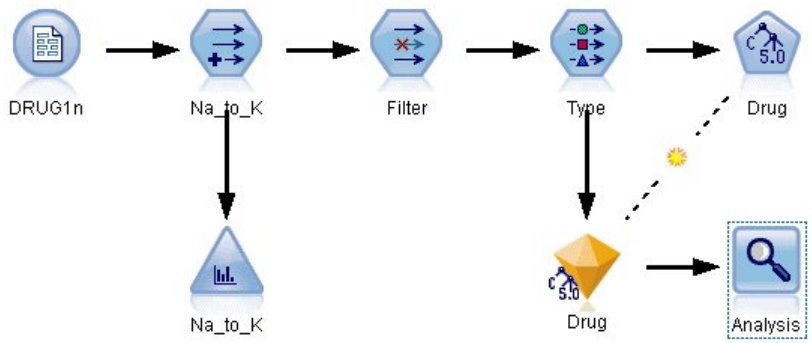


Рисунок 100. Добавление узла Анализ

В выводе узла Анализ показано, что при этом искусственном наборе данных модель правильно предсказывает выбор препарата для каждой записи в наборе данных. При реальном наборе данных вы вряд ли увидите точность 100%, но при помощи узла Анализ сможете узнать, достигла ли модель приемлемой точности для конкретного практического применения.

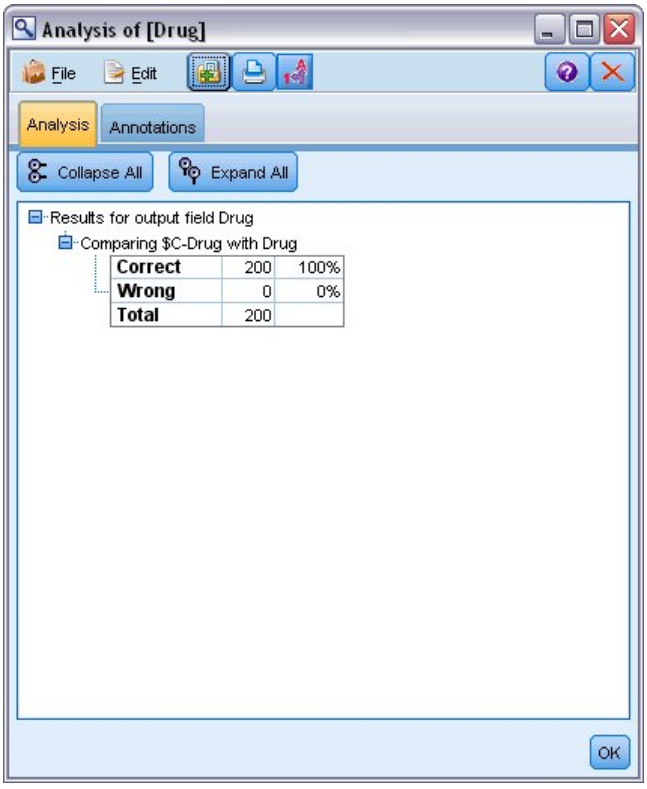


Рисунок 101. Вывод узла анализа

Глава 9. Экранирование предикторов (выбор характеристик)

Узел Выбор характеристик помогает идентифицировать поля, которые будут наиболее важными для предсказания определенных выходных данных. Из набора в сотни и даже тысячи предикторов узел выбора характеристик экранирует, ранжирует и отбирает предикторы, которые могут оказаться самыми важными. В конечном счете вы можете получить более быструю и более эффективную модель, которая использует меньше предикторов, выполняется быстрее и оказывается проще для понимания.

Используемые в этом примере данные представляют собой хранилище данных для гипотетической телефонной компании и содержат информацию об ответах на специальные предложения от пяти тысяч клиентов компании. Эти данные включают в себя большое число полей, содержащих сведения о возрасте, занятости и доходе клиентов, а также статистику их телефонных звонков. Три поля "назначения" показывают, откликнулся ли клиент на каждое из трех предложений. Компания хочет использовать эти данные для помощи в предсказаниях, какие клиенты наиболее вероятно откликнутся на аналогичные предложения в будущем.

Этот пример использует поток *featureselection.str*, в котором используется файл данных *customer_dbase.sav*. Эти файлы находятся в каталоге *Demos* любой установки IBM SPSS Modeler. Доступ к ним можно получить из группы программ IBM SPSS Modeler в меню Пуск Windows. Файл *featureselection.str* находится в каталоге *streams*.

Этот пример фокусируется только на одном их предложений как на целевом. Здесь используется узел выращивания дерева CHAID для разработки модели, описывающей, какие покупатели наиболее вероятно откликнутся на рекламную кампанию. Здесь подчеркивается различие двух подходов:

- Без выбора характеристик. Все поля предикторов в наборе данных используются как входные поля дерева CHAID.
- С выбором характеристик. Узел Выбор характеристик используется для выбора первых 10 предикторов. Затем они станут входными полями дерева CHAID.

При сравнении двух получающихся моделей дерева видно, как выбор характеристик приводит к эффективным результатам.

Построение потока

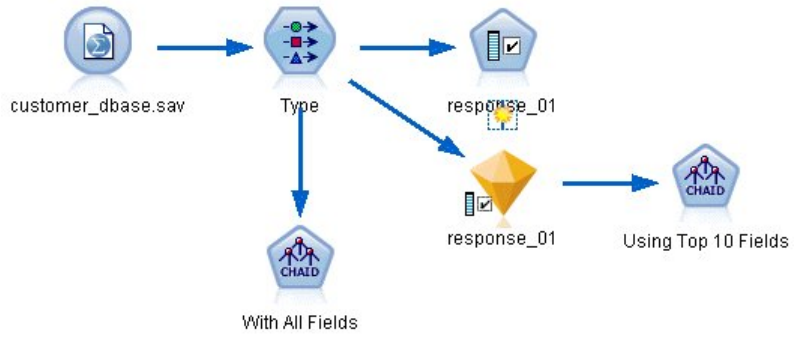


Рисунок 102. Пример потока выбора характеристик

1. Поместите узел источника файла статистики на пустой холст потока. Укажите для этого узла файл с примером данных *customer_dbase.sav*, доступный в каталоге *Demos* вашего каталога установки IBM SPSS Modeler. Можно также открыть файл с примером потока *featureselection.str* в каталоге *streams*.)
2. Добавьте узел Тип. На вкладке Типы прокрутите страницу вниз и измените роль для *response_01* на *Назначение*. Измените на *Нет* роли для других полей ответа (*response_02* и *response_03*), а также для ID покупателя (*custid*) наверху списка. Оставьте значением ролей *Вход* для всех остальных полей и нажмите кнопку **Прочитать значения**, а затем кнопку **OK**.

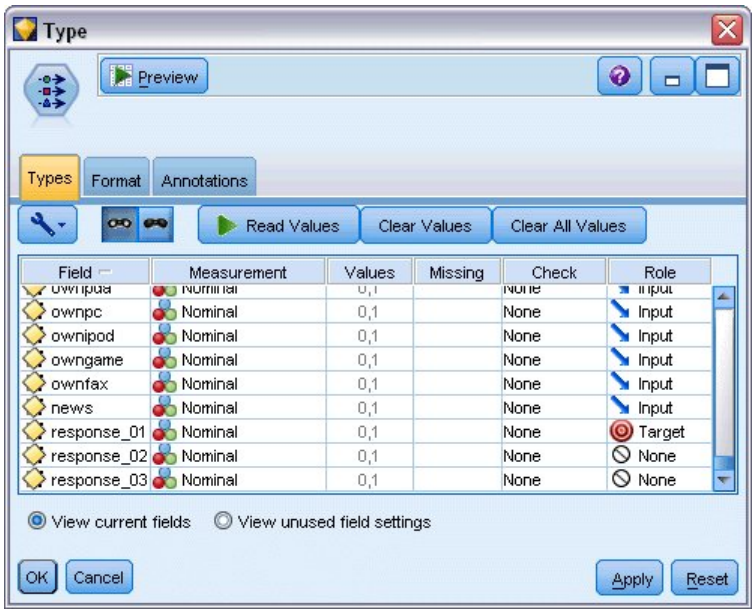


Рисунок 103. Добавление узла Тип

3. Добавить в поток узел моделирования выбора характеристик. В этом узле можно указать правила и критерии для экранирования, или дисквалификации, полей.
4. Запустите поток, чтобы создать слепок модели выбора характеристик.
5. Щелкните правой кнопкой мыши по слепку модели в потоке или на палитре Модели и выберите опцию **Изменить** или **Обзор**, чтобы просмотреть результаты.

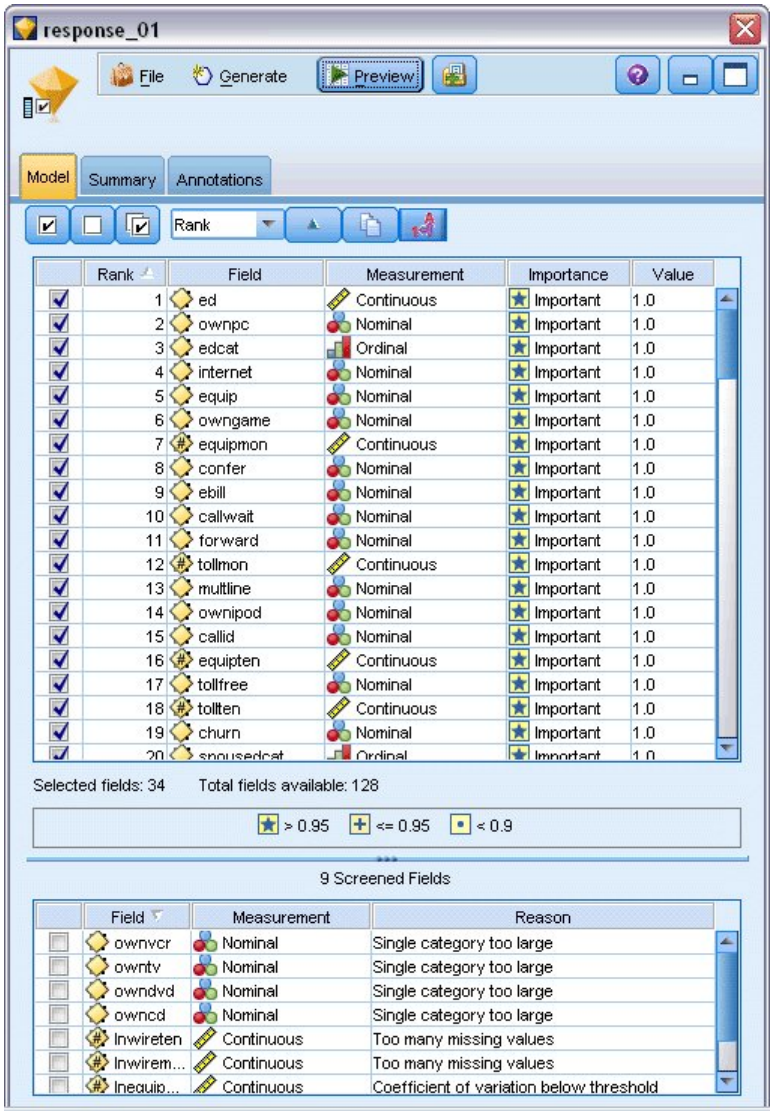


Рисунок 104. Вкладка Модель в слепке модели выбора характеристик

На верхней панели показываются поля, оказавшиеся полезными для предсказания. Они ранжированы по своей важности. На нижней панели показано, какие поля были экранированы при анализе и почему. Изучая поля на верхней панели, можно решить, какие из них использовать в последующих сеансах моделирования.

6. Теперь можно выбрать поля для использования нисходящего потока. Хотя исходно 34 поля были определены как важные, мы хотим сократить набор предикторов еще сильнее.
7. Выберите только первые 10 предикторов, используя галочки в первом столбце, чтобы отменить выбор нежелательных предикторов. (Щелкните по галочке в строке 11, удерживайте нажатой клавишу Shift и щелкните по галочке в строке 34.) Закройте слепок модели.
8. Для сравнения результатов без выбора характеристик необходимо добавить в поток два узла моделирования CHAID: один узел, использующий выбор характеристик, и один, не использующий.
9. Соедините один узел CHAID с узлом Тип, а другой со слепком модели выбора характеристик.
10. Откройте каждый узел CHAID, перейдите на вкладку Опции построения и убедитесь, что на панели Цели выбраны опции **Построить новую модель**, **Построить одно дерево** и **Запустить интерактивный сеанс**.

На панели Основы убедитесь, что для **Максимальной глубины дерева** задано значение 5.

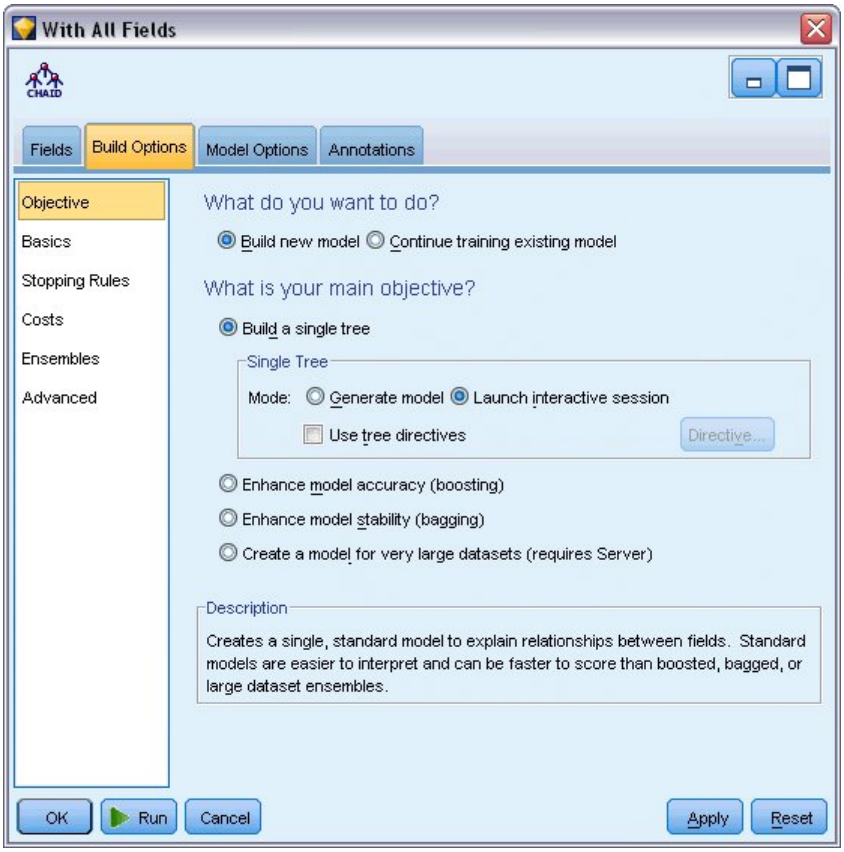


Рисунок 105. Целевые параметры для узла моделирования CHAID для всех полей предикторов

Построение модели

1. Выполните узел CHAID, использующий все предикторы набора данных (он соединен с узлом Тип). При выполнении обратите внимание, сколько времени потребовалось. В окне результатов появится таблица.
2. Выберите пункт меню **Дерево > Вырастить дерево**, чтобы вырастить и вывести развернутое дерево.

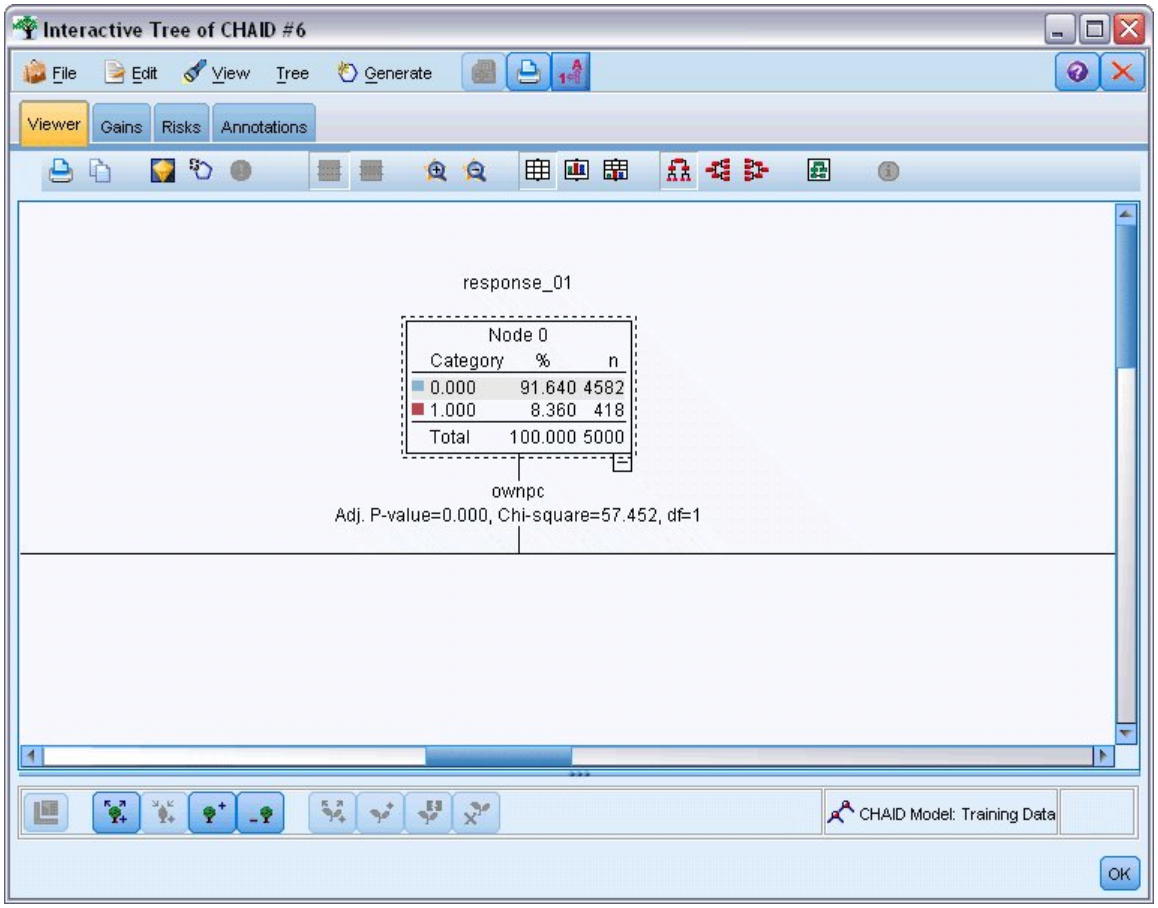


Рисунок 106. Выращивание дерева в построителе деревьев

3. Теперь сделайте то же самое для другого узла CHAID, который использует только 10 предикторов. Снова вырастите дерево, когда открыт построитель деревьев.

Вторая модель должна выполняться быстрее первой. Так как набор данных сравнительно невелик, различие во времени выполнения может составить несколько секунд; но для крупных реальных наборов данных различие может быть очень существенным - минуты и даже часы. Использование выбора характеристик может существенно ускорить обработку.

Кроме этого, у второго дерева меньше узлов, чем у первого. Его проще осмыслить. Но прежде чем принять решение о его использовании, нужно выяснить, насколько оно эффективно, и сравнить его с моделью, использующей все предикторы.

Сравнение результатов

Для сравнения двух результатов требуется мера эффективности. Для этого мы будем использовать вкладку Выигрыш в построителе деревьев. Посмотрим на показатель **рост**, измеряющий, насколько более вероятно записи из узла попадут в категорию назначения по сравнению со всеми записями в наборе данных. Например, значение роста в 148% обозначает, что записи из узла с вероятностью в 1,48 раза большей попадают в категорию назначения, чем все записи в наборе данных. Рост указывается в столбце *Индекс* на вкладке Выигрыш.

1. В построителе деревьев для всего набора предикторов перейдите на вкладку Выигрыш. Измените категорию назначения на 1.0. Измените вывод на квартили первым нажатием кнопки панели инструментов Квантили. Затем выберите **Квартиль** из выпадающего списка справа от этой кнопки.

2. Повторите эту процедуру в построителе деревьев для набора 10 предикторов, чтобы у вас было две аналогичные таблицы Выигрыш для сравнения, как показано на следующих рисунках.

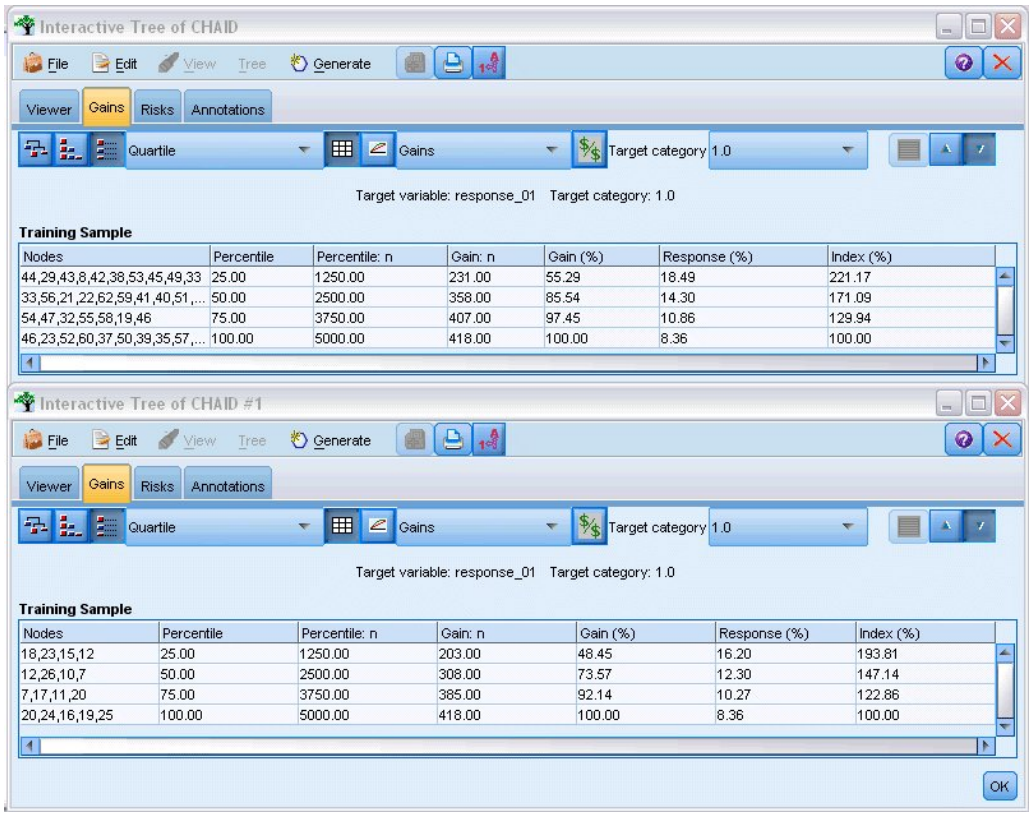


Рисунок 107. Диаграммы выигрыша для двух моделей CHAID

Каждая таблица Выигрыш группирует терминальные узлы для своего дерева в квартили. Чтобы сравнить эффективность двух моделей, посмотрите на рост (значение *Индекс*) для верхней квартили в каждой таблице.

Когда включены все предикторы, модель показывает рост 221%. Таким образом, наблюдения с характеристиками в этих узлах в 2,2 раза более вероятны для получения отклика на целевую рекламную кампанию. Чтобы просмотреть эти характеристики, щелкните для выбора по верхней строке. Затем переключитесь на вкладку Средство просмотра, где соответствующие узлы теперь выделены черным цветом. Следуйте по дереву вплоть до каждого терминального узла, чтобы увидеть, как расщеплялись предикторы. Одна верхняя квартиль содержит 10 узлов. При переводе на модели оценок с реальными показателями это означает, что для 10 различных профилей покупателей будут сложности с управлением.

При включении только 10 верхних предикторов (определенных при выборе возможностей), рост составляет около 194%. Хотя эта модель не так хороша, как модель, использующая все предикторы, она определенно полезна. Здесь верхняя квартиль включает в себя только 4 узла, то есть она проще. Тем самым, мы можем утверждать, что модель выбора характеристик предпочтительна по сравнению с моделью со всеми предикторами.

Итог

Рассмотрим преимущества выбора характеристик. Использование меньшего числа предикторов менее затратно. Это означает, что вы должны собирать, обрабатывать и вводить в модель меньше данных. Сокращается время вычислений. В этом примере, даже с учетом шага выбора характеристик построение модели было существенно быстрее при меньшем числе предикторов. Для гораздо более крупных реальных наборов данных экономия времени окажется очень существенной.

Использование меньшего числа предикторов упрощает скоринг. Как показано в примере, можно идентифицировать только четыре профиля покупателей, которые вероятно откликнутся на рекламную кампанию. Обратите внимание на то, что при большем числе предикторов возникает риск переобучения вашей модели. Более простую модель лучше обобщается на другие наборы данных (хотя для уверенности это нужно проверить).

Чтобы работал выбор характеристик вам могло потребоваться использовать алгоритм выращивания дерева, позволяющий идентифицировать на дереве наиболее важные предикторы. На самом деле алгоритм CHAID часто используется для этой цели, и даже можно наращивать дерево уровень за уровнем, чтобы контролировать его глубину и сложность. Однако использовать узел Выбор характеристик быстрее и проще. Здесь все предикторы ранжируются на одном быстром шаге, что позволяет быстро идентифицировать самые важные поля. Можно изменять также число предикторов для включения в модель. Вы могли бы легко запустить этот пример снова, используя 15 или 20 предикторов вместо 10 и сравнивая результаты для определения оптимальной модели.

Глава 10. Сокращение длины входной строки данных (узел повторной классификации)

Сокращение длины входной строки данных (переклассификация)

Для биномиальной логистической регрессии и моделей автоклассификации, включающих в себя модель биномиальной логистической регрессии, строковые поля ограничены максимальной длиной в восемь символов. Строки длиннее восьми символов можно перекодировать, используя узел Переклассификация.

Этот пример использует поток *reclassify_strings.str*, в котором используется файл данных *drug_long_name*. Эти файлы находятся в каталоге *Demos* любой установки IBM SPSS Modeler. Доступ к ним можно получить из группы программ IBM SPSS Modeler в меню Пуск Windows. Файл *reclassify_strings.str* находится в каталоге *streams*.

Предмет этого примера - небольшая часть потока, на которой мы покажем, какого сорта ошибки могут возникнуть при наличии более длинных строк и как использовать узел Переклассификация, чтобы изменить подробности строки до приемлемой длины. Хотя в этом примере используется узел Биномиальная логистическая регрессия, он также применим при использовании узла Автоклассификация для генерирования модели биномиальной логистической регрессии.

Переклассификация данных

1. Используя узел источников файлов переменных, соединитесь с набором данных *drug_long_name* в папке *Demos*.

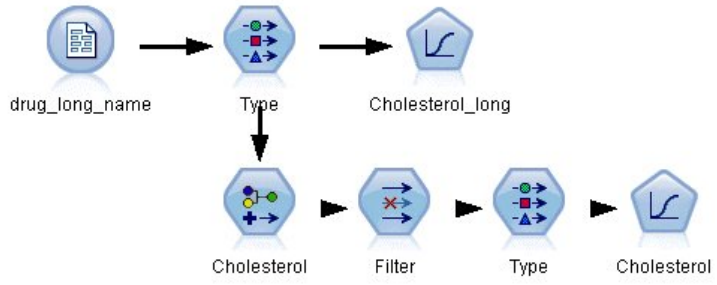


Рисунок 108. Поток примера с переклассификацией строк для биномиальной логистической регрессии

2. Добавьте узел Тип к узлу Источник и выберите в качестве поля назначения **Cholesterol_long**.
3. Добавьте узел Логистической регрессии к узлу Тип.
4. В узле Логистическая регрессия щелкните по вкладке Модель и выберите процедуру **Биномиальная**.

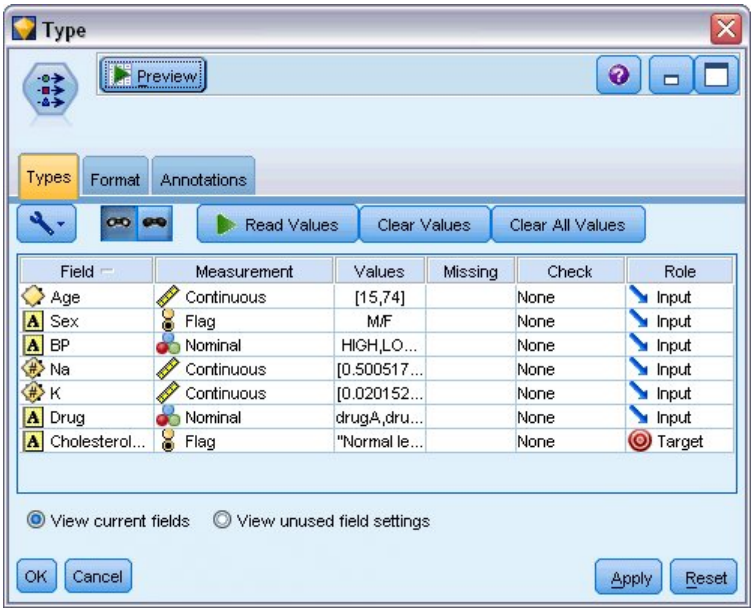


Рисунок 109. Длинные строковые значения в поле "Cholesterol_long"

5. При выполнении узла Логистическая регрессия в потоке *reclassify_strings.str* появляется сообщение об ошибке, предупреждающее, что строковые значения поля **Cholesterol_long** слишком длинные.

   Если встретится сообщение об ошибке такого типа, для изменения данных используйте описанную далее в этом примере процедуру.

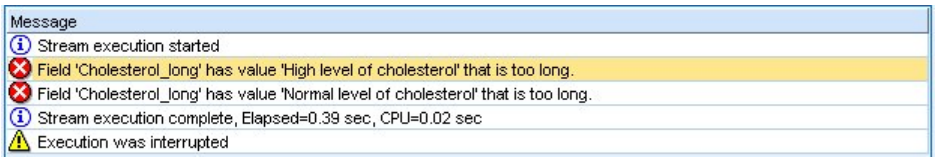


Рисунок 110. При выполнении узла узла биномиальной логистической регрессии выводится сообщение об ошибке

6. Добавьте узел Переклассификация к узлу Тип.
7. В поле Переклассификация выберите **Cholesterol_long**.
8. Введите **Cholesterol** как новое имя поля.
9. Нажмите кнопку **Получить**, чтобы добавить значения **Cholesterol_long** в столбец исходных значений.
10. В столбце новых значений введите **Высокий** рядом с исходным значением **Высокий уровень холестерина** и **Норма** рядом с исходным значением **Нормальный уровень холестерина**.

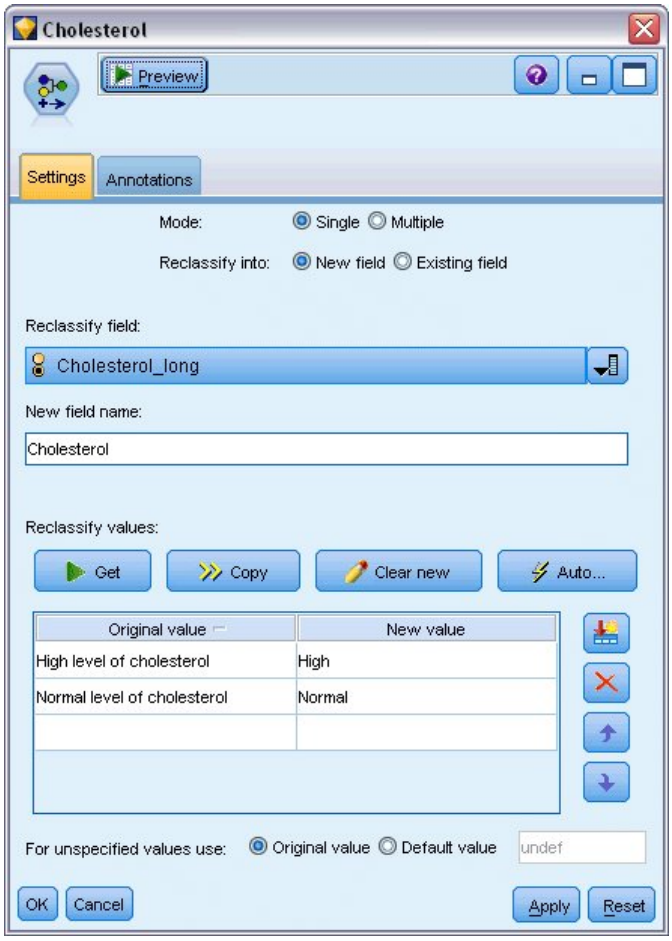


Рисунок 111. Переклассификация длинных строк

11. Добавьте узел Фильтр к узлу Переклассифицировать.
12. Щелкните в столбце Фильтр, чтобы удалить **Cholesterol_long**.

Рисунок 112. Фильтрация поля "Cholesterol_long" из данных

13. Добавьте узел Тип к узлу Фильтр и выберите в качестве поля назначения **Cholesterol**.

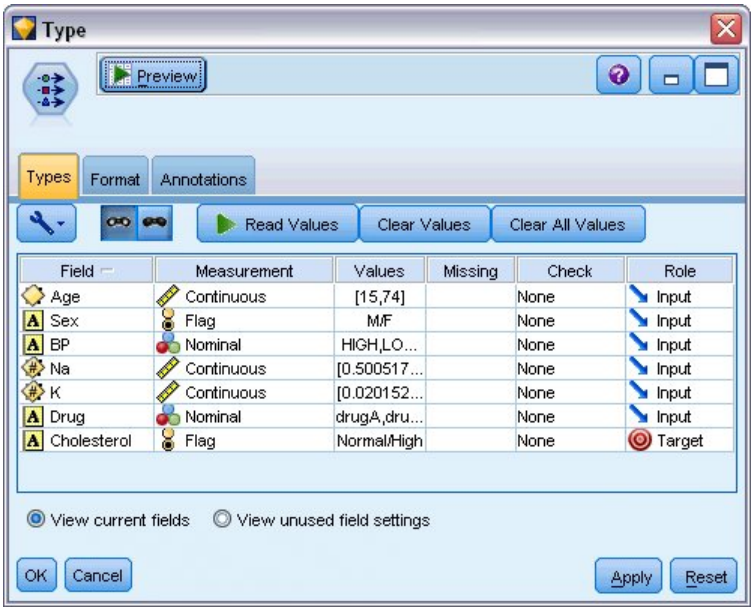


Рисунок 113. Короткие строковые значения в поле "Cholesterol"

14. Добавьте Логистический узел к узлу Тип.
15. В Логистическом узле щелкните по вкладке Модель и выберите процедуру **Биномиальная**.
16. Теперь можно выполнить Биномиальный логистический узел и сгенерировать модель без вывода сообщения об ошибке.

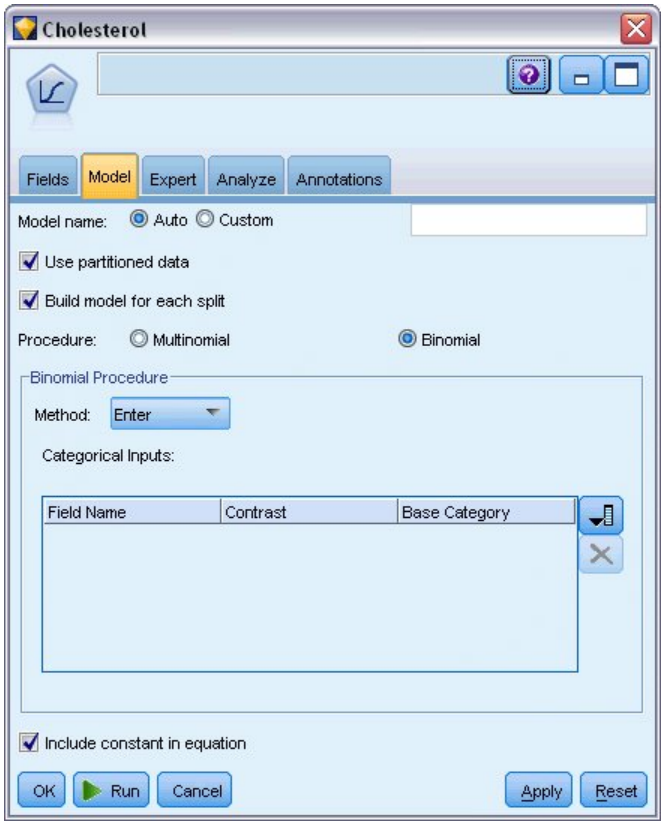


Рисунок 114. Выбор Биномиальной процедуры

В этом примере показана только часть потока. Если вам потребуется дополнительная информация о типах потоков, в которых может потребоваться переклассификация длинных строк, доступны следующие примеры:

- Узел автоклассификации. Дополнительную информацию смотрите в разделе “Моделирование ответа покупателя (автоматический классификатор)” на стр. 39.
- Узел Биномиальная логистическая регрессия. Дополнительную информацию смотрите в разделе Глава 13, “Отток клиентов в сфере телекоммуникаций (Биномиальная логистическая регрессия)”, на стр. 141.

Дополнительную информацию о том, как использовать IBM SPSS Modeler, такую как руководство пользователя, справочник по узлам и руководство по алгоритмам, смотрите в каталоге *\Documentation* установочного диска.

Глава 11. Моделирование откликов клиентов (Список решений)

Алгоритм Список решений генерирует правила, определяющие более или менее высокую вероятность данного бинарного выхода (да или нет). Модели списка решений широко используются в управлении связями с клиентами (например, в колл-центрах и в прикладных программах маркетинга).

Этот пример основан на деятельности вымышленной компании, которая хочет достичь более выгодных результатов в будущих маркетинговых кампаниях, подготовив правильное предложение для каждого клиента. В частности, в этом примере используется модель Дерево решений для определения характеристик покупателей, которые наиболее вероятно откликнутся положительно, на основе предыдущих рекламных кампаний, и для генерирования списка рассылки на основании результатов модели.

Модели Список решений в частности хорошо подходят для интерактивного моделирования, позволяя настраивать параметры в модели и сразу же видеть результаты. При другом подходе, когда допускается автоматическое создание нескольких моделей и ранжирование результатов, вместо этого можно использовать узел автоклассификации.

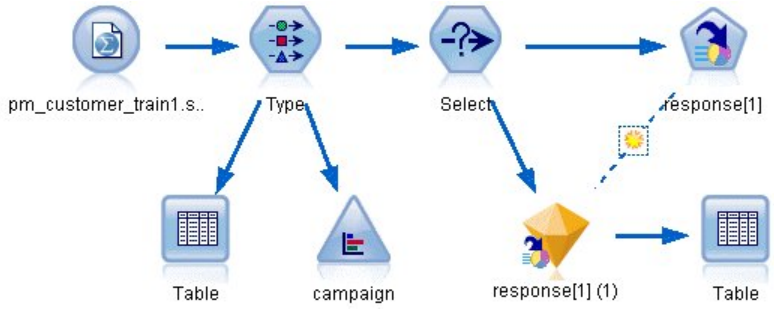
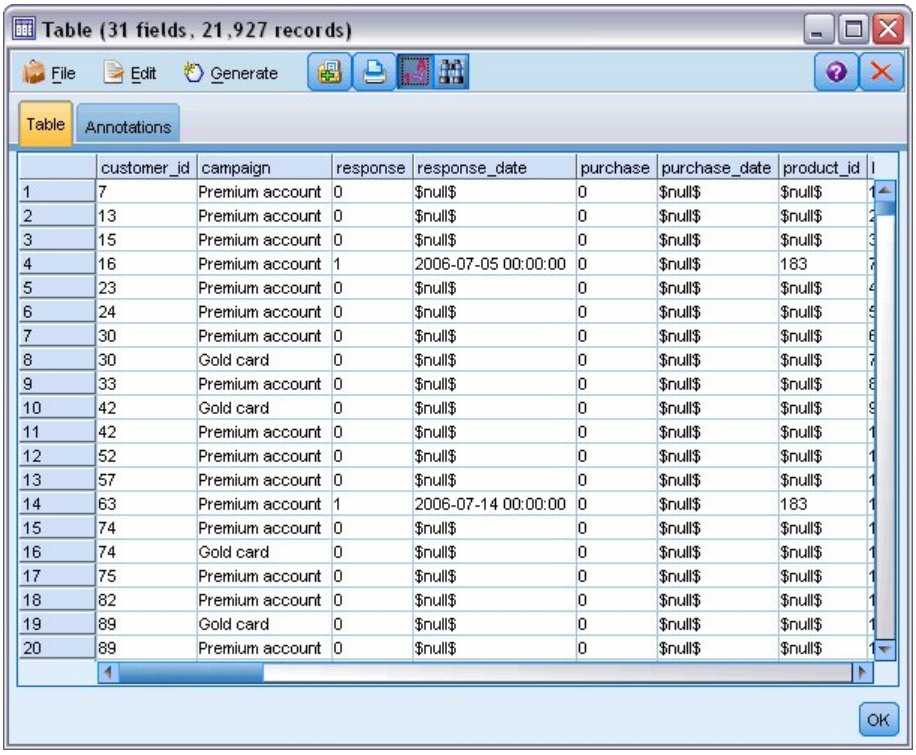


Рисунок 115. Поток примера списка решений

Этот пример использует поток *pm_decisionlist.str*, который ссылается на файл данных *pm_customer_train1.sav*. Эти файлы находятся в каталоге *Demos* любой установки IBM SPSS Modeler. Доступ к ним можно получить из группы программ IBM SPSS Modeler в меню Пуск Windows. Файл *pm_decisionlist.str* находится в каталоге *streams*.

Хронологические данные

В файле *pm_customer_train1.sav* есть хронологические данные, отслеживающие предложения, сделанные для конкретных покупателей в прошлых кампаниях, на что указывает значение в поле *кампания*. Наибольшее число записей подпадает под кампанию *Премиальная учетная запись*.



	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

Рисунок 116. Данные о предыдущих спецпредложениях

Значения поля *campaign* (кампания) в данных закодированы числовыми метками, как определено в узле Тип (например, *2 = учетная запись Premium*). Вывод меток значений можно переключать в таблице при помощи панели инструментов.

Файл также содержит ряд полей с демографической и финансовой информацией о каждом клиенте, которые можно использовать при построении или "обучении" модели, прогнозирующей показатели отклика для различных групп с учетом конкретных характеристик.

Построение потока

1. Добавьте узел Файл статистики со ссылкой на *pm_customer_train1.sav* в папке *Demos* вашей установки IBM SPSS Modeler. (Вместо этой папки в пути файла можно использовать ярлык `$CLEO_DEMOS/`.)

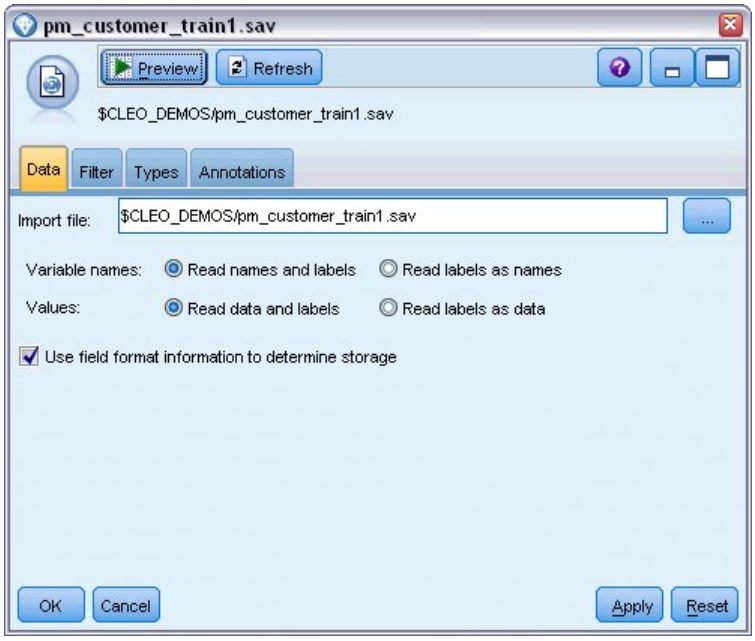


Рисунок 117. Чтение в данных

2. Добавьте узел Тип и выберите *отклик* как целевое поле (Роль = **Назначение**). Задайте для этого поля тип измерений **Флаг**.

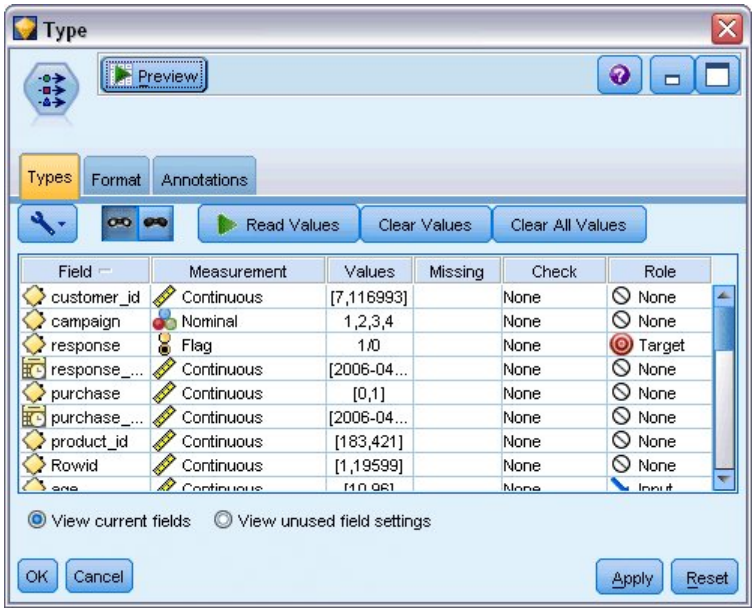


Рисунок 118. Задание уровня измерения и роли

3. Для следующих полей задайте роль **Нет**: *customer_id*, *campaign*, *response_date*, *purchase*, *purchase_date*, *product_id*, *Rowid*, and *X_random* (ID заказчика, кампания, дата отклика, закупка, дата закупки, ID товара, ID строки и X_random). Все эти поля используются при обработке данных, но не будут использоваться в создании фактической модели.
4. Нажмите кнопку **Прочитать значения** в узле Тип, чтобы гарантировать инициирование всех значений.

Хотя эти данные включают в себя информацию о четырех разных кампаниях, вы будете всякий раз фокусироваться на одной кампании. Поскольку наибольшее число записей попадают в кампанию Premium (в

данных кодируется как *campaign = 2*), можно оставить в потоке только эти данные, использовав узел выбора.

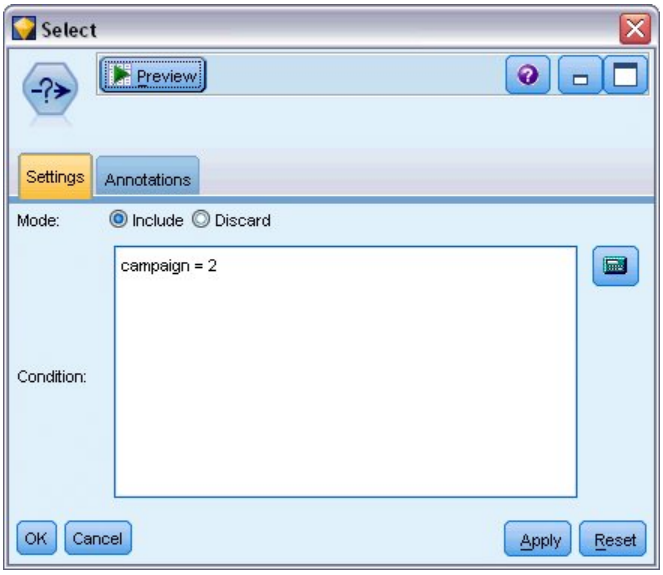


Рисунок 119. Выбор записей для одной кампании

Создание модели

1. Присоедините узел Список решений к потоку. На вкладке Модель задайте для **Значения назначения** значение 1, чтобы отметить выходные поля, для которых вы хотите выполнить поиск. В данном случае вы будете искать пользователей, ответивших *Да* на предыдущее предложение.

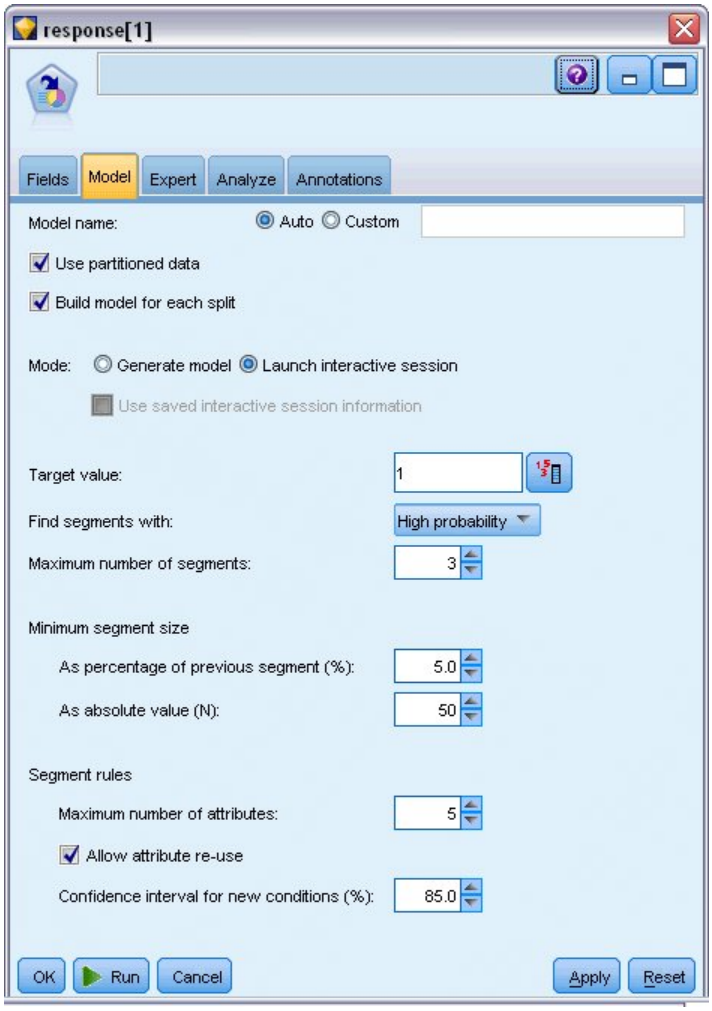


Рисунок 120. Узел Список решений, вкладка Модель

2. Выберите **Запустить интерактивный сеанс**.
3. Чтобы в этом примере модель оставалась простой, задайте для максимального числа сегментов значение 3.
4. Измените доверительный интервал для новых условий на 85%.
5. На вкладке Эксперт задайте для **Режима** значение **Эксперт**.

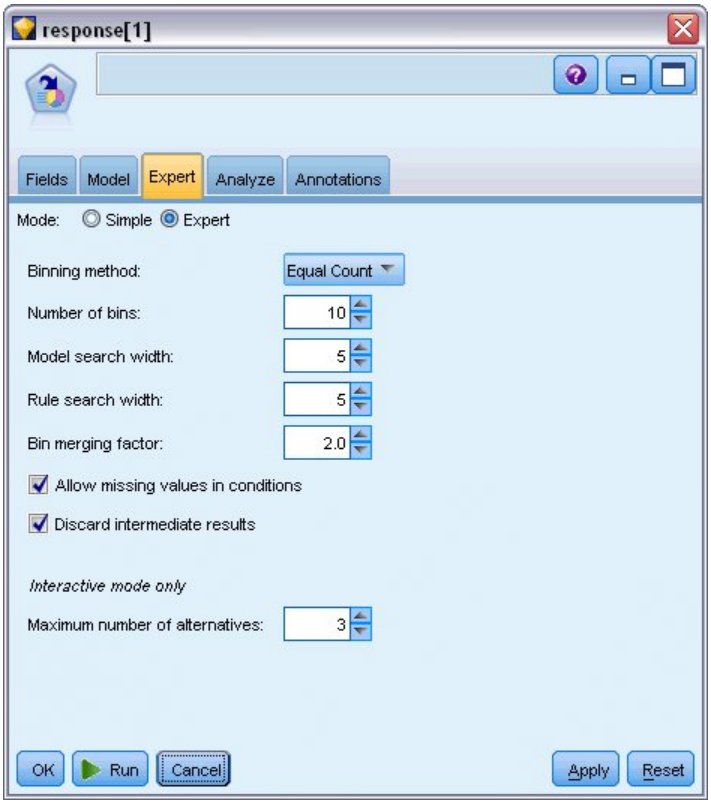


Рисунок 121. Узел Список решений, вкладка Эксперт

6. Увеличьте **Максимальное число альтернативных вариантов** до 3. Эта опция работает совместно с параметром **Запустить интерактивный сеанс**, выбранным на вкладке Модель.
7. Нажмите кнопку **Выполнить**, чтобы вывести средство просмотра интерактивного списка.

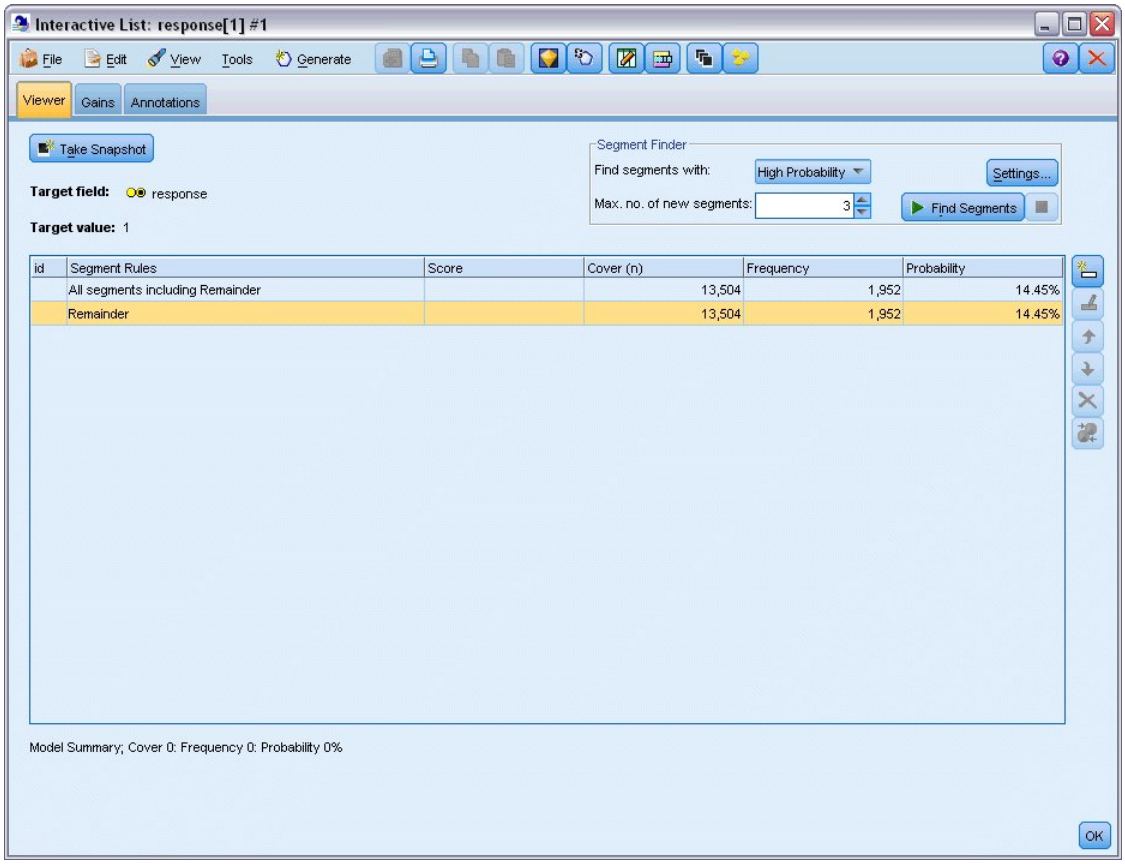


Рисунок 122. Средство просмотра интерактивного списка

Так как никакие сегменты еще не определены, все записи попадут в остаток. Из 13504 записей в выборке есть 1952 ответов *Да*, что дает коэффициент попадания 14,45%. Вы хотите увеличить этот коэффициент, определив сегменты покупателей, которые более (или менее) вероятно дадут желательный отклик.

8. Выберите в меню средства просмотра интерактивного списка пункт:

   Инструменты > Найти сегменты

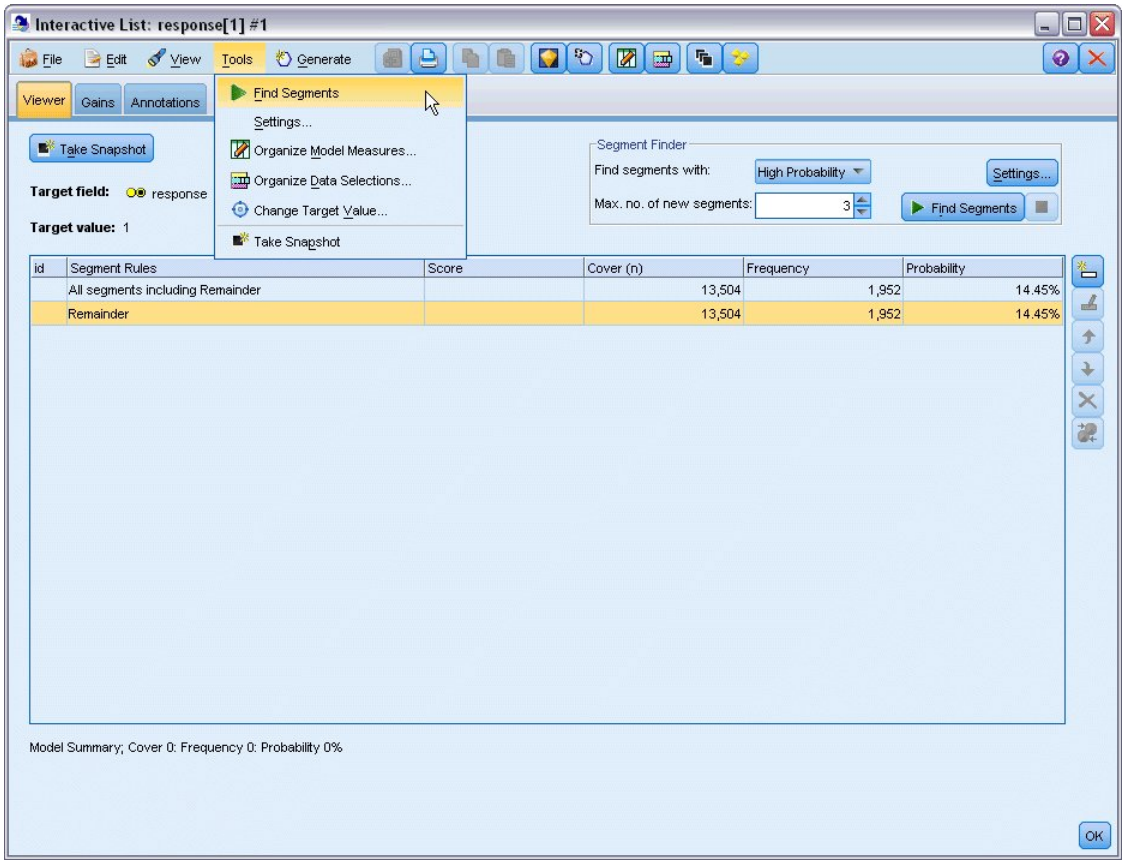


Рисунок 123. Средство просмотра интерактивного списка

Это запускает задачу анализа данных по умолчанию на основании параметров, заданных вами в узле Список решений. Завершенная задача возвратит три альтернативные модели, перечисленные на вкладке Альтернативные варианты диалогового окна Альбомы моделей.

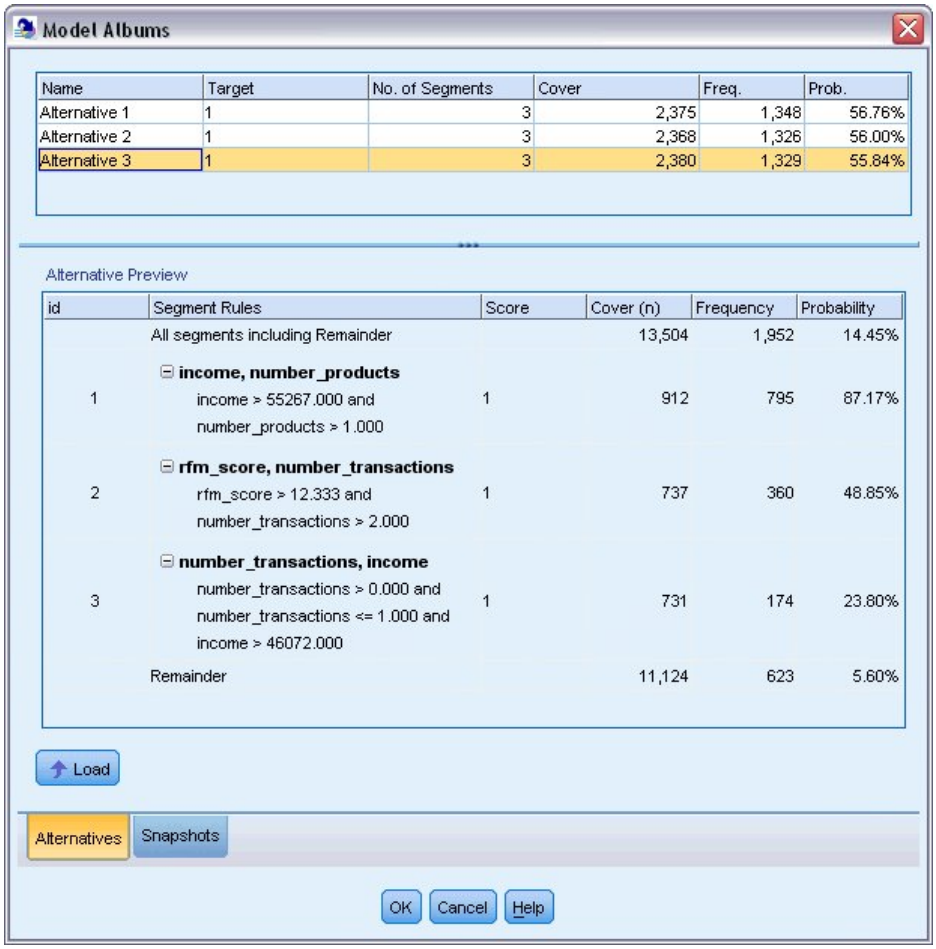


Рисунок 124. Доступные альтернативные модели

9. Выберите первый альтернативный вариант из списка; его подробности показаны на панели Предварительный просмотр альтернативных вариантов.

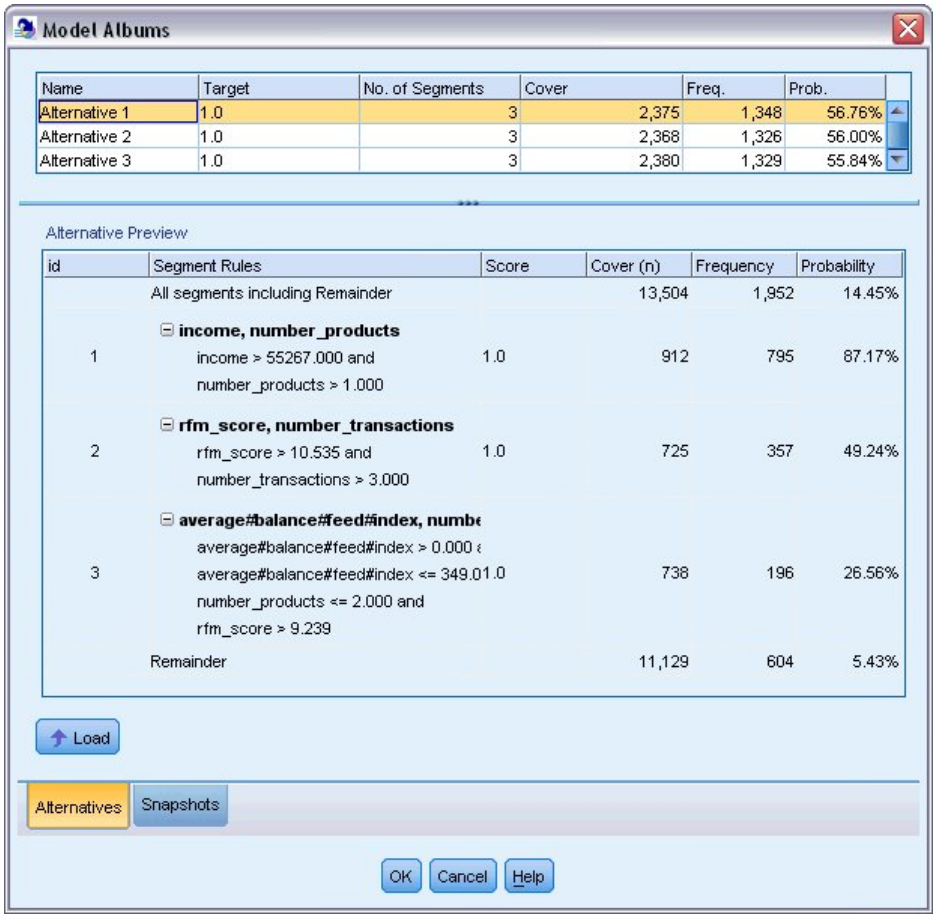


Рисунок 125. Выбрана альтернативная модель

На панели Предварительный просмотр альтернативных вариантов можно быстро просмотреть любое число альтернативных вариантов, не изменяя рабочую модель, что упрощает эксперименты с разными подходами.

Примечание: Чтобы получить лучшее представление модели, вам может потребоваться развернуть панель Предварительный просмотр альтернативных вариантов, как здесь показано. Это можно сделать, перетащив границы панели.

Используя правила на основе предикторов, таких как входные данные, число транзакций за месяц и оценка RFM, модель идентифицирует сегменты с коэффициентом отклика большим, чем в целом по выборке. Если объединить эти сегменты, согласно модели коэффициент попадания повысится до 56,76%. Однако эта модель покрывает только малую часть всей выборки, оставляя свыше 11000 записей, среди которых несколько сотен попаданий, в остатке. Хотелось бы, чтобы модель захватывала больше таких попаданий, по-прежнему исключая малозначимые сегменты.

10. Чтобы испробовать другой подход к моделированию, выберите в меню пункт:

    Инструменты > Параметры

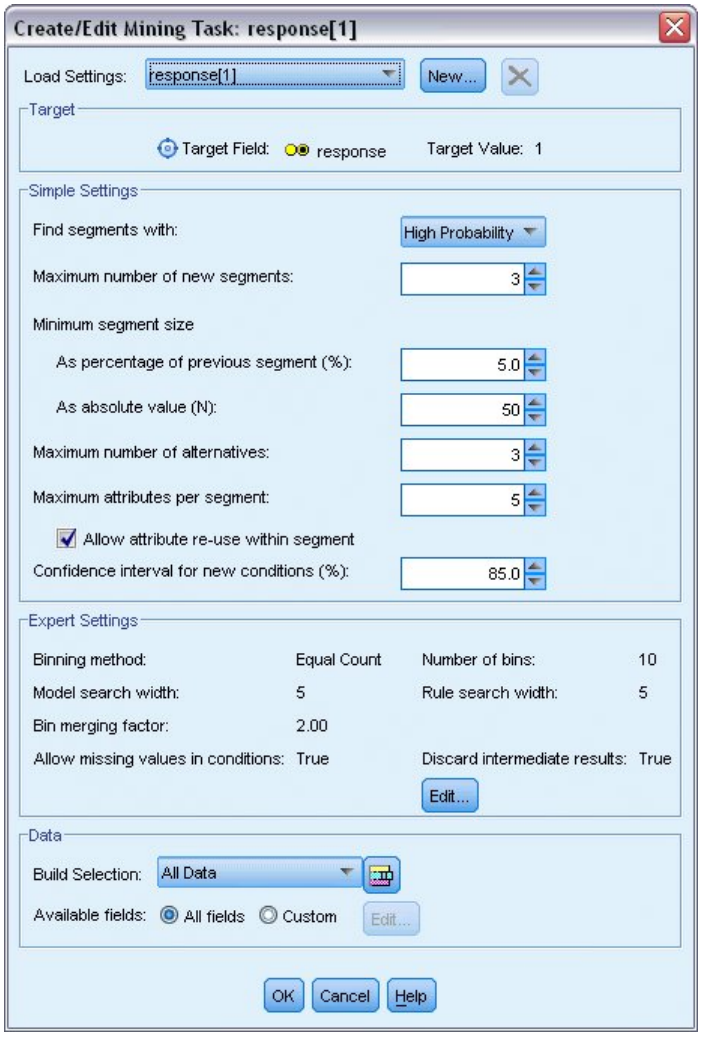


Рисунок 126. Диалоговое окно Создать/изменить задачу исследования данных

11. Нажмите кнопку **Создать** в правом верхнем углу, чтобы создать вторую задачу исследования данных и укажите *Down Search* как имя задачи в диалоговом окне Новые параметры.

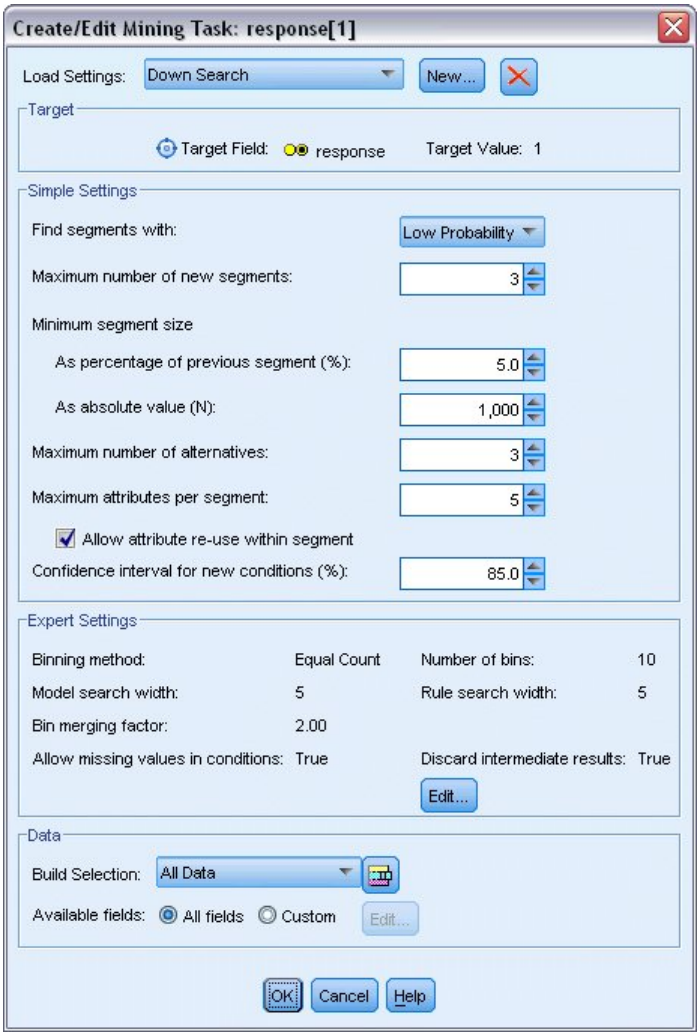


Рисунок 127. Диалоговое окно Создать/изменить задачу исследования данных

12. Для этой задачи измените направление поиска на **Низкая вероятность**. Это приведет к тому, что алгоритм будет искать сегменты с *наименьшими* показателями отклика, а не с наибольшими.
13. Увеличьте минимальный размер сегмента до 1000. Нажмите кнопку **OK**, чтобы вернуться к средству просмотра интерактивного списка.
14. В средстве просмотра интерактивного списка проверьте, что на панели *Средство поиска сегментов* выводятся подробности новой задачи, и щелкните по **Найти сегменты**.

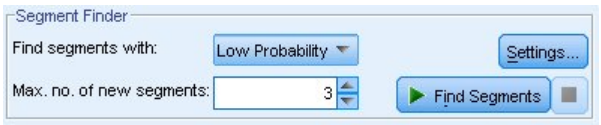


Рисунок 128. Найти сегменты в новой задаче исследования данных

Эта задача возвратит новый набор альтернативных вариантов, которые будут выведены на вкладке Альтернативные варианты диалогового окна Альбомы моделей, и их можно просмотреть так же, как предыдущие результаты.

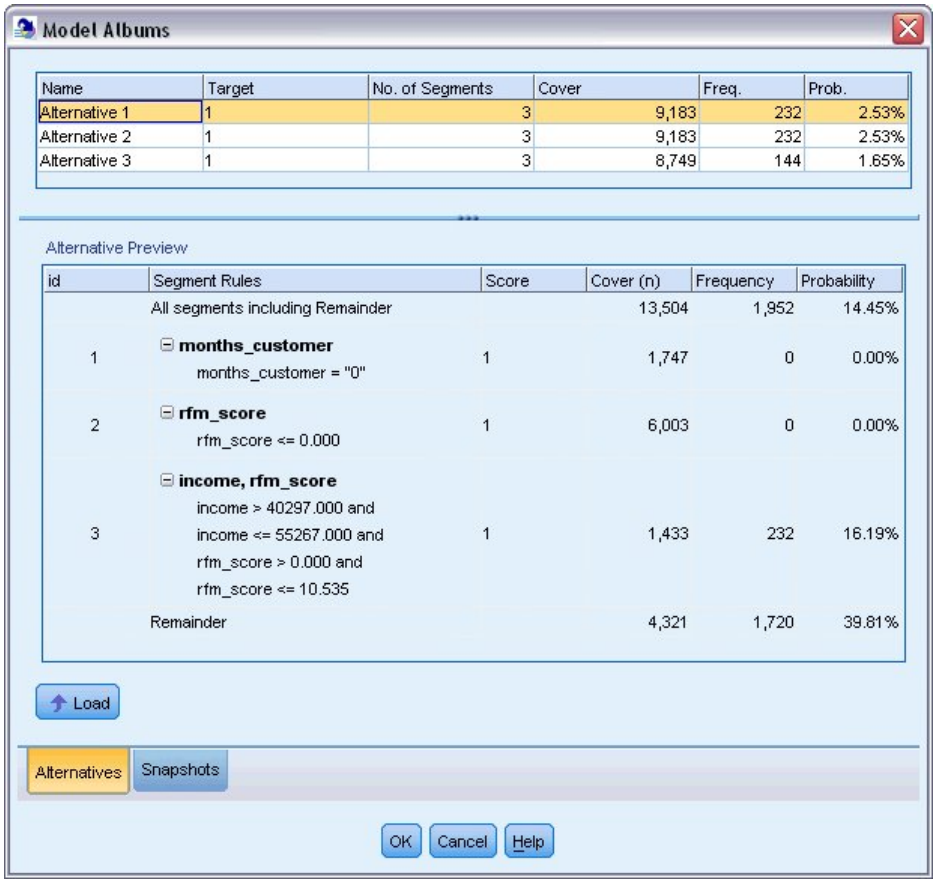


Рисунок 129. Результаты модели Down Search

На этот раз каждая модель идентифицирует сегменты с низкими, а не высокими, вероятностями отклика. Если посмотреть на первый альтернативный вариант, простое исключение таких сегментов увеличит коэффициент попадания для остатка до 39,81%. Это меньше, чем в рассмотренной ранее модели, но покрытие и общее число попаданий больше.

Комбинируя эти два подхода, то есть используя поиск сегментов с низкой вероятностью отклика для отсеивания неинтересных записей, а затем поиск сегментов с высокой вероятностью, можно улучшить этот результат.

15. Щелкните по **Загрузить**, чтобы сделать первый альтернативный вариант Down Search рабочей моделью, и нажмите кнопку **OK**, чтобы закрыть диалоговое окно Альбомы моделей.

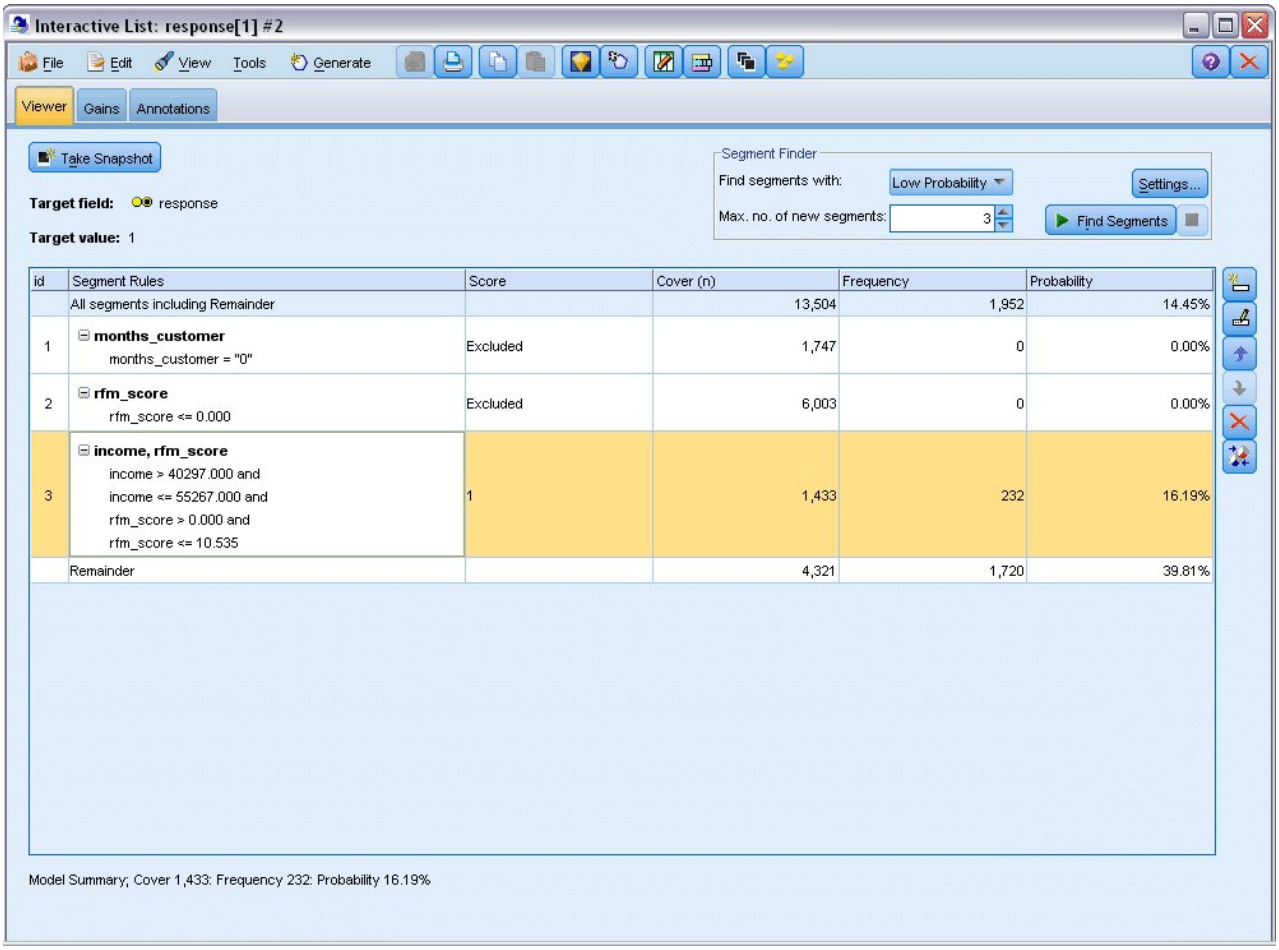


Рисунок 130. Исключение сегмента

16. Щелкните правой кнопкой мыши по каждому из двух первых сегментов и выберите опцию **Исключить сегмент**. Совместно эти два сегмента захватывают почти 8000 записей с нулевыми попаданиями, поэтому есть смысл исключить их из будущих предложений. (Для обозначения этого исключенные сегменты будут оцениваться значением null.)
17. Щелкните правой кнопкой мыши по третьему сегменту и выберите опцию **Удалить сегмент**. Со значением коэффициента попадания в 16,19% этот сегмент не очень отличается от базового уровня в 14,45%, поэтому он не добавит существенной информации, чтобы оставлять его на месте.

    Примечание: Удаление сегмента - это не то же самое, что его исключение. Исключение сегмента просто изменяет его оценку, а удаление полностью выводит сегмент из модели.

    После исключения самых малоценных сегментов можно выполнить поиск сегментов из остатка с большим вкладом.
18. Щелкните по строке остатка в таблице, чтобы выбрать его; и при этом следующая задача исследования данных будет применяться только к остатку.

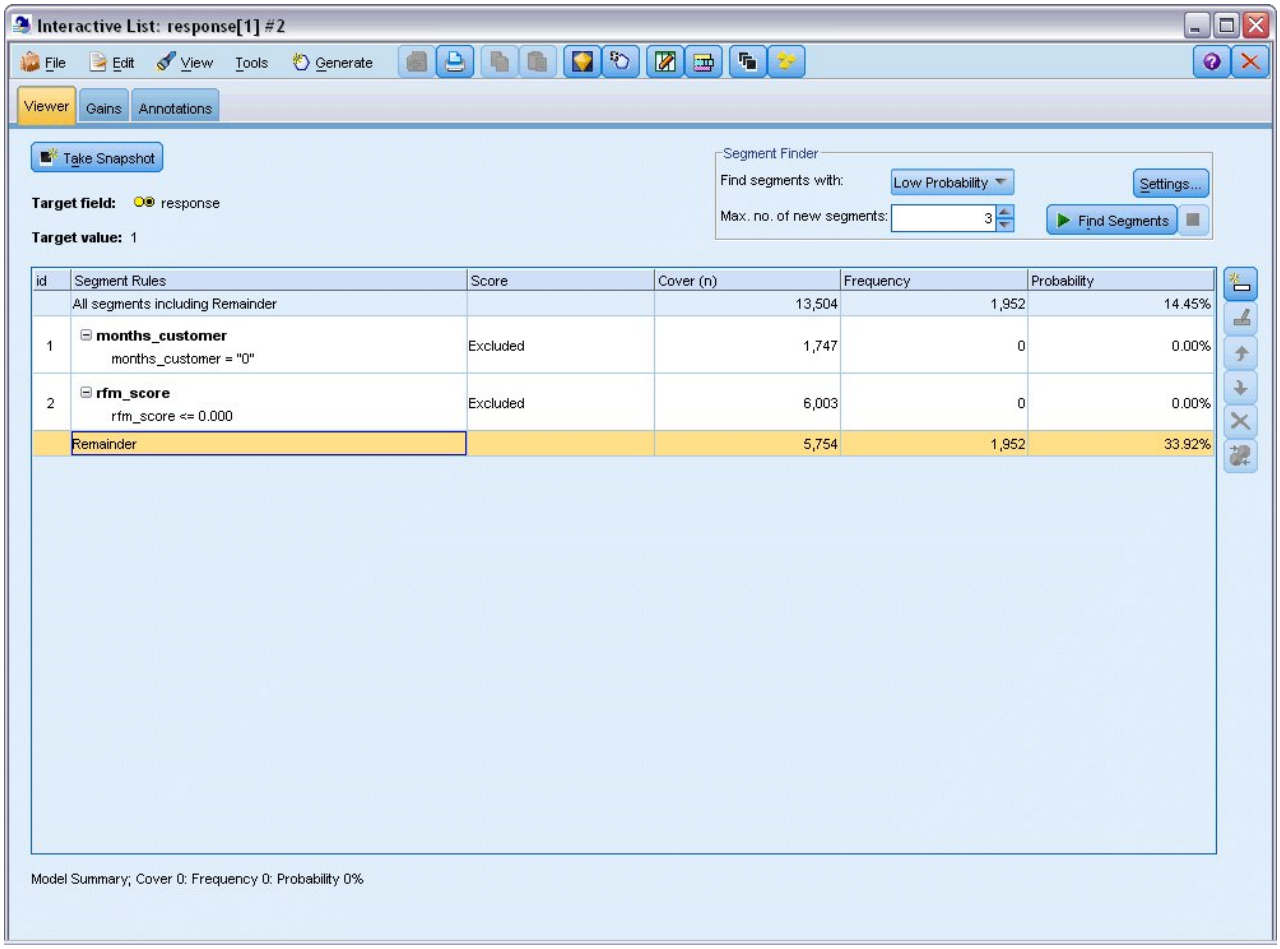


Рисунок 131. Выбор сегмента

19. При выбранном остатке щелкните по **Параметры**, чтобы снова открыть диалоговое окно Создать/изменить задачу исследования данных.
20. В верхней части окна для опции **Загрузить параметры** задайте задачу исследования данных по умолчанию: **response[1]**.
21. Измените **Простые параметры**, чтобы увеличить число новых сегментов до 5, а минимальный размер сегмента до 500.
22. Нажмите кнопку **OK**, чтобы вернуться к средству просмотра интерактивного списка.

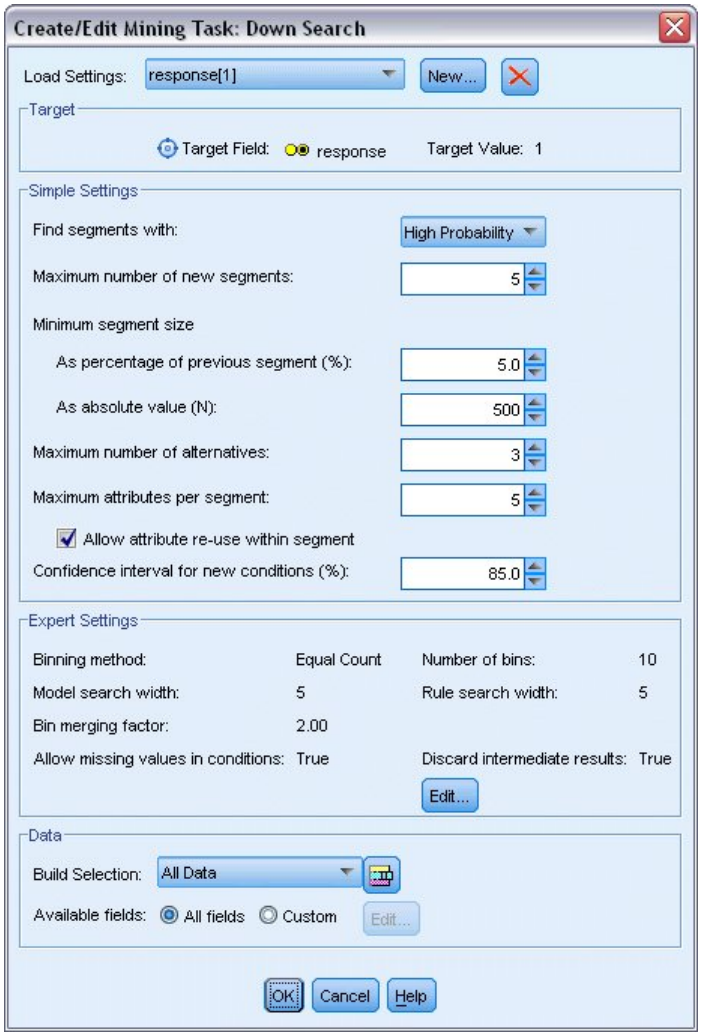


Рисунок 132. Выбор задачи исследования данных по умолчанию

23. Щелкните по **Найти сегменты**.

    При этом будет выведен еще один набор альтернативных моделей. Когда результаты одной задачи исследования данных используются для другой такой задачи, эти последние модели будут содержать смесь хорошо работающих и плохо работающих сегментов. Сегменты с низкими показателями откликов исключены, то есть они оцениваются значением null, а включенные сегменты будут оцениваться значением 1. Общая статистика отображает эти включения, когда первая альтернативная модель показывает коэффициент попадания 45,63% с более высоким покрытием (1577 из 3456 записей), чем в любой из предыдущих моделей.

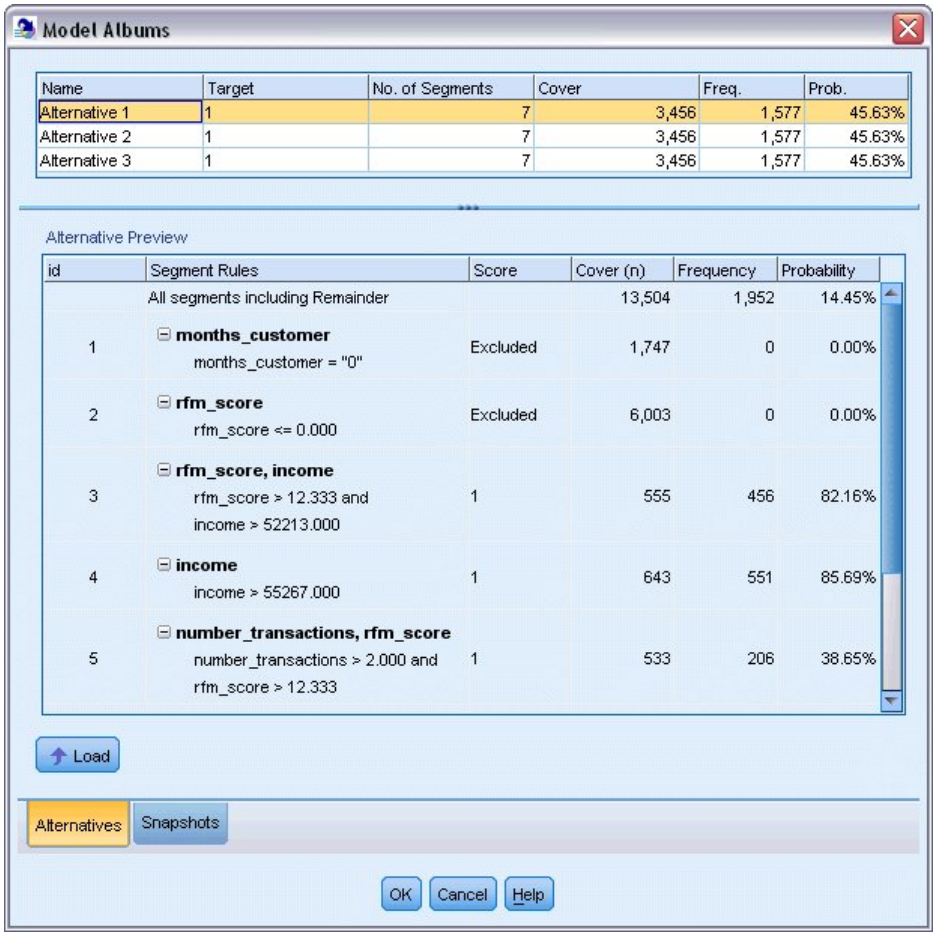


Рисунок 133. Альтернативные варианты для комбинированной модели

24. Предварительно просмотрите первый альтернативный вариант и щелкните по **Загрузить**, чтобы сделать его рабочей моделью.

Вычисление пользовательских показателей с использованием Excel

1. Чтобы чуть подробнее разобраться с работой модели с практической точки зрения, в меню Инструменты выберите опцию **Организовать показатели модели**.

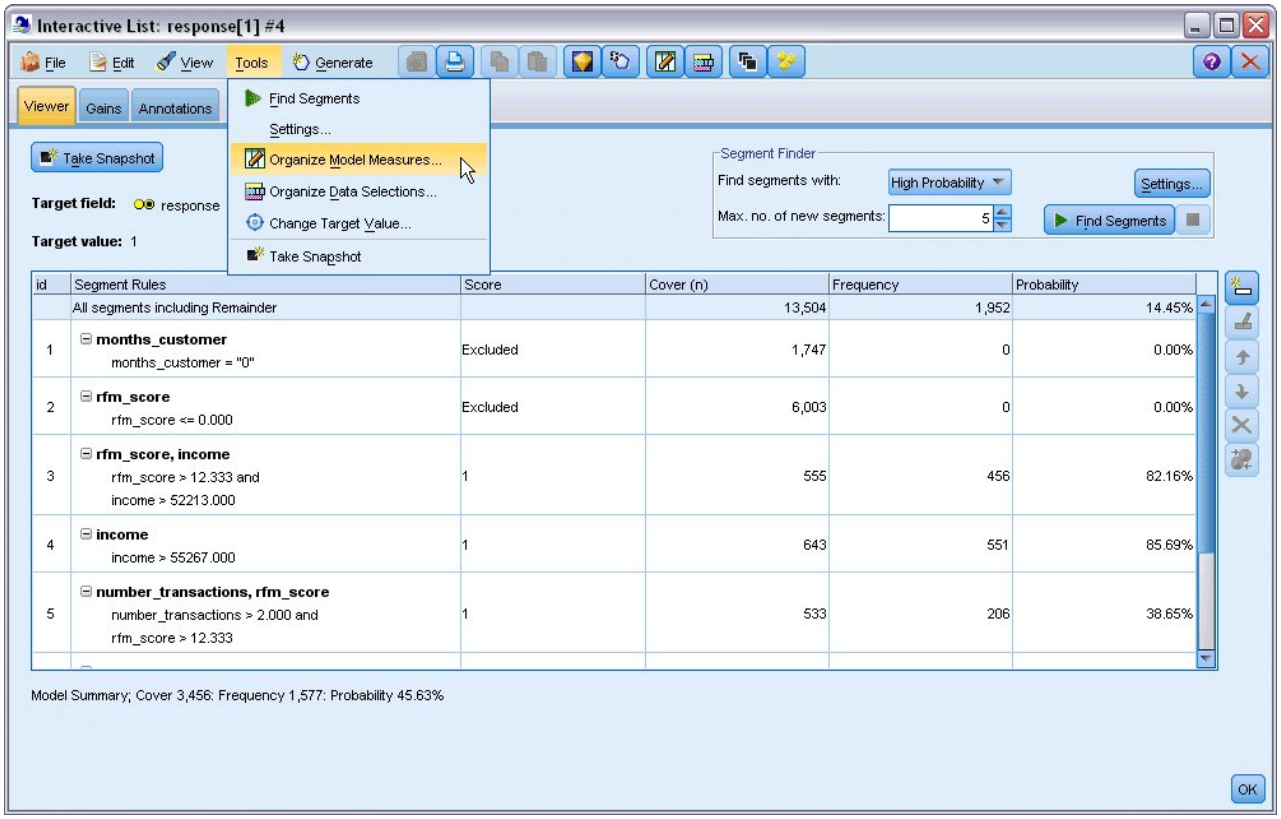


Рисунок 134. Организация показателей модели

В диалоговом окне Организовать показатели модели можно выбрать показатели (или столбцы), которые будут выводиться в средстве просмотра интерактивного списка. Можно указать также, будут ли вычисляться показатели для всех записей или только для выбранного подмножества, и кроме этого выбрать вывод круговой диаграммы вместо чисел, где это возможно.

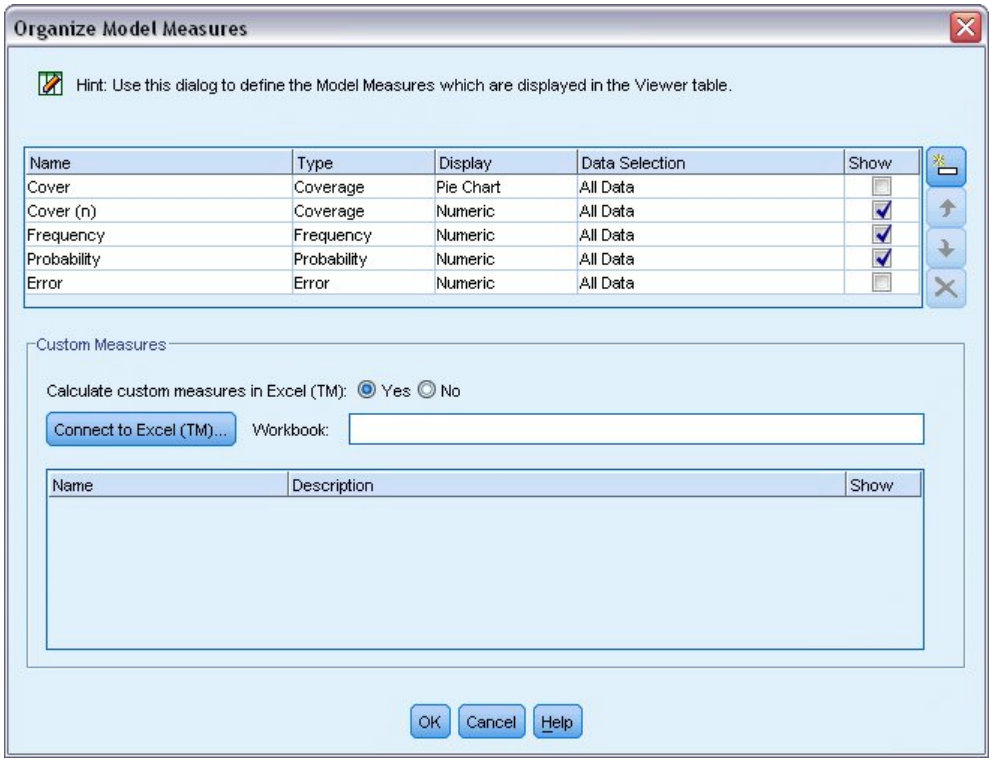


Рисунок 135. Диалоговое окно Организовать показатели модели

Дополнительно, если у вас установлен Microsoft Excel, можно определить связь с шаблоном Excel, в котором будут вычисляться пользовательские показатели, и добавить их в интерактивный вывод.

2. В диалоговом окне Организовать показатели модели задайте для опции **Вычислять пользовательские показатели в Excel (TM)** значение **Да**.
3. Щелкните по **Соединиться с Excel (TM)**
4. Выберите рабочую книгу *template_profit.xlt* из подпапки *streams* папки *Demos* вашего каталога установки IBM SPSS Modeler и щелкните по **Открыть**, чтобы запустить электронную таблицу.

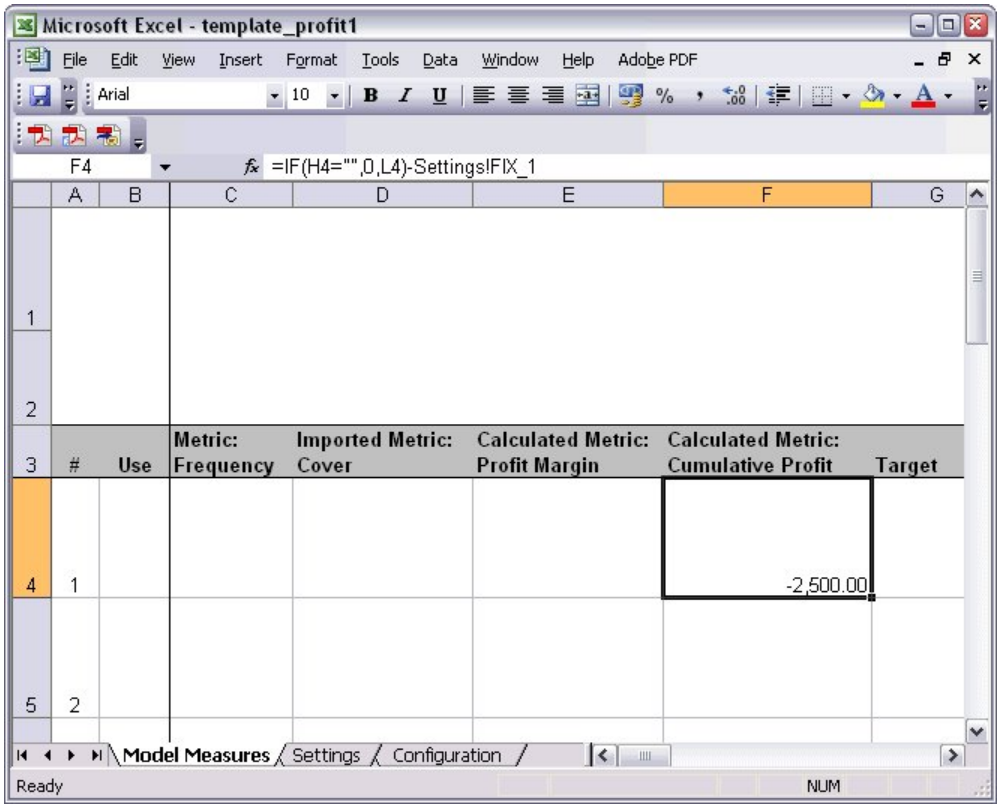


Рисунок 136. Рабочий лист Excel Показатели модели

Шаблон Excel содержит три рабочих листа:

- На рабочем листе **Показатели модели** выводятся показатели модели, импортированные из самой модели, и вычисляются пользовательские показатели для экспорта обратно в модель.
- Рабочий лист **Параметры** содержит параметры, которые будут использоваться в вычислении пользовательских показателей.
- На рабочем листе **Конфигурация** определяются показатели, которые будут импортироваться из модели и экспортироваться в нее.

Обратно в модель экспортируются следующие показатели:

- **Маржа прибыли.** Чистый доход от сегмента
- **Кумулятивная прибыль.** Общая прибыль от кампании

Определяется по следующим формулам:

```
Маржа прибыли = Частота * Доход от респондента - Покрытие * Переменные затраты

Кумулятивная прибыль = Полная маржа прибыли - Фиксированные затраты
```

Обратите внимание на то, что значения Частота и Покрытие импортируются из модели.

Параметры затрат и дохода задаются пользователем на рабочей странице Параметры.

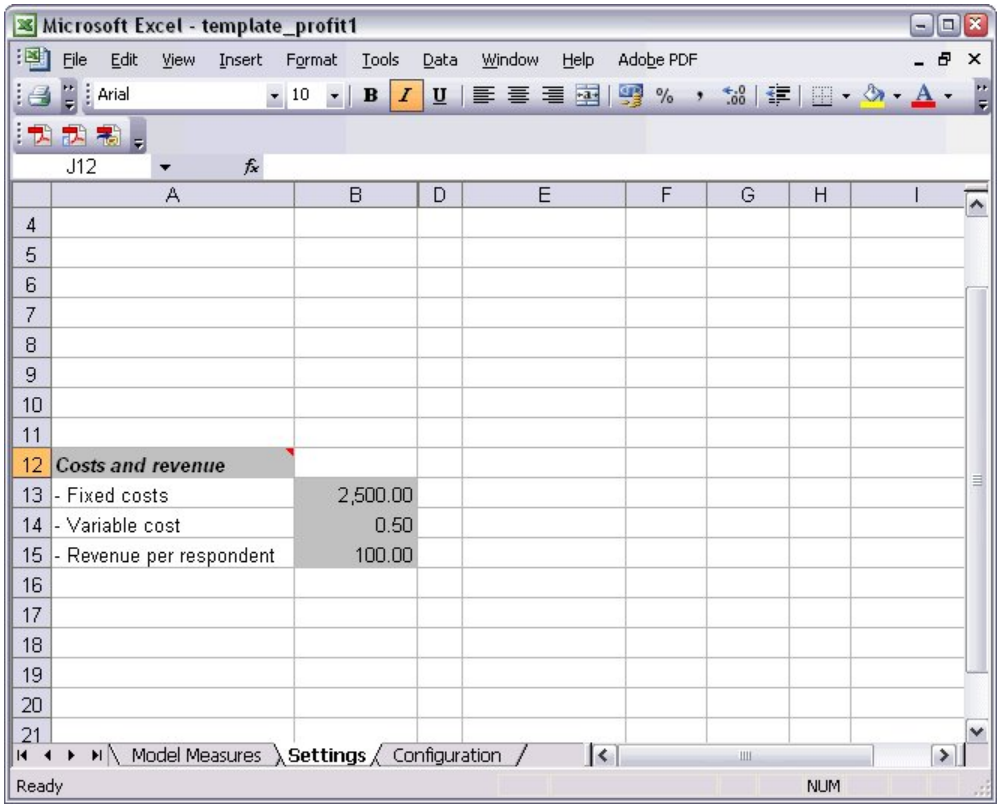


Рисунок 137. Рабочий лист Excel Параметры

Фиксированные затраты - это затраты на организацию кампании, например, на проектирование и планирование.

Переменные затраты - это затраты на расширение предложения для каждого покупателя, например, на конверты и марки.

Доход на респондента - это чистый доход от покупателя, откликнувшегося на предложение.

5. Для завершения обратной связи с моделью используйте панель задач Windows (или нажмите клавиши Alt+Tab), чтобы вернуться назад в средство просмотра интерактивного списка.

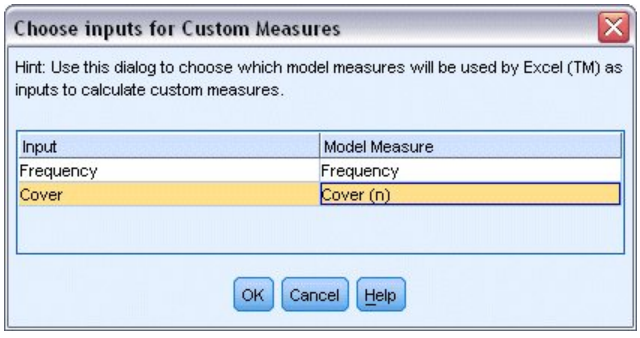


Рисунок 138. Выбор входных данных для пользовательских показателей

Выводится диалоговое окно Выбор входных данных для пользовательских показателей, где вы можете отобразить входные данные из модели на отдельные параметры, определенные в шаблоне. В левом столбце перечислены доступные показатели, а в правом столбце они отображаются на параметры электронной таблицы, определенные на рабочем листе Конфигурация.

6. В столбце **Показатели модели** выберите параметры **Частота** и **Покрытие (n)** для соответствующих входных полей и нажмите кнопку **OK**.

   В этом случае имена параметров в шаблоне Частота и Покрытие (n) окажутся совпадающими с входными полями, но могут использоваться и другие имена.

7. Нажмите кнопку **OK** в диалоговом окне Организовать показатели модели, чтобы обновить средство просмотра интерактивного списка.

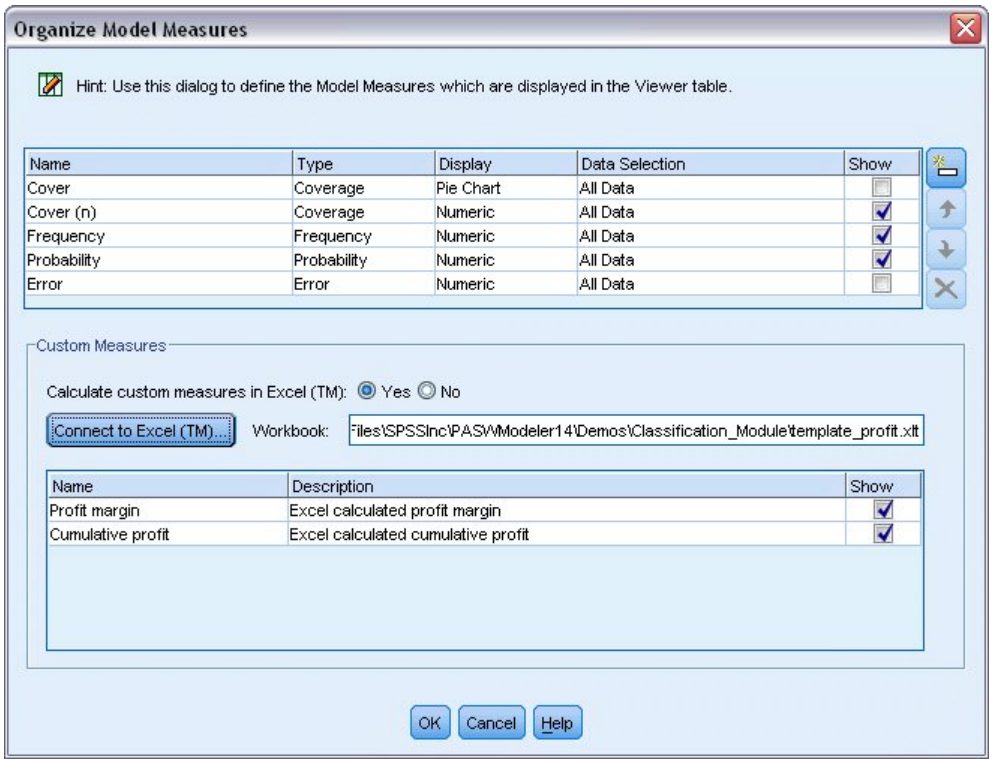


Рисунок 139. Диалоговое окно Организовать показатели модели, показывающее пользовательские показатели из Excel

Теперь новые показатели будут добавлены как новые столбцы в окне и будут пересчитываться при всяком изменении модели.

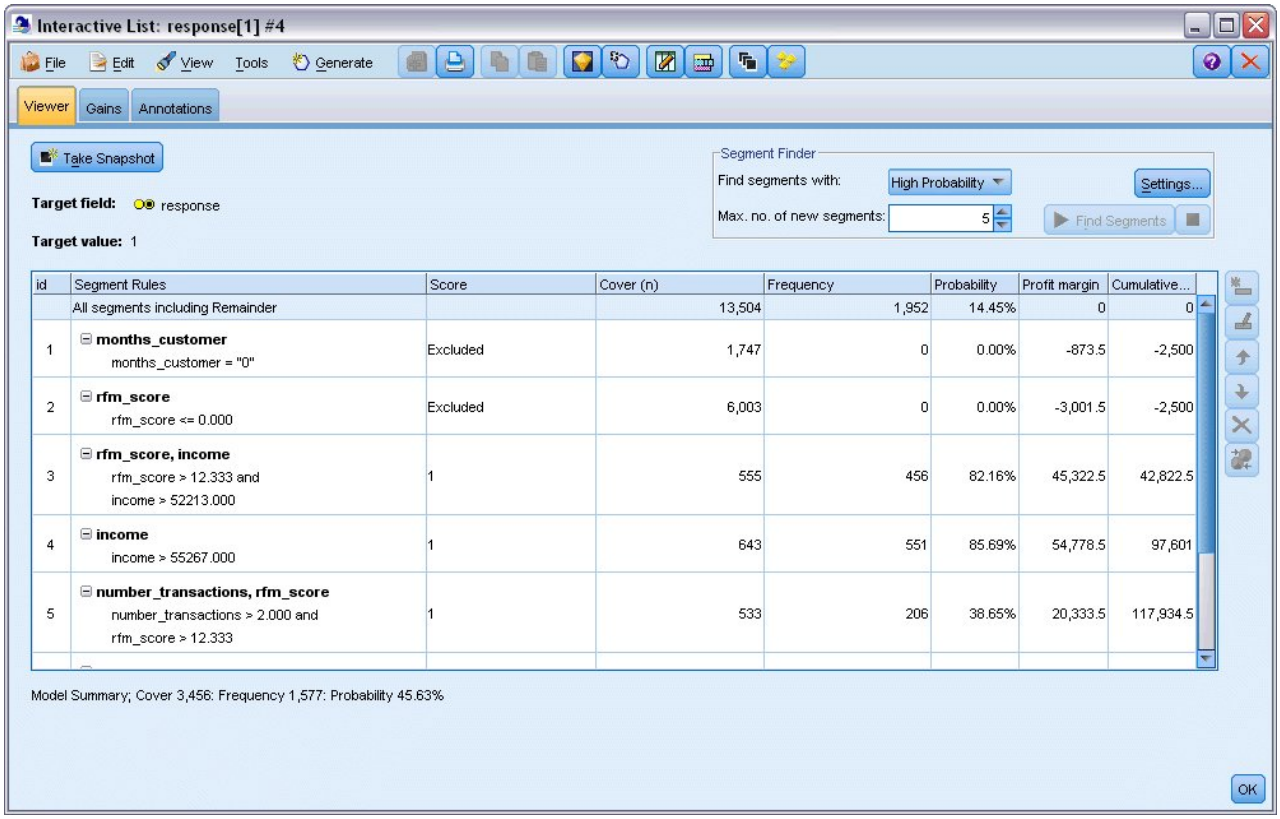


Рисунок 140. Пользовательские показатели из Excel, выведенные в средстве просмотра интерактивного списка

Изменяя шаблон Excel, можно создать любое количество пользовательских показателей.

Изменение шаблона Excel

Хотя IBM SPSS Modeler поставляется с шаблоном Excel по умолчанию для использования со средством просмотра интерактивного списка, вам может потребоваться изменить некоторые параметры или добавить свои собственные. Например, затраты, заданные в этом шаблоне, могут оказаться неверными для вашей организации, и их нужно будет исправить.

Примечание: Если вы изменяете существующий шаблон или создаете свой собственный, не забудьте сохранить его с суффиксом Excel 2003 *.xlt*.

Чтобы заменить в шаблоне подробности затрат и доходов на новые и обновить средство просмотра интерактивного списка новыми данными:

1. В средстве просмотра интерактивного списка выберите в меню Инструменты опцию **Организовать показатели модели**.
2. В диалоговом окне Организовать показатели модели щелкните по **Соединиться с Excel™.**
3. Выберите рабочую книгу *template_profit.xlt* и щелкните по **Открыть**, чтобы запустить электронную таблицу.
4. Выберите рабочий лист Параметры.
5. Измените **Фиксированные затраты** на 3250,00, а **Доход на респондента** на 150,00.

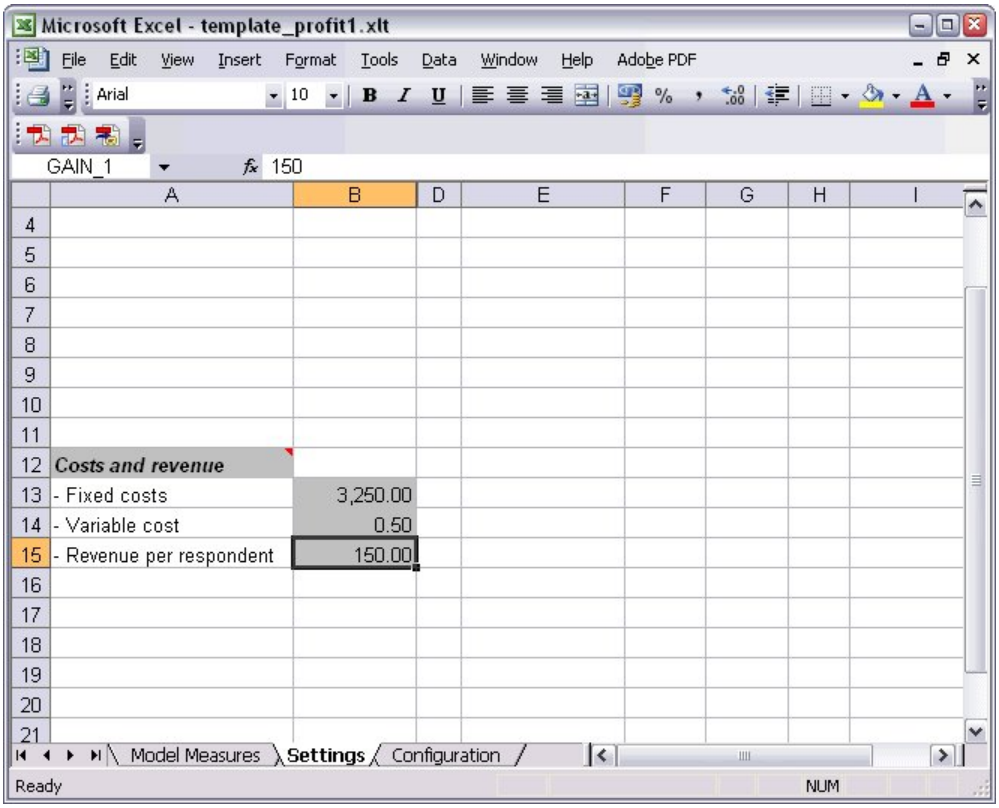


Рисунок 141. Измененные значения на рабочем листе Параметры Excel

6. Сохраните измененный шаблон под уникальным осмысленным именем. Убедитесь, что у него расширение Excel 2003 *.xlt*.

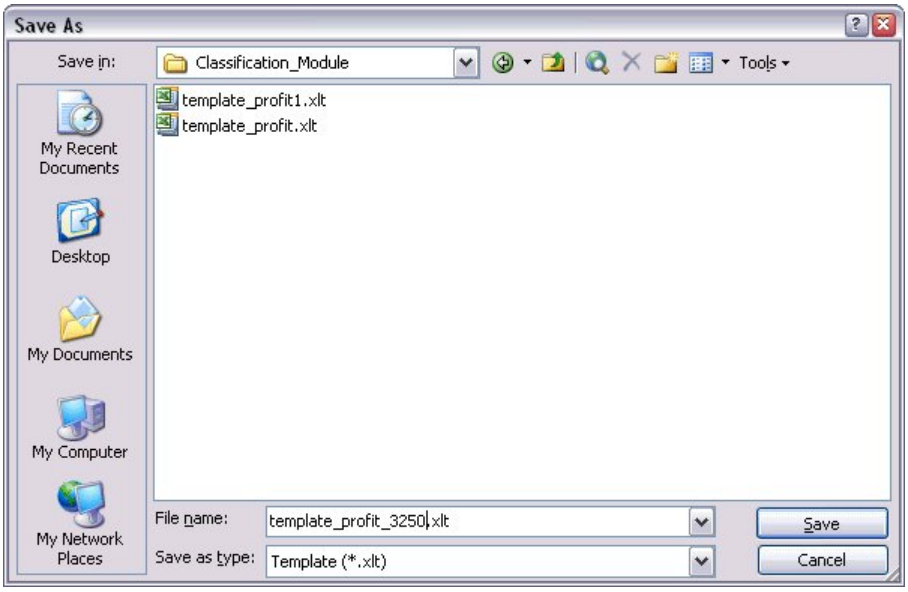


Рисунок 142. Сохранение измененного шаблона Excel

7. Используйте панель задач Windows (или нажмите клавиши Alt+Tab), чтобы снова перейти в средство просмотра интерактивного списка.

В диалоговом окне Выбрать входные данные для пользовательских показателей выберите показатели, которые вы хотите выводить, и нажмите кнопку **OK**.

8. В диалоговом окне Организовать показатели модели нажмите кнопку **OK**, чтобы изменить средство просмотра интерактивного списка.

Конечно, в этом примере показан только один способ изменения шаблона Excel; можно внести другие изменения, которые будут извлекать данные из средства просмотра интерактивного списка или передавать в него данные, или работать в Excel для создания других видов вывода, например, диаграмм.

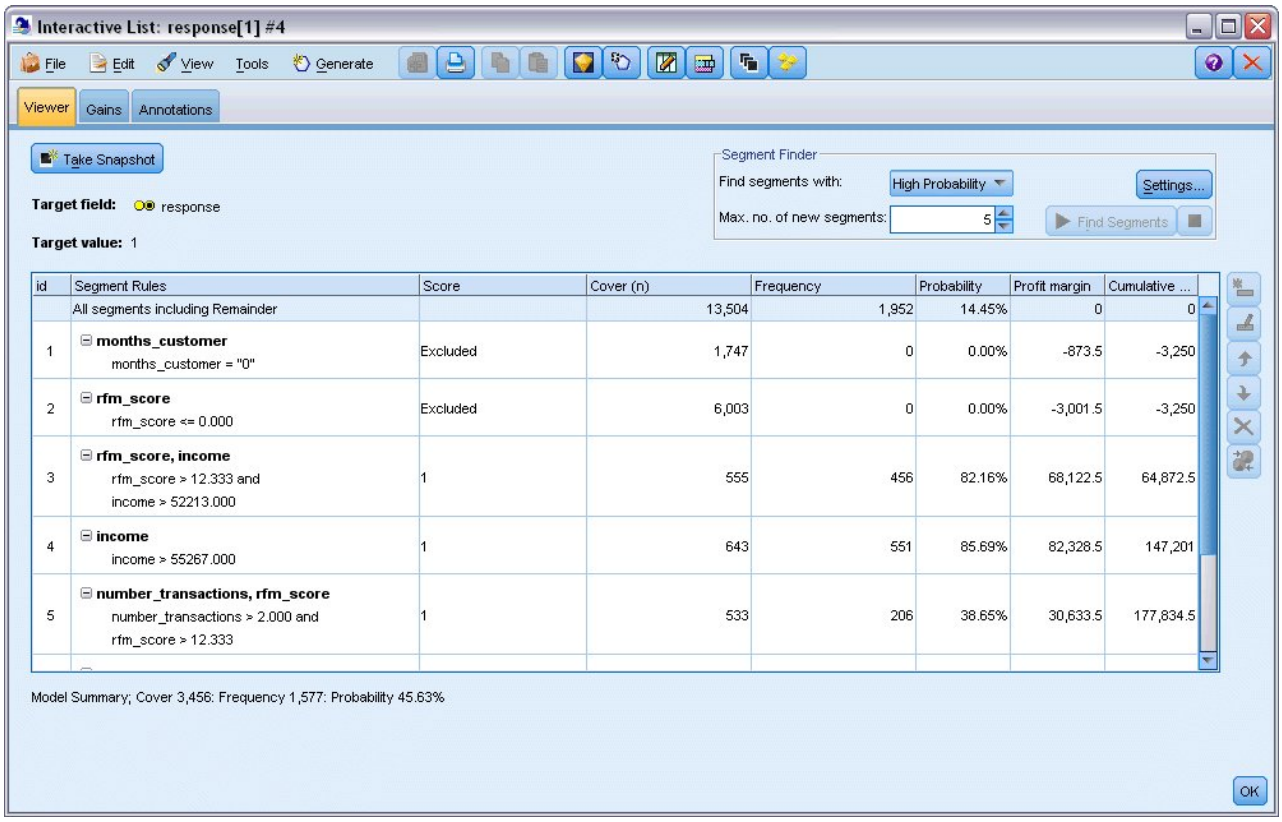


Рисунок 143. Измененные пользовательские показатели из Excel, выведенные в средстве просмотра интерактивного списка

Сохранение результатов

Чтобы сохранить модель для дальнейшего использования в вашем интерактивном сеансе, можно сделать снимок модели, который будет перечислен на вкладке Снимки. К каждому из сохраненных снимков можно вернуться в любое время интерактивного сеанса.

Продолжая таким образом можно экспериментировать с дополнительными задачами исследования данных, чтобы искать дополнительные сегменты. Можно изменять также существующие сегменты, вставлять пользовательские сегменты на основе ваших собственных бизнес-правил, создавать варианты выбора данных, чтобы оптимизировать модель для отдельных групп, и настраивать модель многими другими способами. В конце концов, вы можете явно включить или исключить каждый сегмент в качестве подходящего, чтобы указать, как каждый из них будет оцениваться.

Если достигнуты удовлетворительные результаты, можно использовать меню Генерировать, чтобы сгенерировать модель, которую можно будет добавить в поток или внедрить в целях скоринга.

Вместо этого можно сохранить текущее состояние вашего интерактивного сеанса на следующий день, выбрав опцию **Изменить узел моделирования** в меню Файл. При этом узел моделирования Список решений будет обновлен текущими параметрами, в том числе задачами исследования данных, снимками моделей, вариантами выбора данных и пользовательскими показателями. При следующем запуске потока просто убедитесь, что в узле моделирования Список решений выбрана опция **Использовать сохраненную информацию о сеансе**, чтобы восстановить сеанс до его текущего состояния.

Глава 12. Классификация клиентов в сфере телекоммуникаций (полиномиальная логистическая регрессия)

Логистическая регрессия - это статистический метод для классификации записей на основании значений входных полей. Она аналогична линейной регрессии, но логистическая регрессия использует категориальные поля назначения вместо числовых.

Например, предположим, что провайдер связи сегментировал базу своих клиентов по шаблонам использования сервисов, категоризуя клиентов в четыре группы. Если демографические данные можно использовать для прогноза состава группы, то можно настроить предложения по отдельным возможным заказчикам.

Этот пример использует поток *telco_custcat.str*, в котором используется файл данных *telco.sav*. Эти файлы находятся в каталоге *Demos* любой установки IBM SPSS Modeler. Доступ к ним можно получить из группы программ IBM SPSS Modeler в меню Пуск Windows. Файл *telco_custcat.str* находится в каталоге *streams*.

Этот пример фокусируется на использовании демографических данных для предсказания паттернов использования. У поля назначения *категория клиента* есть четыре возможных значения, соответствующих четырем группам клиентов:

Значение	Метка
1	Базовое обслуживание
2	Интернет-обслуживание
3	Дополнительное обслуживание
4	Полное обслуживание

Так как у поля назначения есть несколько категорий, используется полиномиальная модель. В том случае, когда у поля назначения есть две отдельные категории, такие как да/нет, true/false или есть отток клиентов/нет оттока клиентов, вместо этого можно создавать биномиальную модель. Дополнительную информацию смотрите в разделе Глава 13, “Отток клиентов в сфере телекоммуникаций (Биномиальная логистическая регрессия)”, на стр. 141.

Построение потока

1. Добавьте узел источников файла статистики, указывающий на файл *telco.sav* в папке *Demos*.

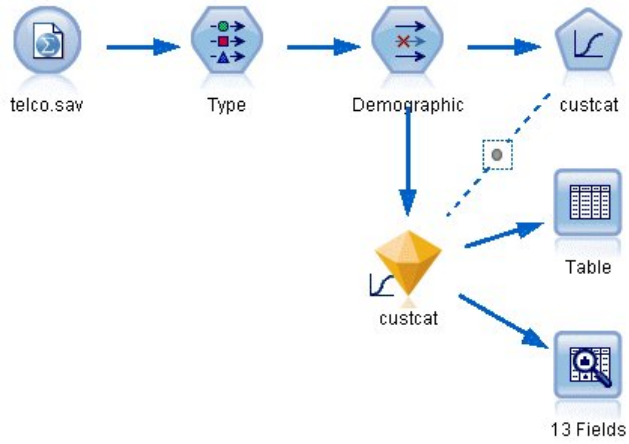


Рисунок 144. Поток примера для классификации покупателей с использованием полиномиальной логистической регрессии

a. Добавьте узел Тип и щелкните по **Прочитать значения**, чтобы убедиться, что все типы измерений заданы правильно. Например, большинство полей со значениями 0 и 1 можно считать флагами.

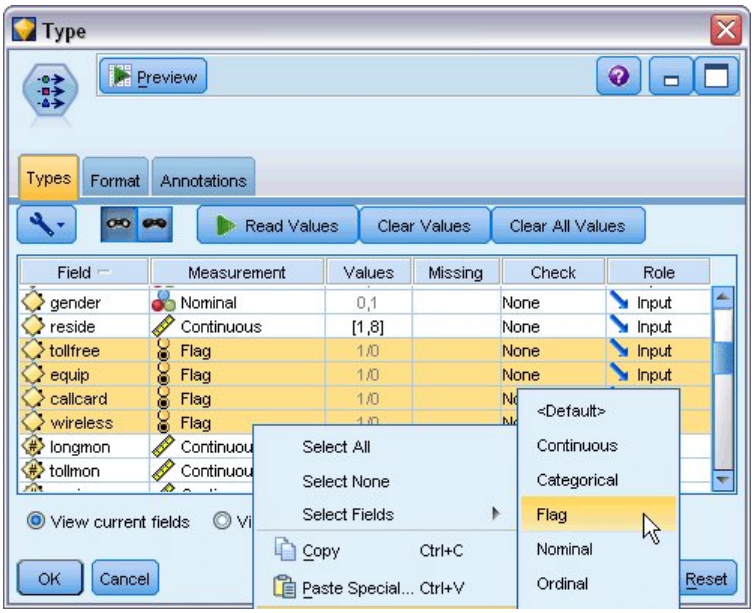


Рисунок 145. Задание уровня измерения для нескольких полей

Совет: чтобы изменить свойства нескольких полей со сходными значениями (например, 0/1), щелкните по заголовку столбца *Значения*, и когда поля будут отсортированы по значению, удерживайте нажатой клавишу Shift и выделите мышью или клавишами со стрелками все поля, которые нужно изменить. Затем можно щелкнуть правой кнопкой мыши по выбранному, чтобы изменить уровень измерения или другие атрибуты выбранных полей.

Заметим, что поле *gender* (пол) правильнее рассматривать не как флаг, а как поле с набором из двух значений, поэтому оставьте Тип измерения для этого поля **Номинальный**.

b. Для поля *custcat* (категория клиента) задайте роль **Назначение**. Для всех остальных полей нужно задать роль **Ввод**.

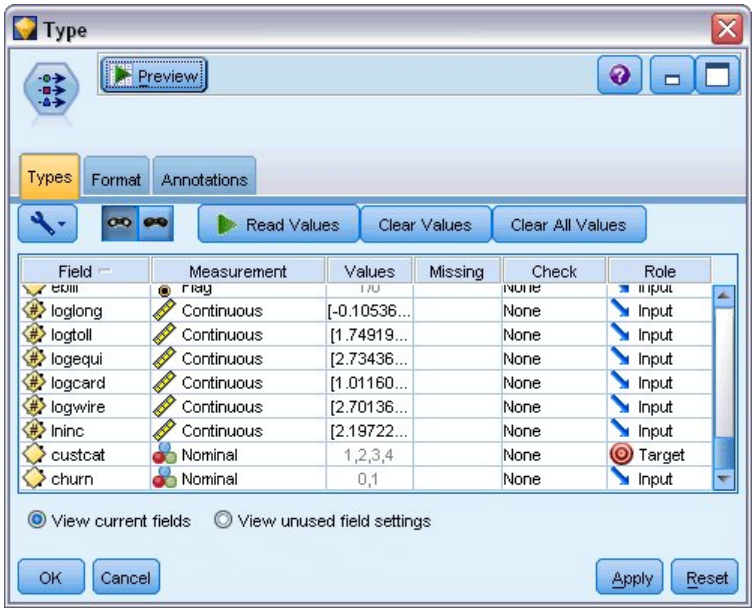


Рисунок 146. Задание роли поля

Поскольку предмет этого примера - демографические показатели, используйте узел Фильтр, чтобы оставить только нужные поля (*region*, *age*, *marital*, *address*, *income*, *ed*, *employ*, *retire*, *gender*, *reside* и *custcat* - регион, возраст, семейное положение, адрес, доход, образование, занятость, на пенсии, пол, место жительства и категория клиента). Остальные поля в ходе данного анализа можно исключить.

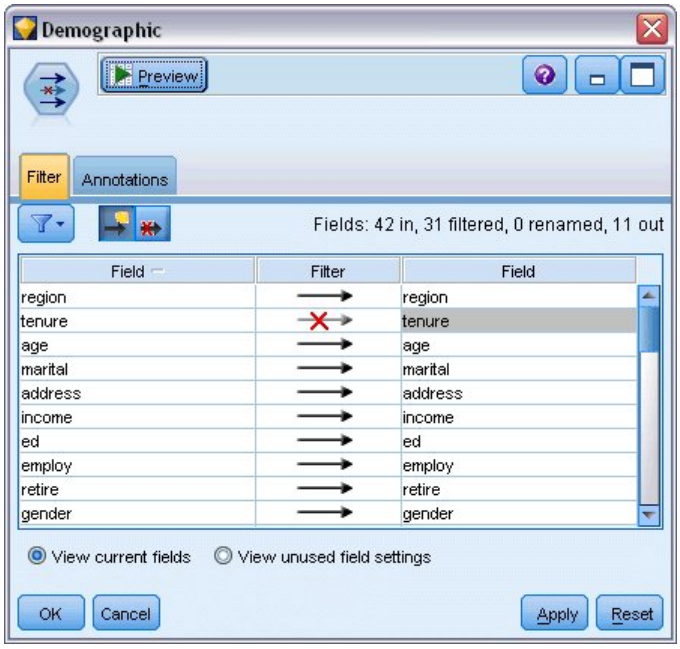


Рисунок 147. Фильтрация демографических полей

(Другой вариант - не исключать эти поля, а задать для них роль **Нет**, или выбрать нужные поля в узле моделирования.)

2. В Логистическом узле щелкните по вкладке **Модель** и выберите **Пошаговый** метод. Выберите опции **Полиноминальная**, **Главные эффекты**, а также **Включить константу в уравнение**.

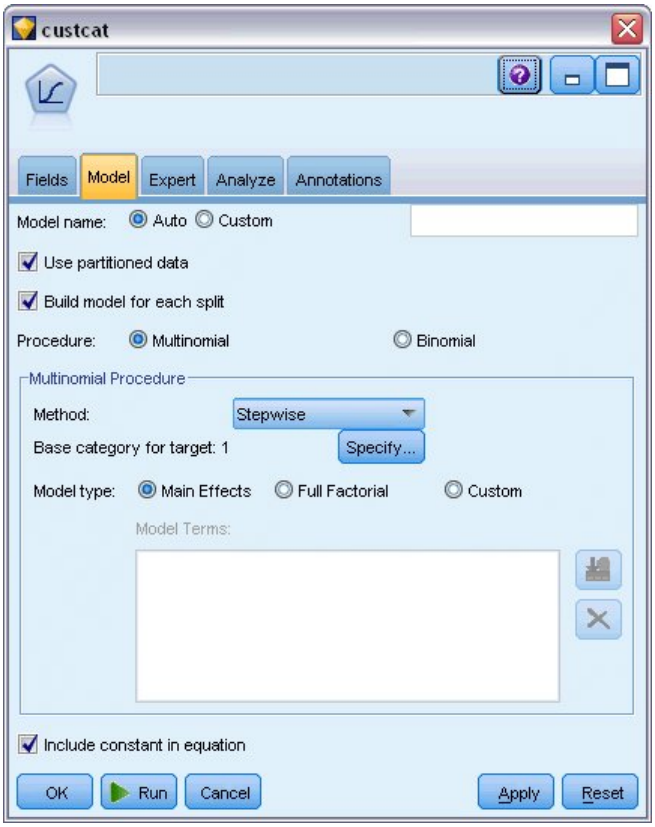


Рисунок 148. Выбор опций модели

Оставьте базовой категорией для поля назначения 1. Модель сравнит других клиентов с теми, кто подписан на Базовое обслуживание.

3. На вкладке Эксперт выберите режим **Эксперт** и опцию **Выходные данные**, а в диалоговом окне Расширенные выходные данные выберите **Таблица классификации**.

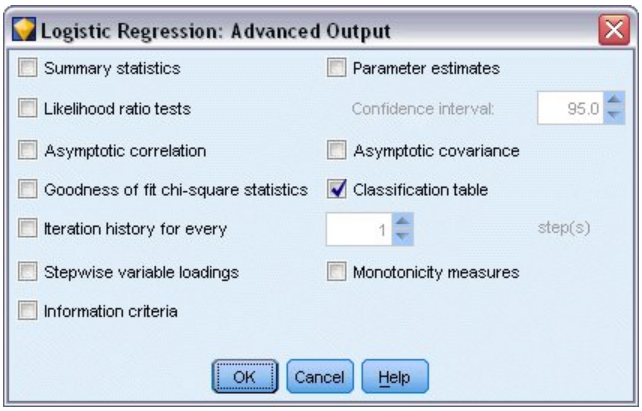


Рисунок 149. Выбор параметров вывода

Просмотр модели

1. Выполните узел, чтобы сгенерировать модель, которая будет добавлена на палитру Модели в правом верхнем углу. Для просмотра ее подробностей щелкните правой кнопкой мыши по сгенерированной модели и выберите опцию **Обзор**.

На вкладке модели будут показаны уравнения, использованные для назначения записей каждой из категорий поля назначения. Есть четыре возможные категории, одна из которых базовая, и для нее подробности уравнения не показываются. Подробности показываются для трех остальных категорий, где категория 3 представляет Дополнительное обслуживание и так далее.

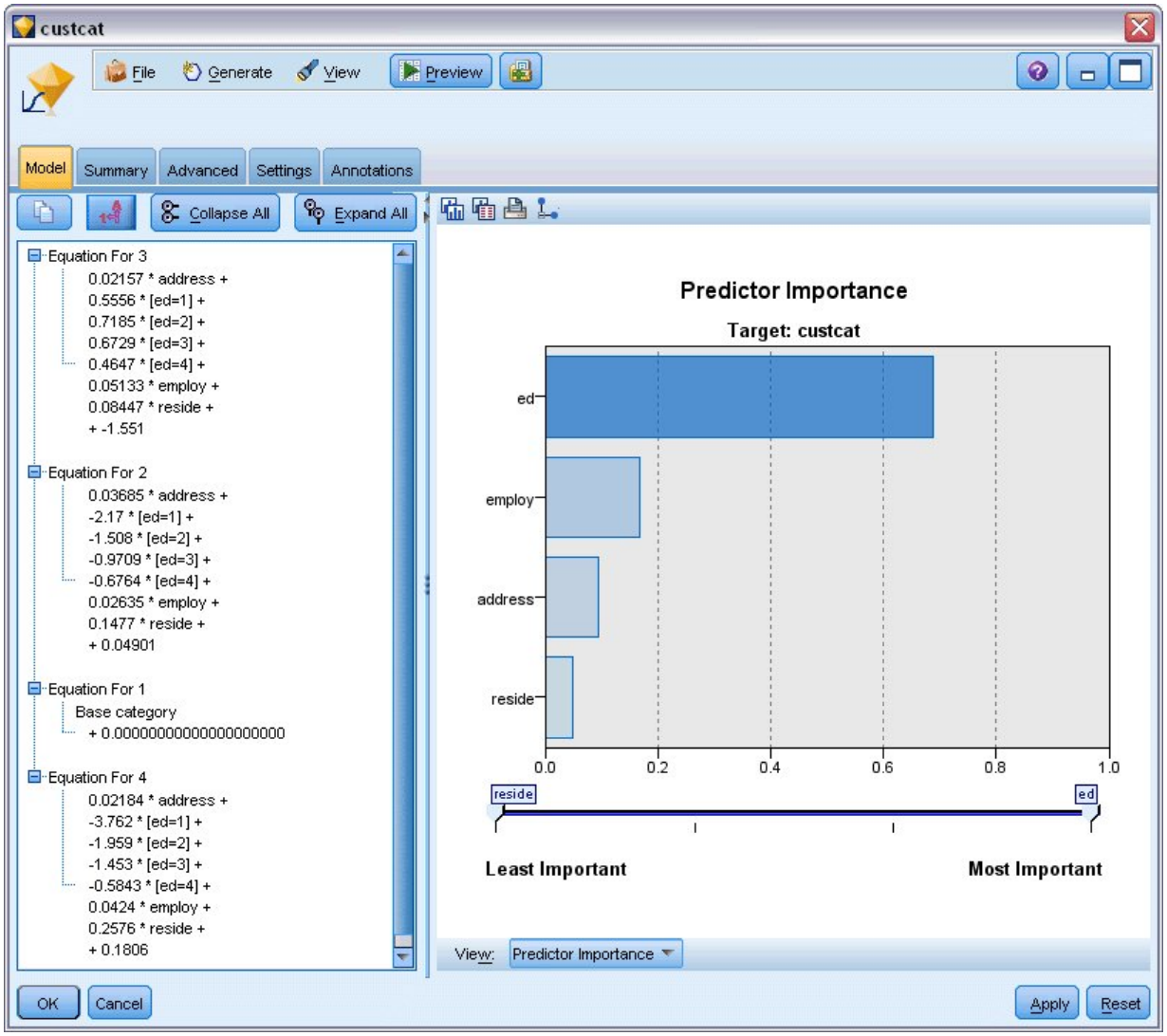


Рисунок 150. Просмотр результатов модели

На вкладке Сводка среди прочего будет показано поле назначения и входные поля (предикторы), используемые моделью. Обратите внимание на то, что эти поля были выбраны на основе пошагового метода, а не представляют собой полный список полей, переданный для рассмотрения.

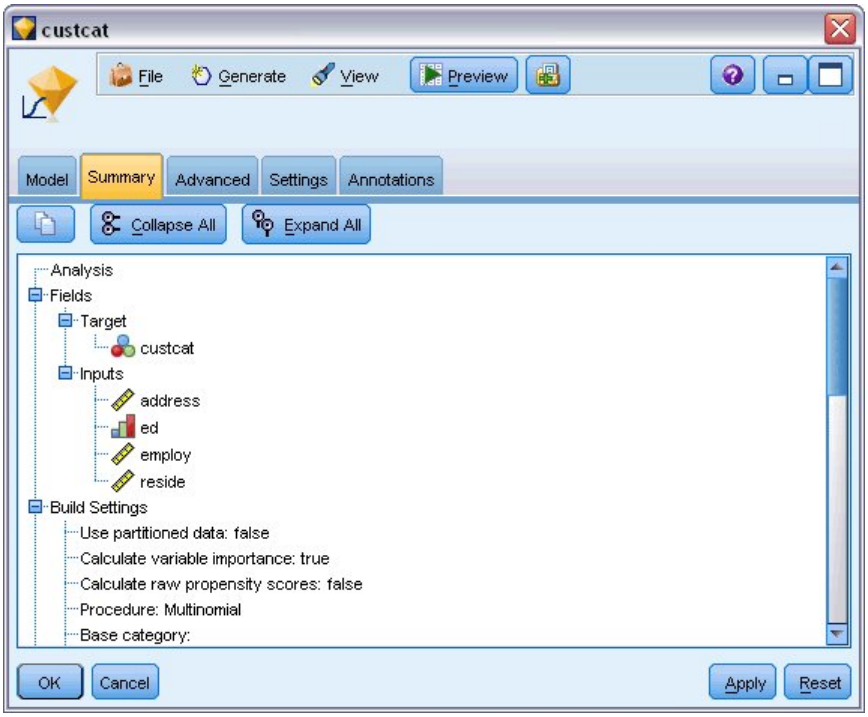


Рисунок 151. Сводка модели, иллюстрирующая целевые и входные поля

Показанные на вкладке Дополнительно элементы зависят от опций, выбранных в диалоговом окне узла моделирования Расширенный вывод.

Один всегда показываемый элемент - это Сводка обработки наблюдений, где выводится процентная доля записей, попадающих в каждую категорию поля назначения. Это дает вам пустую модель, которая будет использоваться как база для сравнения.

Без построения модели, использующей предикторы, наилучшим предположением будет назначение всех клиентов самой общей группе, то есть группе для Дополнительных услуг.

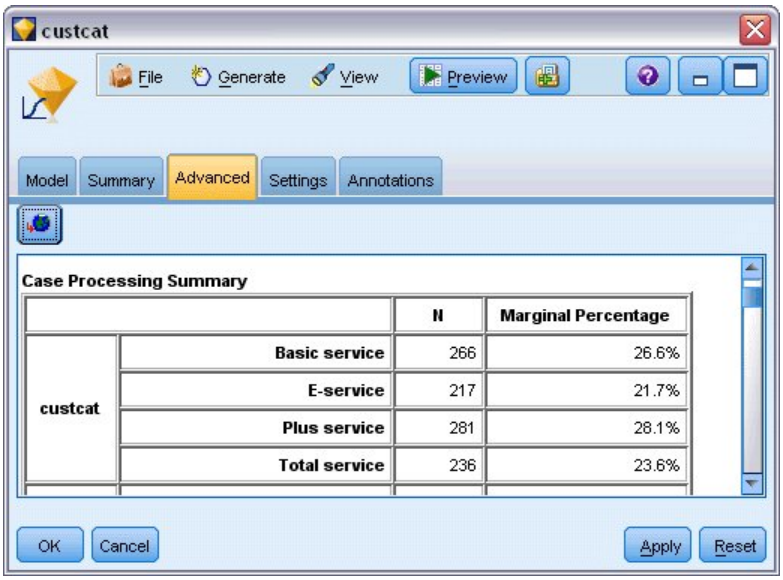


Рисунок 152. Сводный отчет обработки наблюдений

На основании данных обучения, если все клиенты назначены пустой модели, вы будете правы в 281/1000 = 28,1% случаев. Вкладка Дополнительно содержит еще информацию, которая позволяет проверить предсказания модели. Поэтому можно сравнить предсказания с результатами пустой модели, чтобы увидеть, насколько хорошо эта модель работает с вашими данными.

Внизу вкладки Дополнительно таблица Классификация показывает результаты вашей модели, которые правильны в 39,9% случаев.

В частности, ваша модель отлично работает при идентификации клиентов Полного обслуживания (категория 4), но дает очень слабые результаты при идентификации клиентов Интернет-обслуживание (категория 2). Если требуется большая точность для клиентов категории 2, для их идентификации требуется найти другой предиктор.

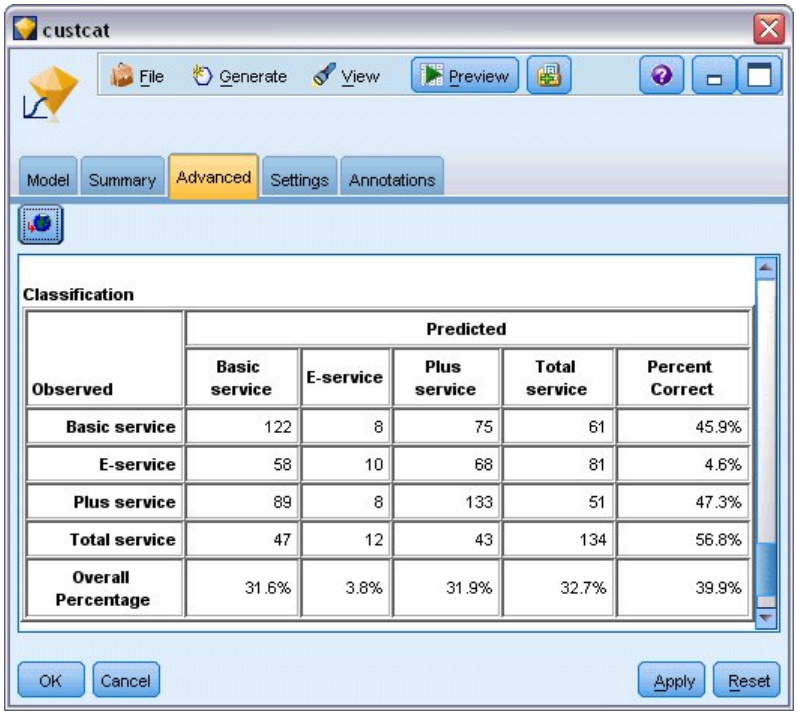


Рисунок 153. Таблица классификации

В зависимости от того, что вы хотите предсказать, модель может идеально подходить для ваших потребностей. Например, если идентификация клиентов в категории 2 не существенна, модель может быть достаточно точной для вас. Это может случиться в такой ситуации, когда при обслуживании Интернет-обслуживание продажи идут с большими скидками и приносят малую прибыль.

Например, если максимальная прибыль на инвестиции приходит от клиентов, попадающих в категорию 3 или 4, эта модель может дать вам всю необходимую информацию.

Чтобы оценить, насколько хорошо модель на самом деле подгоняет данные, при построении модели в диалоговом окне Расширенный вывод доступно несколько диагностических средств. Объяснение математических основ методов моделирования, используемых в IBM SPSS Modeler, смотрите в публикации *IBM SPSS Modeler: Руководство по алгоритмам* в каталоге *\Documentation* на установочном диске.

Учтите, что эти результаты основаны только на обучающих данных. Чтобы оценить, насколько хорошо модель обобщается на другие данные реального мира, можно применить узел Разбиение, который будет содержать поднабор записей в целях проверки.

Глава 13. Отток клиентов в сфере телекоммуникаций (Биномиальная логистическая регрессия)

Логистическая регрессия - это статистический метод для классификации записей на основании значений входных полей. Она аналогична линейной регрессии, но логистическая регрессия использует категориальные поля назначения вместо числовых.

Этот пример использует поток *telco_churn.str*, в котором используется файл данных *telco.sav*. Эти файлы находятся в каталоге *Demos* любой установки IBM SPSS Modeler. Доступ к ним можно получить из группы программ IBM SPSS Modeler в меню Пуск Windows. Файл *telco_churn.str* находится в каталоге *streams*.

Например, предположим, что провайдер телекоммуникационных услуг озабочен количеством клиентов, теряемых из-за конкурирующих компаний. Если данные об использовании услуг можно использовать для предсказания, какие клиенты склонны перейти к другому поставщику услуг, можно настроить специальные предложения, чтобы сохранить как можно больше клиентов.

Предмет этого примера - применение существующих данных об использовании услуг для предсказания оттока клиентов. Так как у поля назначения есть две отдельные категории, используется биномиальная модель. В том случае, если у поля назначения есть несколько категорий, вместо этого можно создавать полиномиальную модель. Дополнительную информацию смотрите в разделе Глава 12, “Классификация клиентов в сфере телекоммуникаций (полиномиальная логистическая регрессия)”, на стр. 133.

Построение потока

1. Добавьте узел источников файла статистики, указывающий на файл *telco.sav* в папке *Demos*.

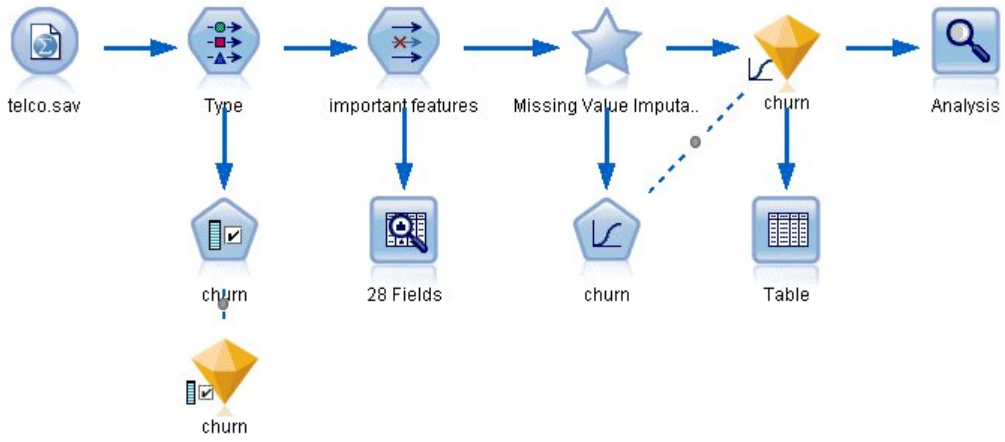


Рисунок 154. Поток примера для классификации покупателей с использованием биномиальной логистической регрессии

2. Добавьте узел Тип для определения полей, убедившись, что все уровни измерений заданы правильно. Например, большинство полей со значениями 0 и 1 можно рассматривать как флаги, но некоторые поля, такие как пол, точнее считать номинальными полями с двумя значениями.

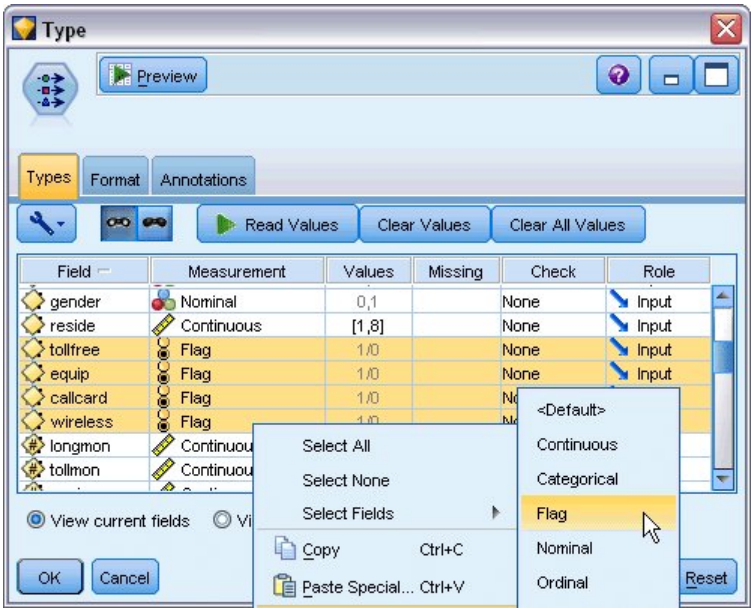


Рисунок 155. Задание уровня измерения для нескольких полей

Подсказка: Чтобы изменить свойства для нескольких полей с аналогичными значениями (такими как 0/1), щелкните по заголовку столбца *Значения*, чтобы отсортировать поля по значениям, а затем удерживайте нажатой кнопку Shift, используя мышь или клавишу со стрелкой для выбора всех полей, которые вы хотите изменить. Затем можно щелкнуть правой кнопкой мыши по выбранному, чтобы изменить уровень измерения или другие атрибуты выбранных полей.

3. Задайте для поля *churn* (отток клиентов) уровень измерения **Флаг** и роль **Назначение**. Для всех остальных полей нужно задать роль **Ввод**.

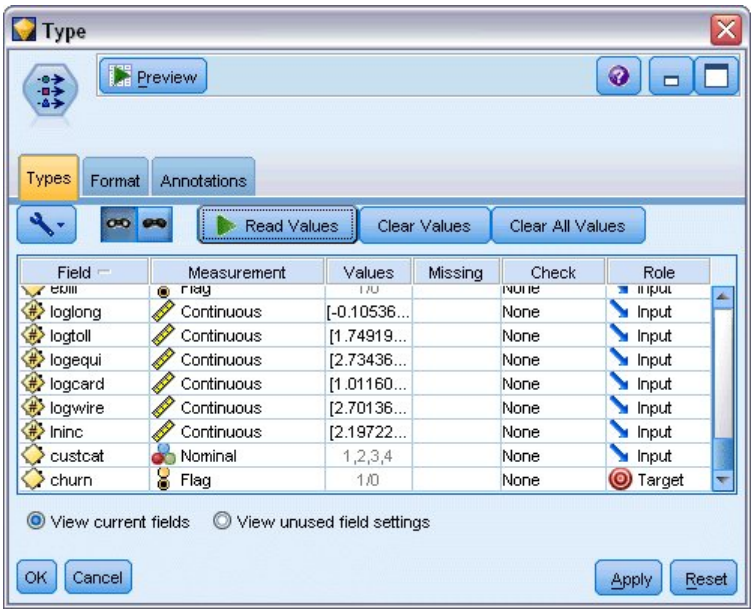


Рисунок 156. Задание уровня измерения и роли для поля оттока клиентов

4. Добавьте узел моделирования выбора возможностей к узлу Тип.

   Использование узла Выбор возможностей позволяет удалить предикторы или данные, которые не добавляют полезной информации с точки зрения взаимосвязей предиктор/назначение.

5. Выполните поток.
6. Откройте получившийся слепок модели и в меню **Генерировать** выберите опцию **Фильтр**, чтобы создать узел Фильтр.

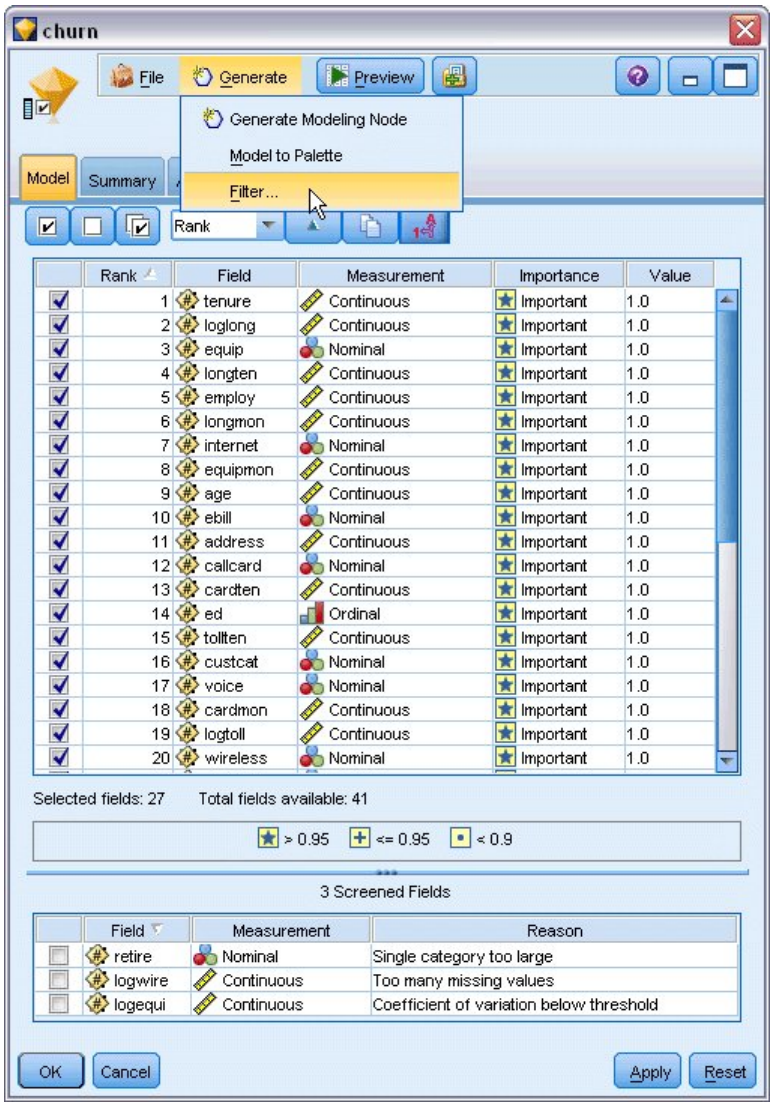


Рисунок 157. Генерирование узла Фильтр из узла Выбор возможностей

Не все данные в файле *telco.sav* будут полезны для предсказания оттока клиентов. Можно использовать фильтр для выбора только тех рассматриваемых данных, которые будут полезны при использовании в качестве предикторов.

7. В диалоговом окне Сгенерировать фильтр выберите опцию **Все отмеченные поля: важно** и нажмите кнопку **OK**.
8. Присоедините сгенерированный узел Фильтр к узлу Тип.

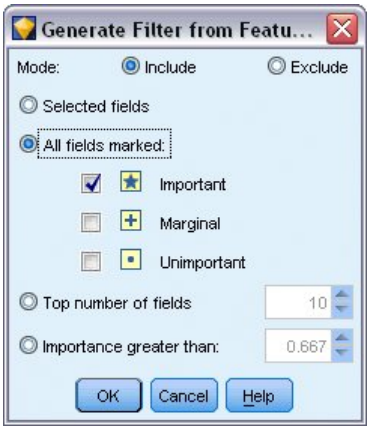
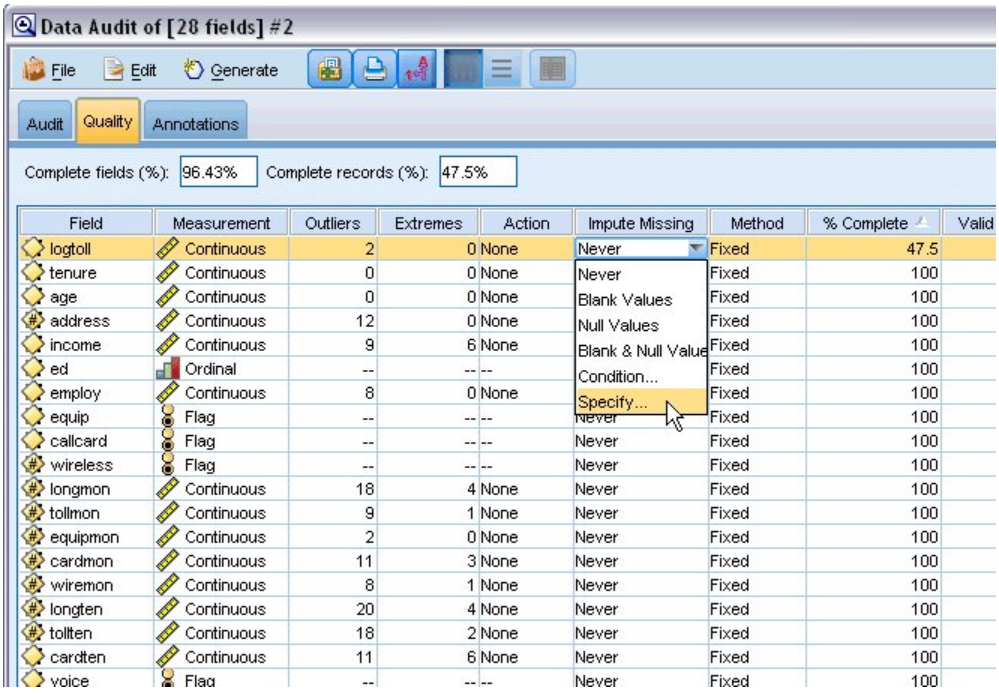


Рисунок 158. Выбор важных полей

9. Присоедините узел Аудит данных к сгенерированному узлу Фильтр.

   Откройте узел Аудит данных и нажмите кнопку **Выполнить**.

10. На вкладке Качество браузера Аудит данных щелкните по столбцу *% заполнения* для сортировки столбца в порядке возрастания чисел в нем. Это позволяет идентифицировать все поля с большим количеством отсутствующих данных; в этом случае единственное поле, которое нужно исправить, - это *logtoll*, процент заполнения в котором меньше 50.
11. В столбце *Импутировать отсутствующие* для поля *logtoll* щелкните по **Задать**.



Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid
logtoll	Continuous	2	0	None	Never	Fixed	47.5	
tenure	Continuous	0	0	None		Fixed	100	
age	Continuous	0	0	None		Fixed	100	
address	Continuous	12	0	None		Fixed	100	
income	Continuous	9	6	None		Fixed	100	
ed	Ordinal	--	--	--		Fixed	100	
employ	Continuous	8	0	None		Fixed	100	
equip	Flag	--	--	--		Fixed	100	
callcard	Flag	--	--	--	Never	Fixed	100	
wireless	Flag	--	--	--	Never	Fixed	100	
longmon	Continuous	18	4	None	Never	Fixed	100	
tollmon	Continuous	9	1	None	Never	Fixed	100	
equipmon	Continuous	2	0	None	Never	Fixed	100	
cardmon	Continuous	11	3	None	Never	Fixed	100	
wiremon	Continuous	8	1	None	Never	Fixed	100	
longten	Continuous	20	4	None	Never	Fixed	100	
tollten	Continuous	18	2	None	Never	Fixed	100	
cardten	Continuous	11	6	None	Never	Fixed	100	
voice	Flag	--	--	--	Never	Fixed	100	

Рисунок 159. Импутация отсутствующих значений для logtoll

12. Для опции **Когда импутировать** выберите значение **Пробелы и пустые значения**. Для опции **Исправить как** выберите значение **Среднее** и нажмите кнопку **OK**.

    Выбор опции **Среднее** гарантирует, что импутированные значения не будут неблагоприятно влиять на среднее всех значений в общих данных.

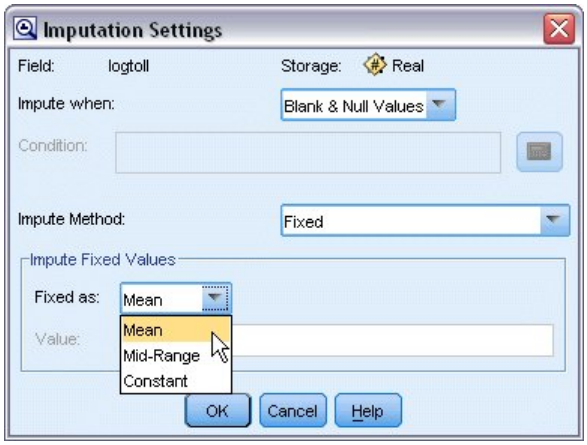


Рисунок 160. Выбор параметров импутации

13. На вкладке Качество браузера Аудит данных сгенерируйте надузел Пропущенные значения. Для этого выберите пункт меню:

    Создать > Надузел пропущенных значений

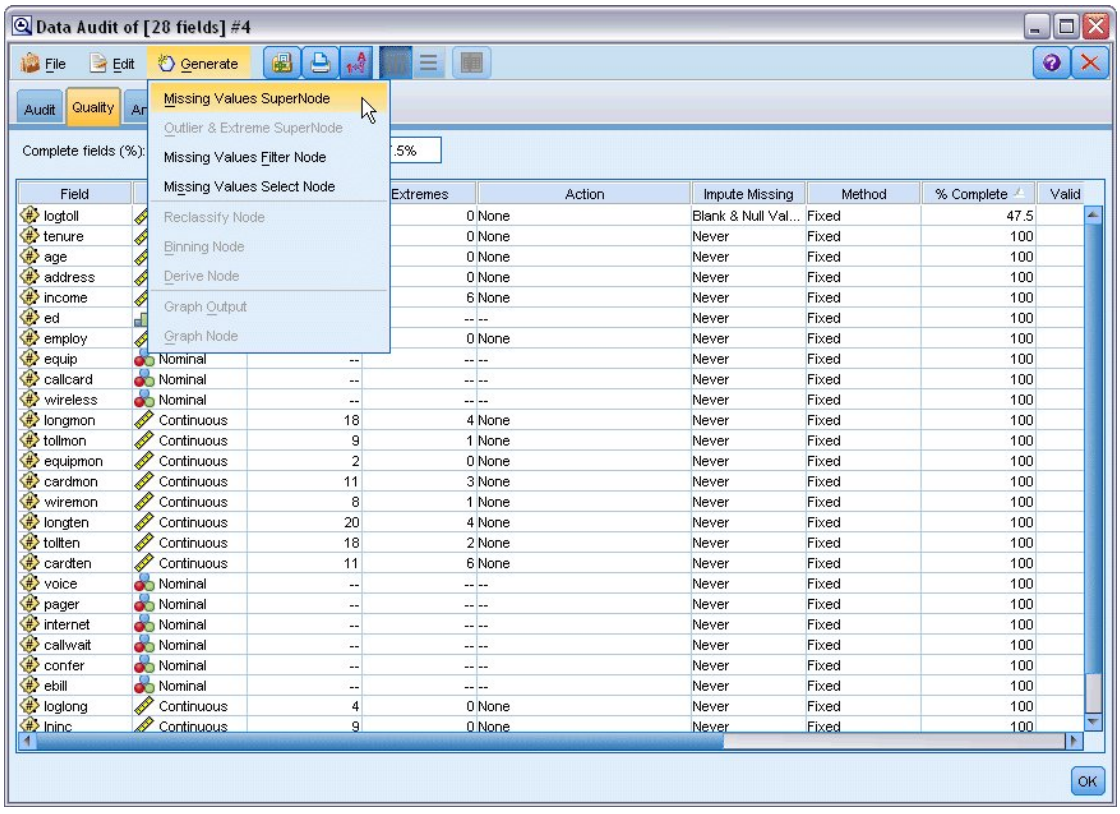


Рисунок 161. Надузел Генерирование отсутствующих значений

В диалоговом окне надузла пропущенных значений увеличьте **Размер выборки** до 50% и нажмите кнопку **OK**.

Этот надузел будет показан на холсте потока с заголовком: *Импутация пропущенных значений*.

14. Присоедините этот надузел к узлу Фильтр.

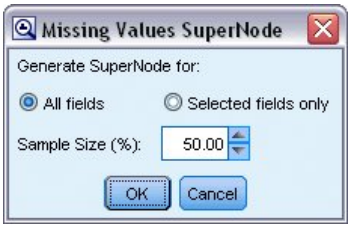


Рисунок 162. Задание размера выборки

15. Добавьте Логистический узел к надузлу.
16. В Логистическом узле щелкните по вкладке Модель и выберите процедуру **Биномиальная**. В области *Биномиальная процедура* выберите **Прямой** метод.

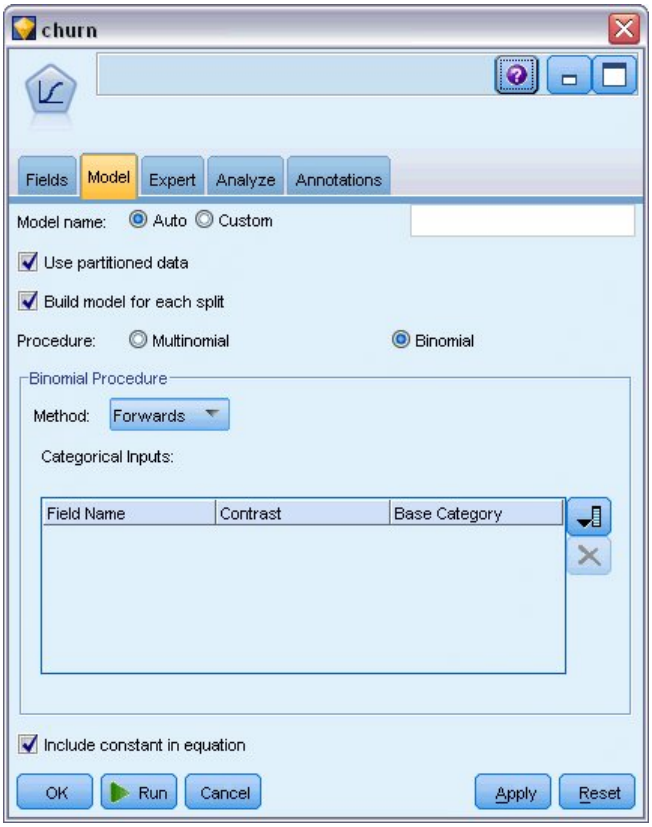


Рисунок 163. Выбор опций модели

17. На вкладке Эксперт выберите режим **Эксперт** и щелкните по **Выходные данные**. Появится диалоговое окно Расширенный вывод.
18. В диалоговом окне Расширенный вывод выберите в качестве типа *Вывода* значение **На каждом шаге**. Выберите опции **Хронология итераций** и **Оценки параметров** и нажмите кнопку **OK**.

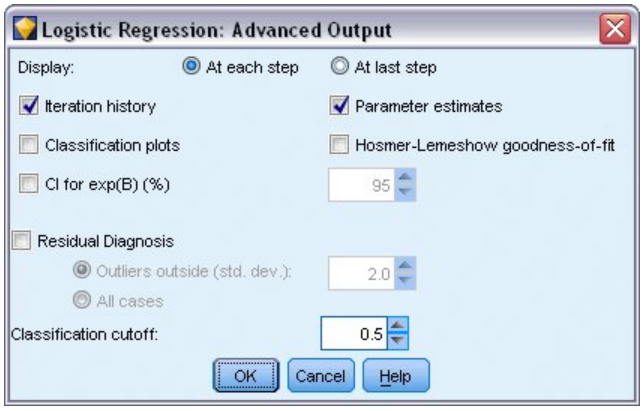


Рисунок 164. Выбор параметров вывода

Просмотр модели

1. В Логистическом узле щелкните по **Выполнить**, чтобы создать модель.

   Слепок модели будет добавлен на холст потока и также на палитру Модели в правом верхнем углу. Для просмотра его подробностей щелкните правой кнопкой мыши по слепку модели и выберите опцию **Изменить** или **Обзор**.

   На вкладке Сводка среди прочего будет показано поле назначения и входные поля (предикторы), используемые моделью. Обратите внимание на то, что эти поля были выбраны на основе прямого метода (Forwards), а не представляют собой полный список полей, переданный для рассмотрения.

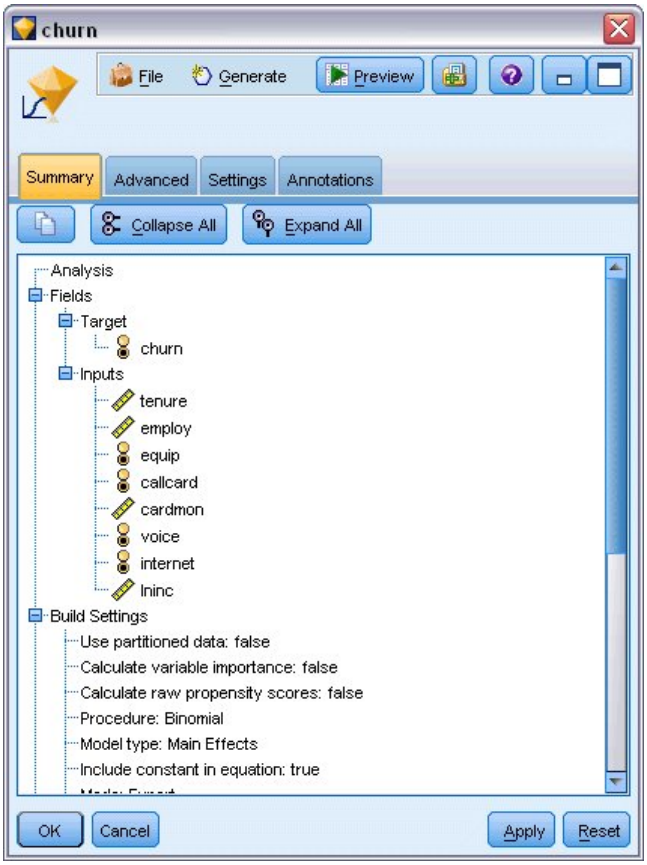


Рисунок 165. Сводка модели, иллюстрирующая целевые и входные поля

Показанные на вкладке Дополнительно элементы зависят от опций, выбранных в диалоговом окне Логистического узла Расширенный вывод. Один всегда показываемый элемент - это Сводка обработки наблюдений; здесь выводится количество и процентная доля записей, включенных в анализ. Кроме этого, здесь указывается число пропущенных наблюдений (если такие есть), когда одно или несколько входных полей были недоступны, и число невыбранных наблюдений.

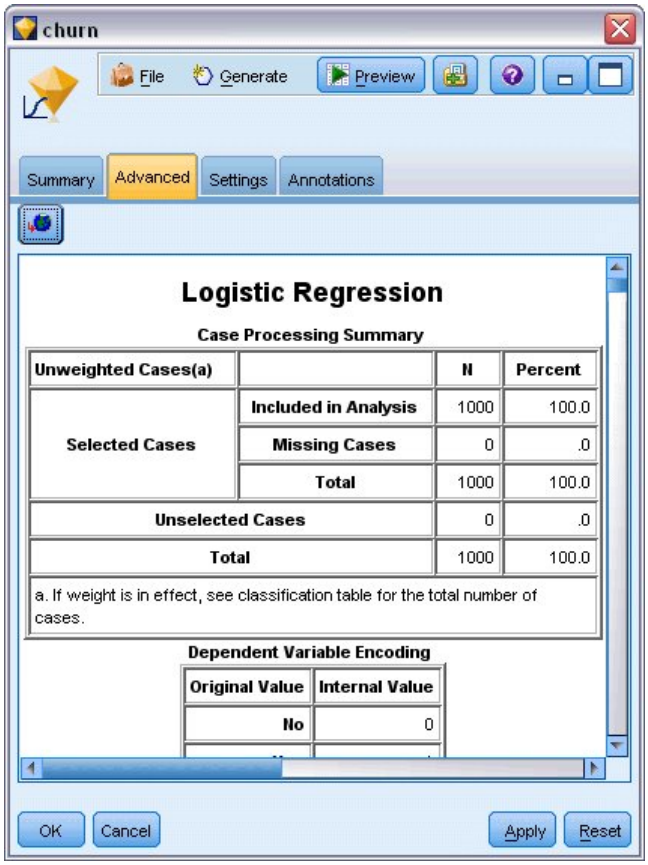


Рисунок 166. Сводный отчет обработки наблюдений

2. Прокрутите страницу вниз от Сводки обработки наблюдений, чтобы вывести таблицу классификации для Блока 0: Начальный блок.

   Прямой пошаговый метод запускается с пустой моделью, то есть с моделью без предикторов, что можно использовать как основу для сравнения с окончательной построенной моделью. По определению пустая модель для всех полей назначения предсказывает 0, то есть точность пустой модели 72,6%, просто потому, что 726 клиентов, не ушедших к конкурентам, были предсказаны правильно. Но ушедшие клиенты вовсе не были предсказаны.

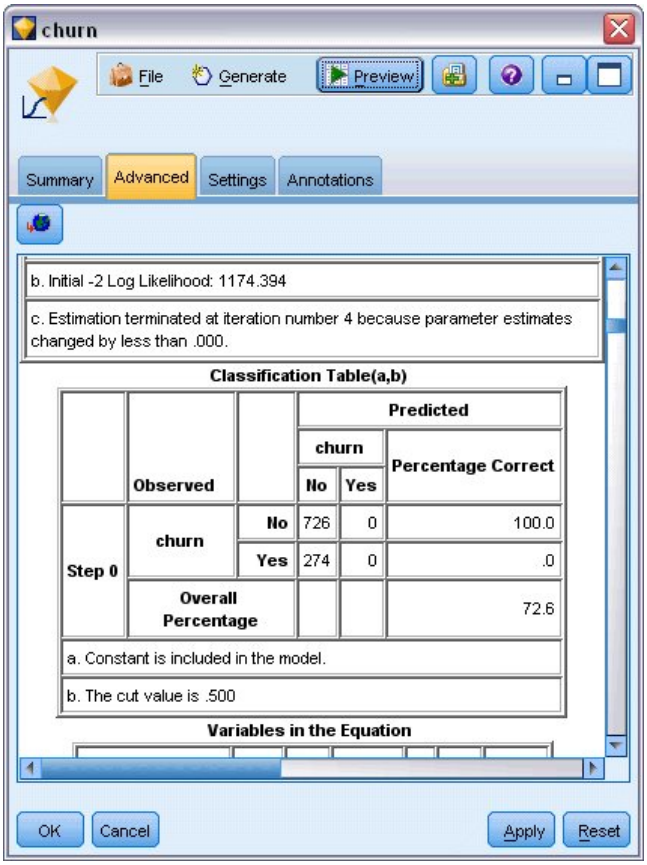


Рисунок 167. Начало таблицы классификации - Блок 0

3. Теперь прокрутите страницу еще дальше до вывода Таблицы классификации в Блоке 1: Метод = прямой пошаговый.

   Эта таблица классификации показывает результаты для вашей модели, когда на каждом шаге добавляется по предиктору. Уже на первом шаге, после использования только одного предиктора, модель увеличила точность предсказания оттока клиентов с 0,0% до 29,9%

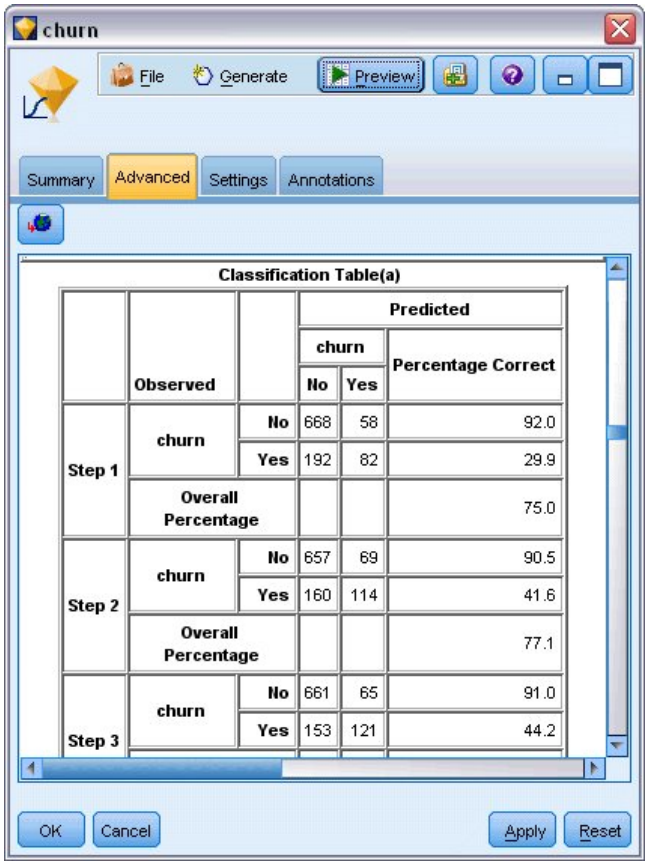


Рисунок 168. Таблица классификации - Блок 1

4. Прокрутите страницу до конца этой Таблицы классификации.

В таблице классификации видно, что последний шаг - это шаг 8. На этой стадии алгоритм решил, что больше не нужно добавлять предикторы в модель. Хотя точность предсказания оставшихся клиентов слегка уменьшилась до 91,2%, точность предсказания ухода клиентов к конкурентам выросла от 0% до 47,1%. Это существенное улучшение по сравнению с исходной пустой моделью, в которой предикторы не использовались.

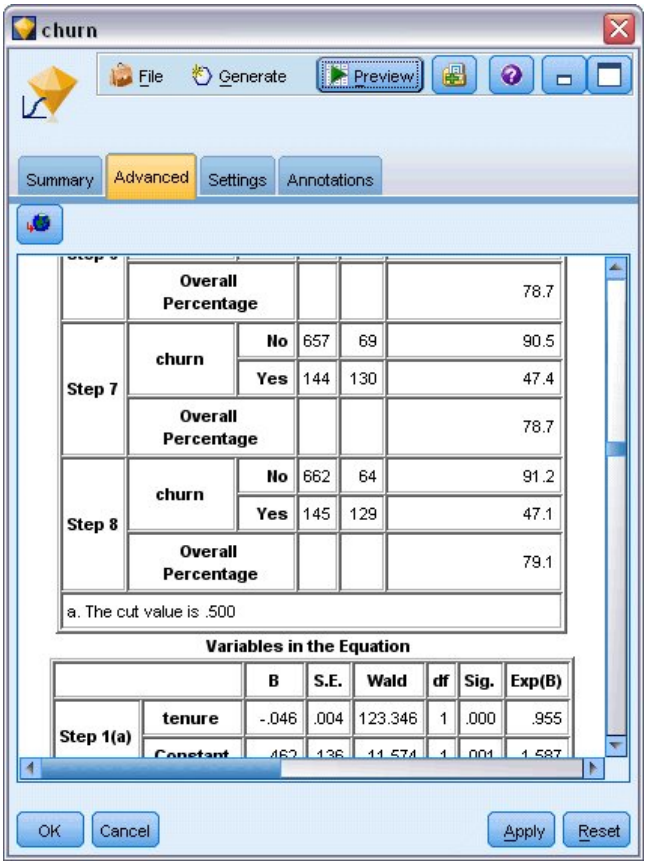


Рисунок 169. Таблица классификации - Блок 1

Для заказчика, желающего сократить отток своих клиентов, возможность сократить число потерь почти вдвое была бы важным шагом в защите своих доходов.

Примечание: Этот пример показывает также, что использование Общей процентной доли как оценки точности модели может в некоторых случаях вводить в заблуждение. У исходной пустой модели общая точность составляла 72,6%, а у получившейся предсказательной модели - 79,1%; но как мы видели, точность фактических предсказаний в отдельных категориях была существенно отличающейся.

Чтобы оценить, насколько хорошо модель на самом деле подгоняет данные, при построении модели в диалоговом окне Расширенный вывод доступно несколько диагностических средств. Объяснение математических основ методов моделирования, используемых в IBM SPSS Modeler, смотрите в публикации *IBM SPSS Modeler: Руководство по алгоритмам* в каталоге *\Documentation* на установочном диске.

Учтите, что эти результаты основаны только на обучающих данных. Чтобы оценить, насколько хорошо модель обобщается на другие данные реального мира, рекомендуется применить узел Разбиение, который будет содержать поднабор записей в целях проверки.

Глава 14. Прогноз использования пропускной способности (временной ряд)

Прогнозирование с использованием узла временных рядов

Поставщику услуг широкополосного доступа в стране требуется аналитик для прогнозов подписок пользователей с целью предсказания использования пропускной способности. Прогнозы требуются для каждого из локальных рынков, составляющих базу подписчиков по стране. Вы можете, применив моделирование временных рядов, составить прогнозы на следующие три месяца для ряда локальных рынков. Во втором примере показано, как можно преобразовать данные источника, если их формат не подходит для ввода в узел временных рядов.

В этих примерах используется поток *broadband_create_models.str*, содержащий ссылки на файл данных *broadband_1.sav*. Эти файлы находятся в папке *Demos*, внутри папки, где установлен IBM SPSS Modeler. Доступ к ним можно получить из группы программ IBM SPSS Modeler в меню Пуск Windows. Файл *broadband_create_models.str* находится в папке *streams*.

Последний пример демонстрирует, как применить сохраненные модели к обновленному набору данных, чтобы распространить прогноз на следующие три месяца.

В IBM SPSS Modeler за одну операцию можно сгенерировать несколько моделей временных рядов. В файле источника, который вы будете использовать, есть данные временных рядов для 85 различных рынков, хотя для простоты вы будете моделировать только пять из них, а также общий результат для всех рынков.

В файле *broadband_1.sav* есть данные ежемесячного использования для каждого из 85 локальных рынков. Для целей этого примера будут использоваться только первые пять рядов; отдельная модель будет создана для каждого из этих рядов, и еще одна модель - для всего рынка в целом.

Этот файл содержит также поле даты, где указан месяц и год каждой записи. Это поле будет использовано для создания меток записей. Поле даты считывается в IBM SPSS Modeler как строка, но чтобы использовать его в IBM SPSS Modeler, вы преобразуете тип хранения в числовой формат Дата, используя узел заполнения.

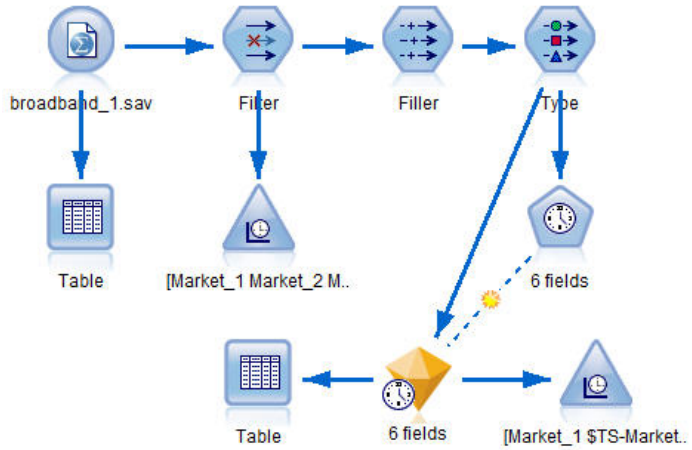
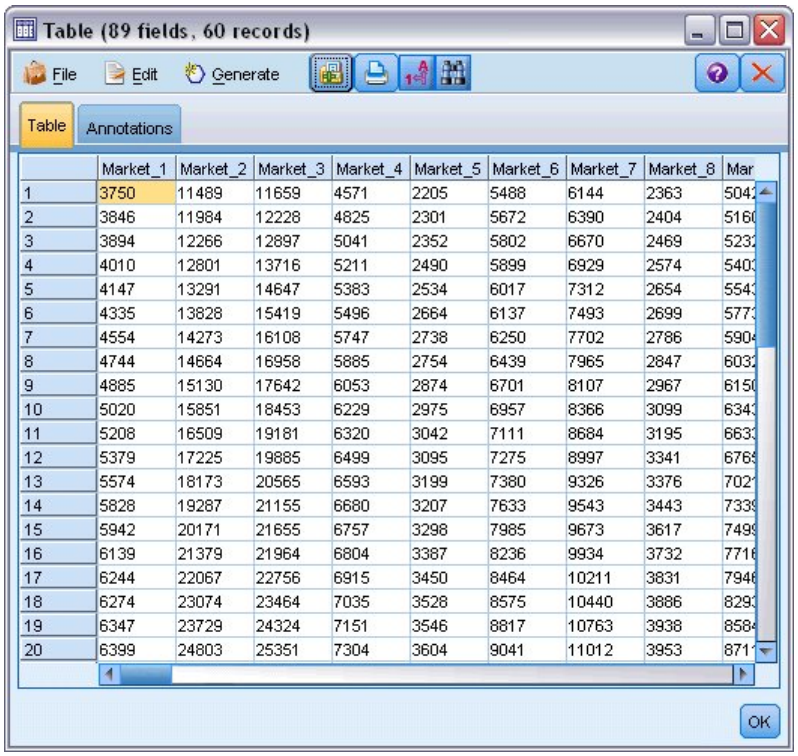


Рисунок 170. Образец потока, демонстрирующий моделирование временного ряда

Для узла временных рядов требуется, чтобы каждый ряд был в отдельном столбце со строкой для каждого интервала. При необходимости IBM SPSS Modeler предоставляет способы преобразования данных для

обеспечения этого формата.



	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	5042
2	3846	11984	12228	4825	2301	5672	6390	2404	5160
3	3894	12266	12897	5041	2352	5802	6670	2469	5232
4	4010	12801	13716	5211	2490	5899	6929	2574	5403
5	4147	13291	14647	5383	2534	6017	7312	2654	5543
6	4335	13828	15419	5496	2664	6137	7493	2699	5773
7	4554	14273	16108	5747	2738	6250	7702	2786	5904
8	4744	14664	16958	5885	2754	6439	7965	2847	6032
9	4885	15130	17642	6053	2874	6701	8107	2967	6150
10	5020	15851	18453	6229	2975	6957	8366	3099	6343
11	5208	16509	19181	6320	3042	7111	8684	3195	6633
12	5379	17225	19885	6499	3095	7275	8997	3341	6765
13	5574	18173	20565	6593	3199	7380	9326	3376	7027
14	5828	19287	21155	6680	3207	7633	9543	3443	7339
15	5942	20171	21655	6757	3298	7985	9673	3617	7499
16	6139	21379	21964	6804	3387	8236	9934	3732	7716
17	6244	22067	22756	6915	3450	8464	10211	3831	7946
18	6274	23074	23464	7035	3528	8575	10440	3886	8293
19	6347	23729	24324	7151	3546	8817	10763	3938	8584
20	6399	24803	25351	7304	3604	9041	11012	3953	8711

Рисунок 171. Ежемесячные данные подписки для локальных рынков широкополосного доступа

Создание потока

1. Создайте новый поток и добавьте узел источника Файл статистики, указывающий на файл *broadband_1.sav*.
2. С помощью узла Фильтр отфильтруйте поля с *Market_6* по *Market_85*, а также поля *MONTH_* и *YEAR_* для упрощения модели.

Совет: Для выбора в одной операции нескольких смежных полей выберите поле *Market_6* и удерживая нажатой левую кнопку мыши, перетащите указатель до поля *Market_85*. Выбранные поля будут выделены синим. Чтобы добавить поля *MONTH_* и *YEAR_*, щелкните по ним, удерживая нажатой кнопку Ctrl.

Рисунок 172. Упрощение модели

Изучение данных

Всегда полезно иметь представление о природе своих данных, прежде чем строить модель. Обнаруживают ли эти данные сезонную изменчивость? Хотя Expert Modeler может автоматически находить наилучшую сезонную или несезонную модель для каждого ряда, часто более быстрые результаты можно получить, ограничив поиск несезонными моделями, если в ваших данных отсутствует сезонность. Не исследуя данные для каждого из локальных рынков, можно получить приблизительную картину наличия или отсутствия сезонности, построив график общего числа подписчиков по всем пяти рынкам.

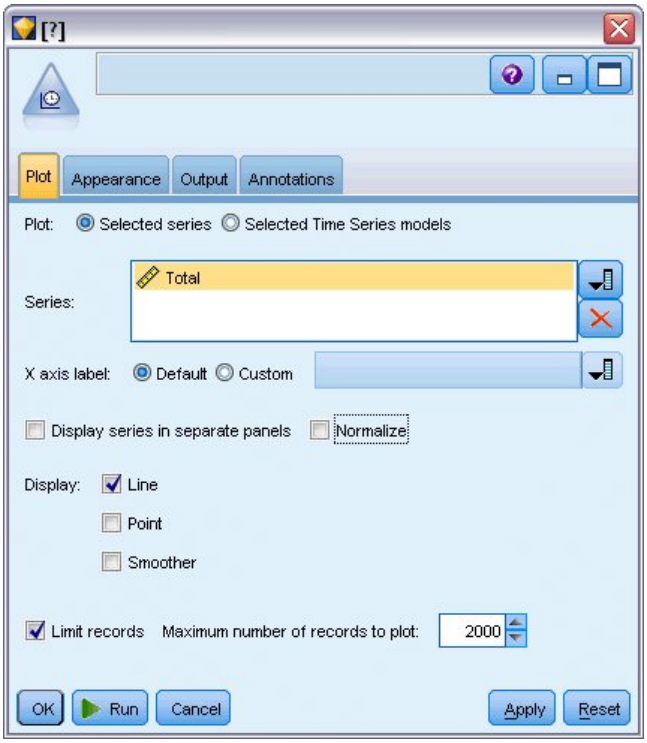


Рисунок 173. Графическое изображение общего количества подписчиков

1. В палитре Диаграммы присоедините узел График зависимости от времени к узлу Фильтр.
2. Добавьте поле *Total* в список Ряды.
3. Выключите переключатели **Показывать ряды на отдельных панелях** и **Нормализовать**.
4. Нажмите кнопку **Выполнить**.

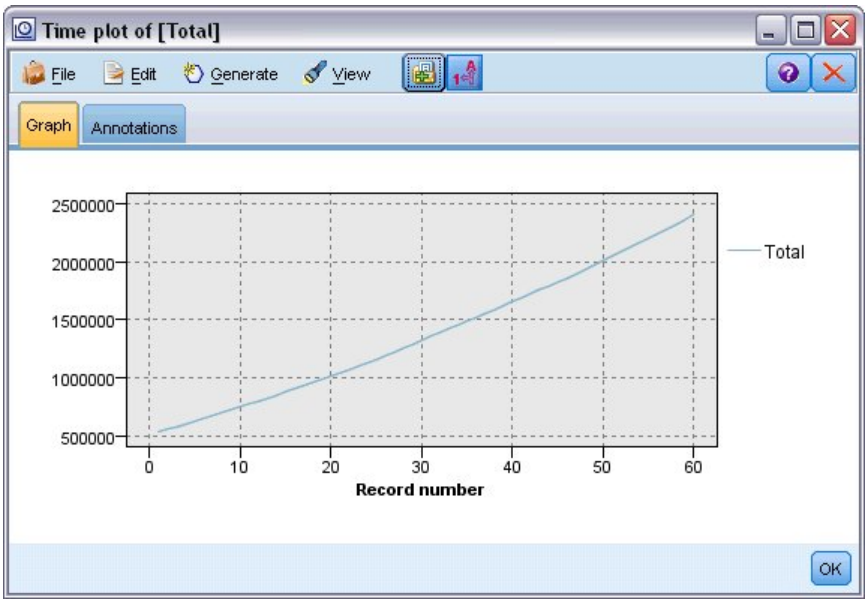


Рисунок 174. Временной график поля Всего

Ряд демонстрирует очень ровный временной тренд без признаков сезонной изменчивости. В отдельных рядах может проявляться сезонность, но судя по всему, она не выражена для всего массива данных.

Естественно, перед тем, как исключить сезонные модели, следует проверить все ряды. После этого можно выделить ряды с сезонной изменчивостью и моделировать их отдельно.

С помощью IBM SPSS Modeler легко строить совместные графики для нескольких рядов.

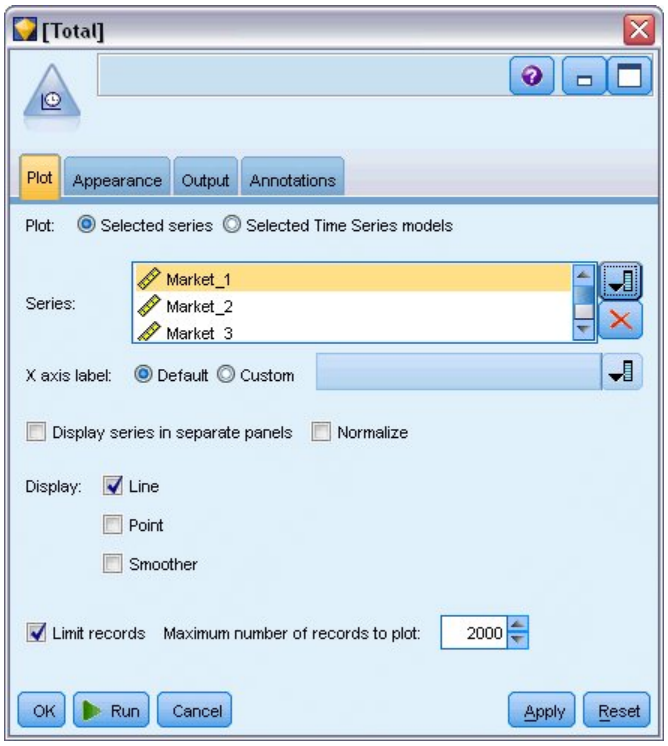


Рисунок 175. Графическое изображение нескольких временных рядов

5. Снова откройте узел Временной ряд.
6. Удалите поле *Total* из списка Ряды (выберите его и нажмите красную кнопку X).
7. Добавьте в список поля с *Market_1* по *Market_5*.
8. Нажмите кнопку **Выполнить**.

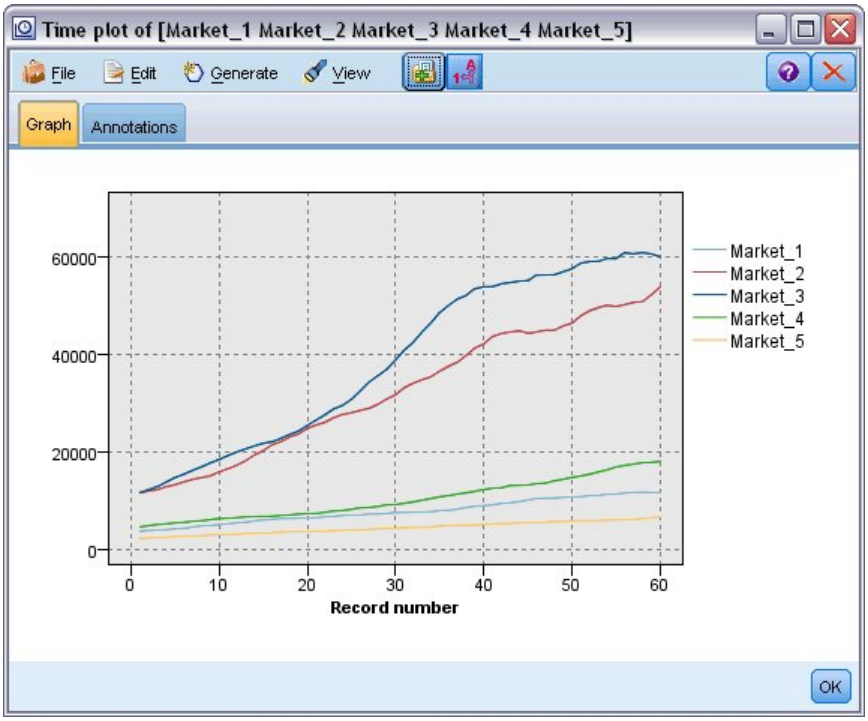


Рисунок 176. Временной график по нескольким полям

Проверка каждого из рынков выявила стойкую тенденцию к возрастанию в каждом случае. Хотя некоторые рынки демонстрируют менее устойчивые показатели, чем остальные, нет никаких видимых признаков сезонности.

Определение дат

Теперь надо изменить тип хранения поля *DATE_* на формат даты.

1. Присоедините узел Заполнитель к узлу Фильтр.
2. Откройте узел Заполнитель и нажмите кнопку выбора полей.
3. Выберите **DATE_** для добавления в **Заполнить поля**.
4. Задайте для условия **Заменять** значение **Всегда**.
5. Задайте для **Заменять на** значение **to_date(DATE_)**.

Рисунок 177. Задание типа хранения даты

Измените формат даты по умолчанию, чтобы он соответствовал формату поля Date. Это требуется для правильного преобразования поля Date.

6. Выберите в меню **Инструменты > Свойства потока > Опции**, чтобы открыть диалоговое окно Опции потока.
7. Выберите панель **Дата/Время** и задайте **Формат даты** по умолчанию как **MON YYYY** .

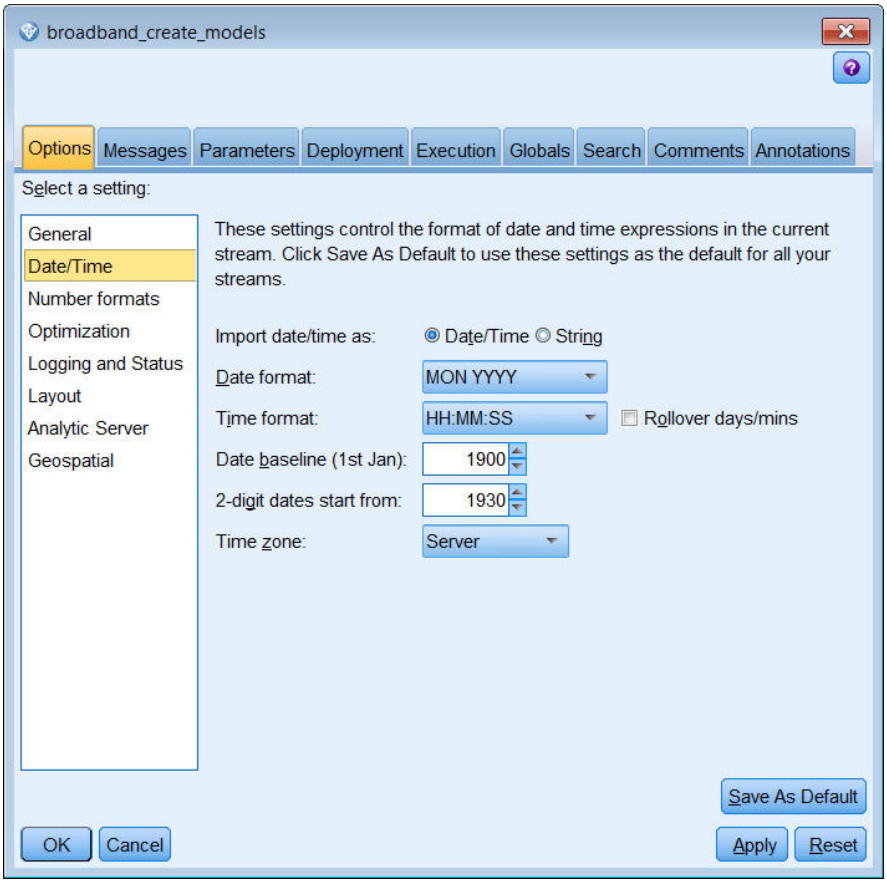


Рисунок 178. Задание формата даты

Определение назначений

1. Добавьте узел Тип и задайте роль **Нет** для поля *DATE_*. Задайте роль **Назначение** для всех остальных полей (*Market_n* и *Total*).
2. Нажмите кнопку **Читать значения**, чтобы заполнить столбец Значения.

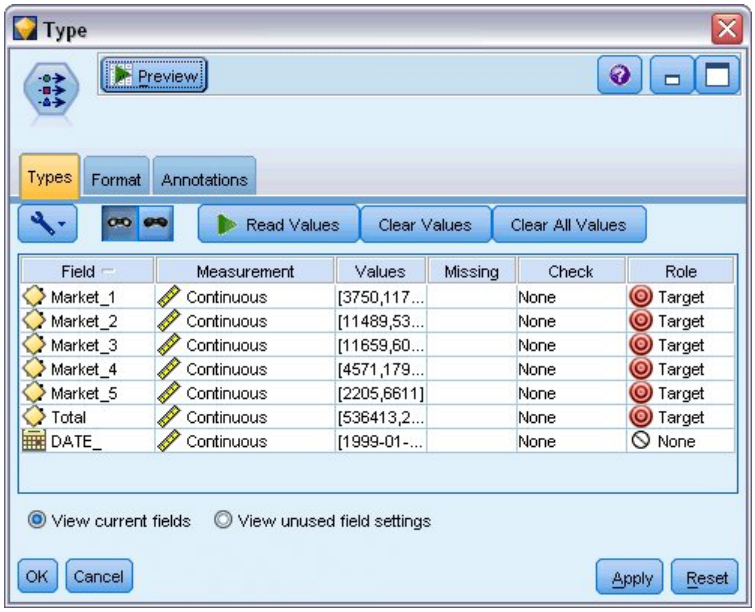


Рисунок 179. Задание роли для нескольких полей

Задание временных интервалов

1. На палитре Моделирование добавьте в поток узел Временные ряды и присоедините его к узлу Тип.
2. На вкладке Спецификации данных в панели Наблюдения выберите DATE_ в качестве поля **Дата/Время**.
3. Выберите Месяцы как **Интервал времени**.

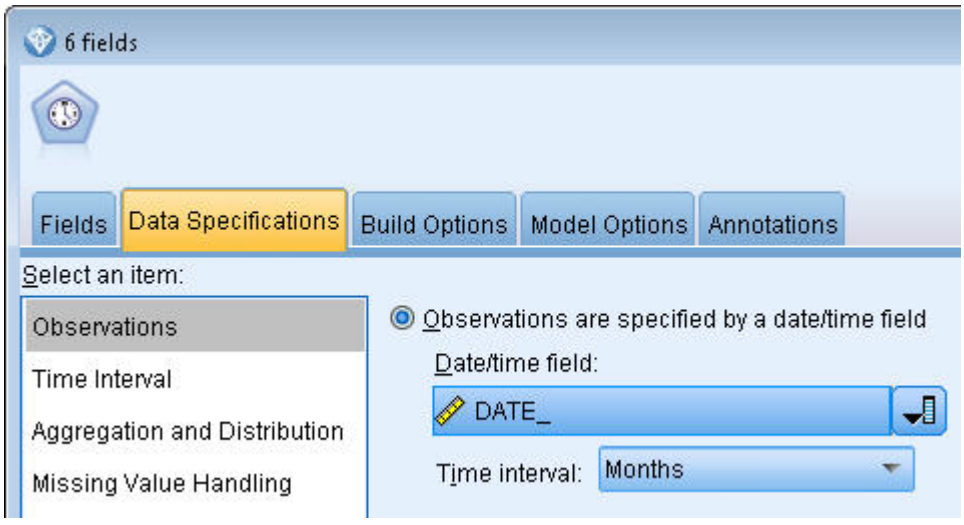


Рисунок 180. Задание временного интервала

4. На вкладке Опции модели включите переключатель **Распространить записи на будущее**.
5. Задайте значение **3**.

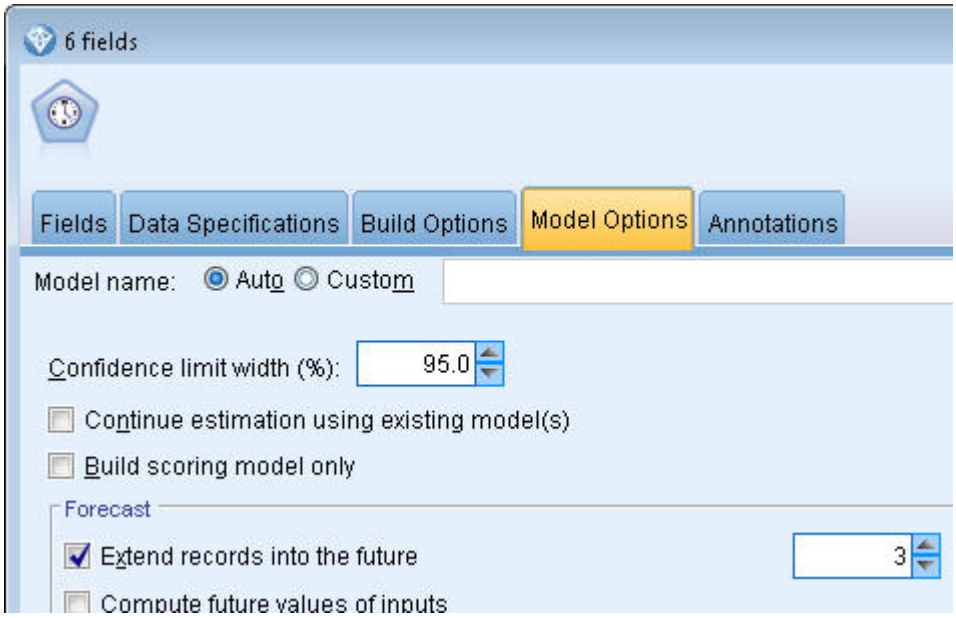


Рисунок 181. Задание периода прогноза

Создание модели

1. В узле Временные ряды выберите вкладку Поля. В списке **Поля** выберите все 5 рынков и скопируйте их в оба списка: **Назначения** и **Входные поля - кандидаты**. Кроме того, выберите поле Total и скопируйте его в список **Назначения**.
2. Выберите вкладку Опции построения и на панели Общие проверьте, что выбран **Метод** Expert Modeler с использованием всех параметров по умолчанию. Это позволяет Expert Modeler предлагать наиболее подходящую модель для работы с каждым временным рядом. Нажмите кнопку **Выполнить**.

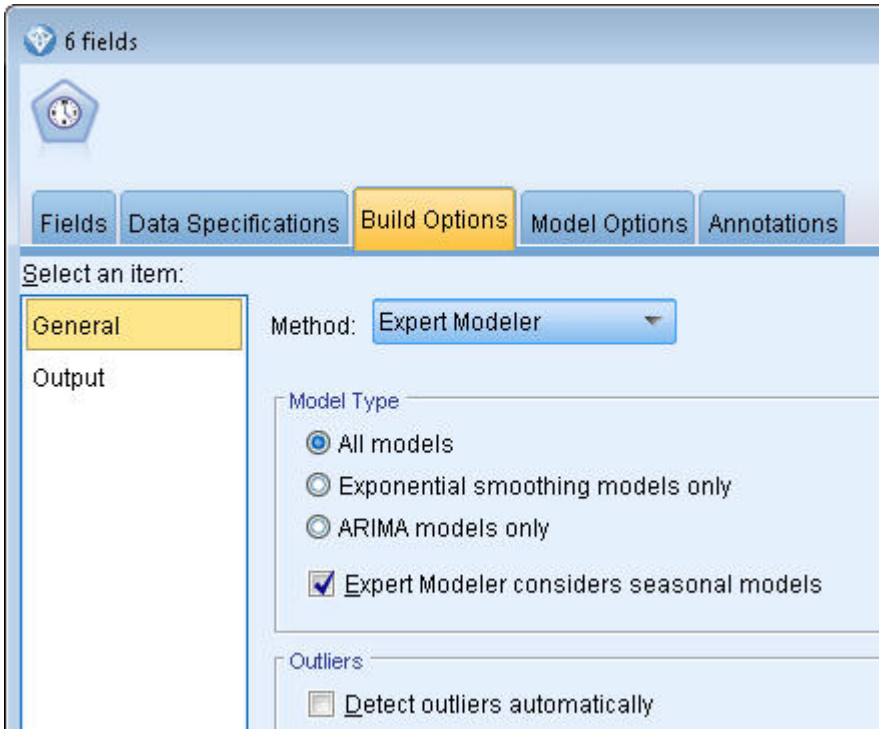


Рисунок 182. Выбор Expert Modeler для временных рядов

3. Присоедините слепок модели временного ряда к узлу Временные ряды.
4. Присоедините узел Таблица к слепку модели временного ряда и нажмите кнопку **Выполнить**.

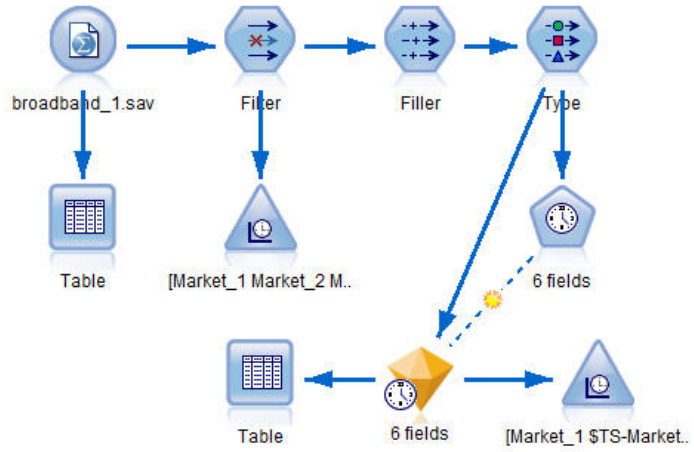


Рисунок 183. Образец потока, демонстрирующий моделирование временного ряда

Теперь к исходным данным добавлены три новых строки (с 61 по 63). Это строки для прогностического периода, в данном случае с января по март 2004 года.

Появилось также несколько новых столбцов, столбцы с префиксом *$TS-* добавлены узлом Временные ряды. В этих столбцах содержится следующая информация для каждой строки (то есть для каждого интервала в данных временного ряда):

Столбец	Описание
$TS-*имя_столбца*	Данные сгенерированной модели для каждого столбца исходных данных.
$TSLCI-*имя_столбца*	Нижнее значение доверительного интервала для каждого столбца данных сгенерированной модели.
$TSUCI-*имя_столбца*	Верхнее значение доверительного интервала для каждого столбца данных сгенерированной модели.
$TS-Итог	Итог по всем значениям $TS-*имя_столбца* для данной строки.
$TSLCI-Итог	Итог по всем значениям $TSLCI-*имя_столбца* для данной строки.*
$TSUCI-Итог	Итог по всем значениям $TSUCI-*имя_столбца* для данной строки.

Наиболее важные столбцы для выполнения прогнозирования - *$TS-Market_n*, *$TSLCI-Market_n* и *$TSUCI-Market_n*. В частности, эти столбцы содержат в строках с 61 по 63 данные прогнозов подписки пользователей и доверительные интервалы для каждого из локальных рынков.

Изучение модели

1. Щелкните дважды по слепку модели временного ряда и выберите вкладку Вывод для вывода данных о моделях, сгенерированных для каждого из рынков.

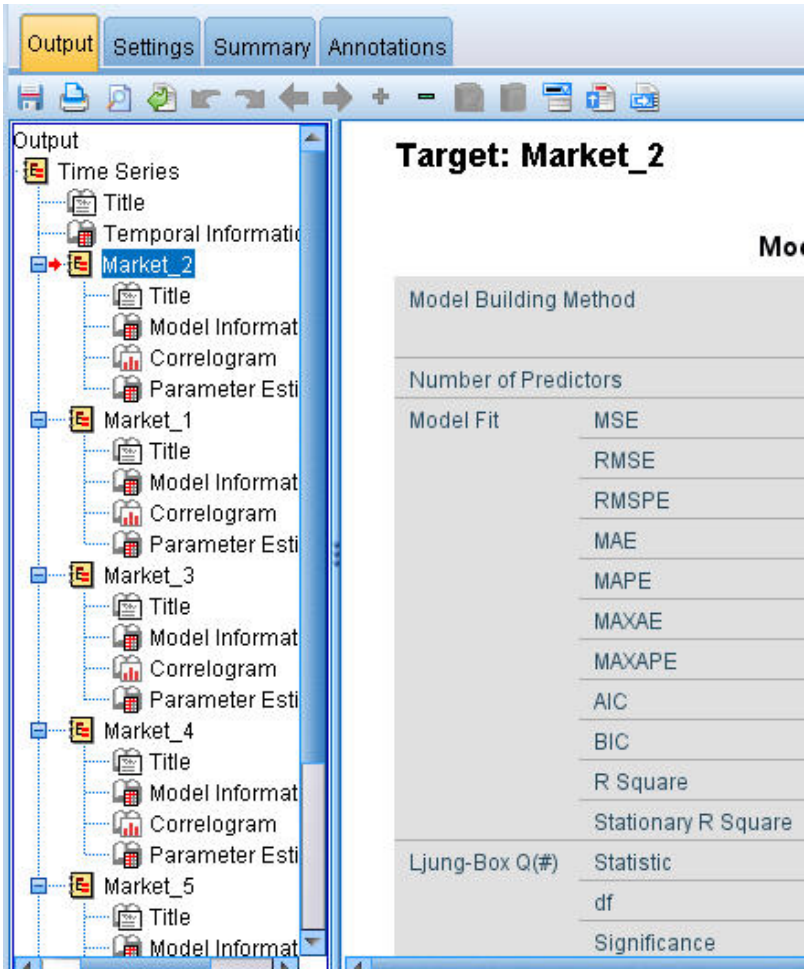


Рисунок 184. Модели временных рядов, сгенерированные для рынков

В левом столбце вкладки Вывод выберите **Информация о модели** для любого рынка. Линия **Число предикторов** показывает, сколько полей было использовано в качестве предикторов для каждого назначения, в данном случае таковых нет.

Остальные линии в таблицах **Информация о модели** показывают различные меры согласия для каждой модели. Значение **Стационарный R-квадрат** дает оценку доли общей вариации в ряду, которая объясняется данной моделью. Чем выше это значение (максимум 1,0), тем лучше согласие модели.

Линии **Q(#)-статистика**, **df** и **Значимость** относятся к статистике Льюнга-Бокса, в тесте на случайность остаточных ошибок модели; чем случайнее эти ошибки, тем выше качество модели. **Q(#)** - это собственно статистика Льюнга-Бокса, а значение **df** (степени свободы) указывает число параметров модели, которые могут варьировать при оценке конкретного назначения.

Линия **Значимость** показывает уровень значимости статистики Льюнга-Бокса, что является еще одним критерием корректности заданной модели. При уровне значимости менее 0,05 остаточные ошибки нельзя считать случайными, что означает, что в наблюденных рядах имеется структура, не описываемая данной моделью.

Судя по значениям **Стационарный R-квадрат** и **Значимость**, модели, выбранные Expert Modeler для объектов *Market_3* и *Market_4* вполне приемлемы. Значения меры **Значимость** для объектов *Market_1*, *Market_2* и *Market_5* меньше 0,05, что указывает на необходимость подбора для этих рынков моделей с лучшим согласием.

На экран выводится несколько дополнительных мер согласия моделей. Значение **R-квадрат** дает оценку общей вариации во временном ряду, которая объясняется данной моделью. Поскольку максимальное значение для этой статистики - 1,0, наши модели в этом отношении удачны.

СКО - это среднеквадратичная ошибка, которая показывает, насколько фактические значения отличаются от значений, предсказанных моделью. Она выражается в тех же единицах, что и значения самого ряда. Поскольку это оценка ошибки, нужно, чтобы ее значение было как можно более низким. На первый взгляд, может показаться, что хотя модели для *Market_2* и *Market_3*, приемлемы, судя по представленным выше статистическим критериям, они менее удачны, чем модели для остальных трех рынков.

Эти дополнительные меры согласия включают среднюю абсолютную ошибку в процентах (**САОП**) и ее максимальное значение (**МСАОП**). Абсолютная ошибка в процентах показывает, насколько исследуемый ряд отклоняется от значений, предсказанных моделью, в процентном выражении. Изучив среднее и максимальное значения для всех моделей, можно оценить уровень погрешности в ваших прогнозах.

Судя по значениям САОП, все модели дают среднюю погрешность порядка 1%, то есть весьма низкую. Значение МСАОП представляет максимальную абсолютную ошибку в процентах и помогает представить наихудший возможный сценарий в ваших прогнозах. Для большинства моделей максимальная ошибка лежит в пределах от 1,8% до 3,7%, что также совсем немного, и лишь для рынка *Market_4* ошибка выше (около 7%).

Значение **САО** (средняя абсолютная ошибка) - это среднее арифметическое абсолютных значений ошибок прогноза. Как и в случае СКО, эта ошибка выражается в тех же единицах, что и значения самого ряда. **МАО** (максимальная абсолютная ошибка) - это наибольшая ошибка прогноза, выраженная в тех же единицах и соответствующая наихудшему возможному сценарию прогноза.

Хотя эти абсолютные значения представляют определенный интерес, в данном случае более полезны относительные ошибки (САОП и МСАОП), поскольку в исследуемом ряду представлено число подписчиков для рынков разного размера.

Представляют ли САОП и МСАОП приемлемый уровень погрешности для моделей? Их значения, безусловно, очень низки. В этой ситуации приходится руководствоваться деловой интуицией, поскольку приемлемый риск может быть разным для разных проблем. Примем, что значения статистик согласия лежат в приемлемых пределах, и перейдем к остаточным ошибкам.

Исследование значений автокорреляционной функции (АКФ) и частной автокорреляционной функции (ЧАКФ) для остатков в модели дает больше возможностей количественного анализа, чем простой просмотр статистик согласия.

Удачно заданная модель временных рядов объясняет всю неслучайную вариацию, включая сезонность, тренды, циклы и другие важные факторы. Если это так, ошибки модели не должны коррелировать сами с собой (автокорреляция) во времени. Заметная структурированность в какой-либо из автокорреляционных функций говорит о неполноте используемой модели.

2. Для четвертого рынка щелкните в левом столбце по **Коррелограмма**, чтобы показать значения автокорреляционной функции (АКФ) и частной автокорреляционной функции (ЧАКФ) для остаточных ошибок в модели.

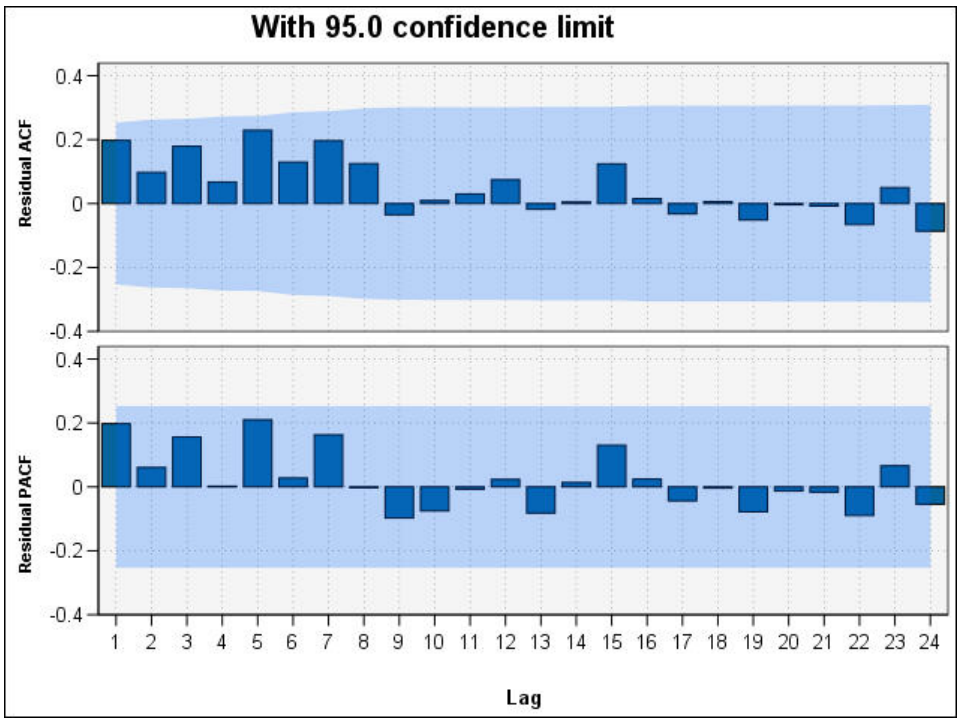


Рисунок 185. Значения АКФ и ЧАКФ для четвертого рынка

На этих графиках значения переменной ошибок сдвинуты относительно исходных на величину до 24 периодов времени и сопоставлены с исходными для выявления корреляций во времени. Модель можно считать приемлемой, если полоски на верхнем графике (АКФ) не выходят за пределы затененной области ни в положительном (вверх), ни в отрицательном (вниз) направлении.

Если это имеет место, нужно посмотреть нижний график (ЧАКФ), чтобы убедиться, что он подтверждает выявленную структуру. График ЧАКФ отображает корреляции после исключения значений ряда в промежуточные моменты времени.

Значения для *Market_4* лежат целиком в затененной области, поэтому можно продолжить и посмотреть значения для других рынков.

3. Щелкните по **Коррелограмма** для каждого из оставшихся полей market и total.

   Значения для других рынков во всех случаях местами выходят за пределы затененной области, подтверждая наши предположения, сделанные при просмотре уровней **Значимости**. Для этих рынков нужно будет поэкспериментировать с другими моделями в поисках лучшей подгонки, но в завершающей части данного примера мы сосредоточимся на дополнительной информации, которую можно извлечь из модели *Market_4*.

4. На палитре Диаграммы присоедините узел График зависимости от времени к слепку модели временного ряда.
5. На вкладке График выключите переключатель **Выводить ряды на отдельных панелях**.
6. В списке **Ряды** нажмите кнопку выбора поля, выберите поля *Market_4* и *$TS-Рынок_4* и нажмите кнопку **OK**, чтобы добавить их в список.
7. Нажмите кнопку **Выполнить** для вывода линейной диаграммы фактических и предсказанных данных по первому из локальных рынков.

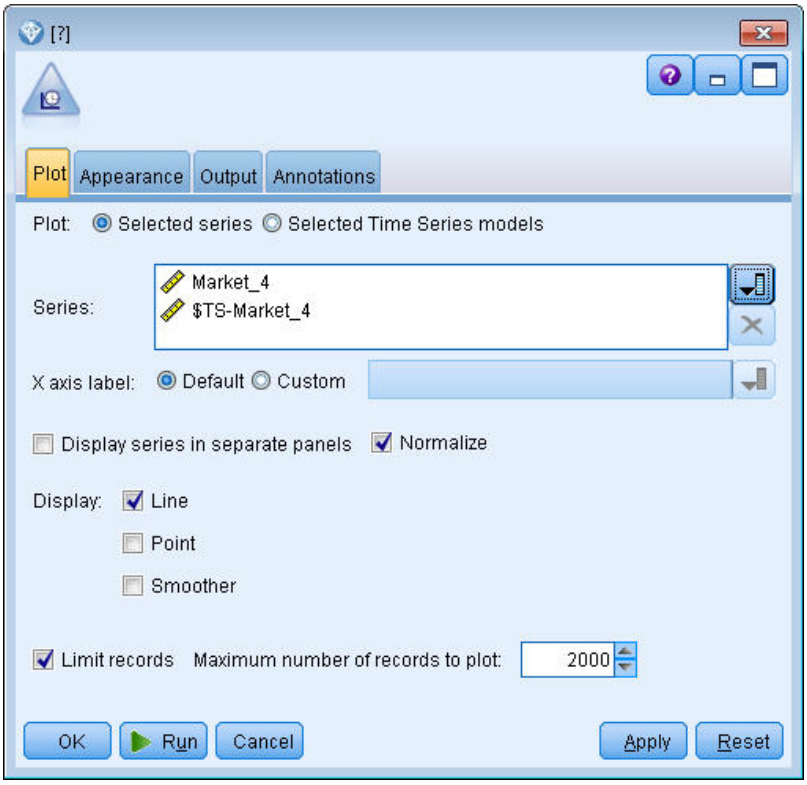


Рисунок 186. Выбор полей для графика

Обратите внимание на ход линии прогноза (*$TS-Market_4*) за пределами диапазона фактических данных. Это прогноз ожидаемого спроса для данного рынка на следующие три месяца.

Линии для фактических и прогнозируемых данных на графике на протяжении всего временного ряда идут очень близко друг к другу, указывая, что данная модель надежна для конкретного временного ряда.

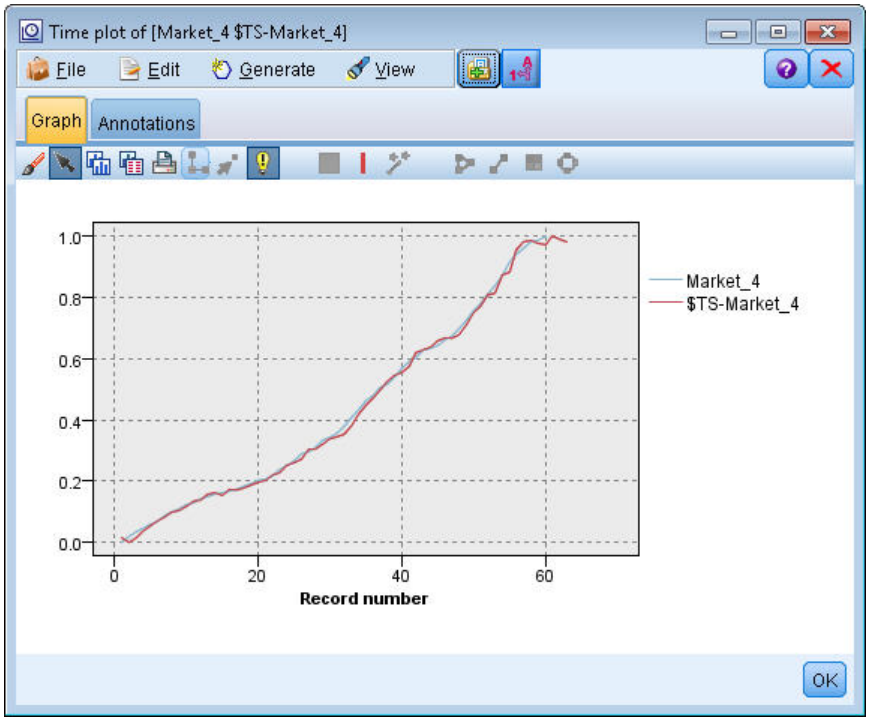


Рисунок 187. Временной график фактических и прогностических данных для Market_4

Сохраните модель в файле для использования в будущем примере:

8. Выберите **ОК**, чтобы закрыть текущую диаграмму.
9. Откройте слепок модели Временной ряд.
10. Выберите **Файл > Сохранить узел** и задайте положение файла.
11. Нажмите кнопку **Сохранить** .

    Теперь у вас есть надежная модель для конкретного рынка, но какова погрешность этого прогноза? Узнать об этом можно, посмотрев значение доверительного интервала.
12. Щелкните дважды по узлу График зависимости от времени в потоке (помеченному как **Market_4 $TS-Market_4**), чтобы снова открыть его диалоговое окно.
13. Нажмите кнопку выбора полей и добавьте поля *$TSLCI-Market_4* и *$TSUCI-Market_4* в список **Ряды**.
14. Нажмите кнопку **Выполнить**.

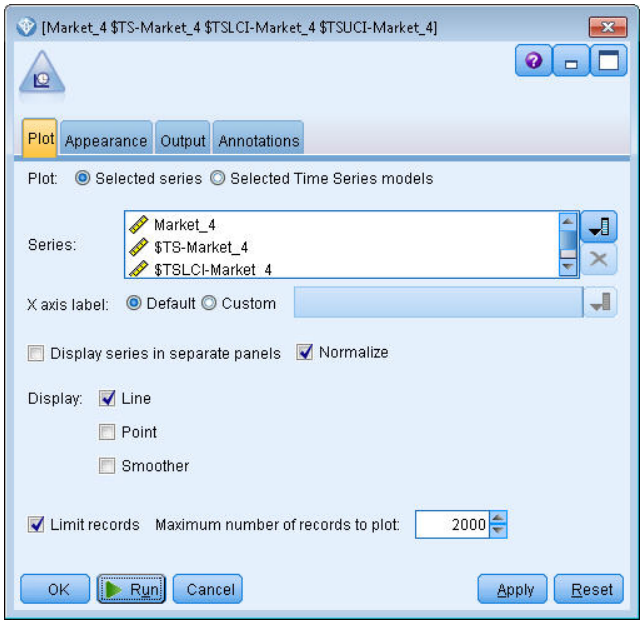


Рисунок 188. Добавление полей к графику

Теперь на вашу диаграмму нанесены верхняя (*$TSUCI*) и нижняя (*$TSLCI*) границы доверительного интервала.

Обратите внимание на то, как границы доверительного интервала расходятся за пределами прогнозируемого периода, указывая на растущую неопределенность при переходе к более долгосрочных прогнозам.

Однако поскольку каждый период времени проходит, у вас появится очередной (в данном случае) месяц с фактическими данными, на которых можно базировать дальнейший прогноз. Теперь, когда известно, что используемая модель надежна, можно считать новые данные в поток и повторно применить модель. Дополнительную информацию смотрите в разделе “Повторное применение модели временных рядов” на стр. 171.

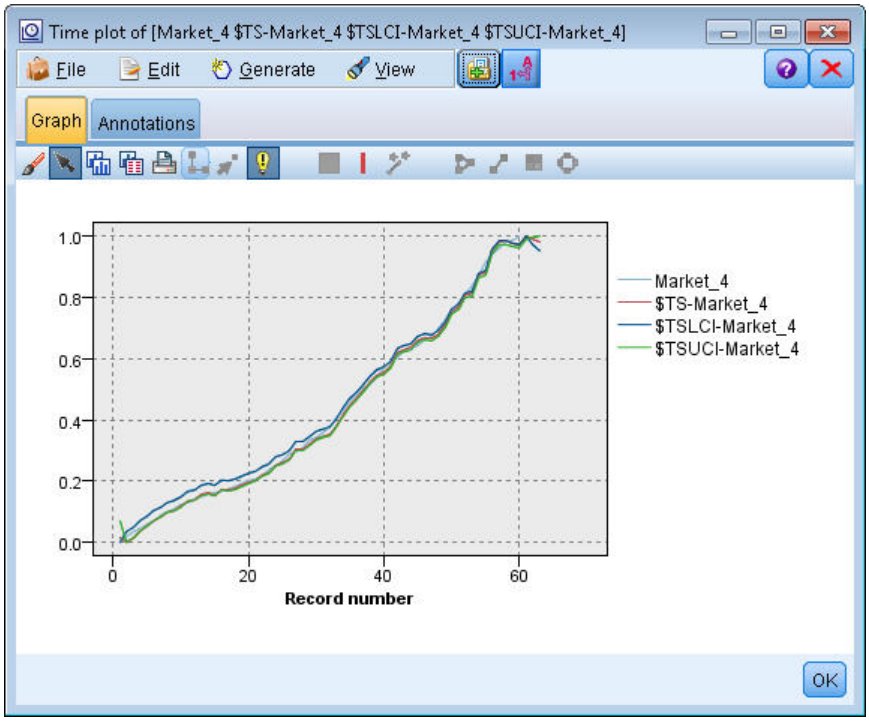


Рисунок 189. Добавлен график зависимости от времени с доверительным интервалом

Итог

Вы научились использовать эксперт построения моделей (Expert Modeler), чтобы создавать прогнозы для нескольких временных рядов, и сохранили получившиеся модели во внешнем файле.

В следующем примере вы увидите, как преобразовывать нестандартные данные временных рядов в формат, подходящий для ввода в узел временных рядов.

Повторное применение модели временных рядов

В этом примере применяются модели временных рядов из первого примера временного ряда, но его можно использовать и независимо. Дополнительную информацию смотрите в разделе “Прогнозирование с использованием узла временных рядов” на стр. 153.

Как и в исходном сценарии, поставщику услуг широкополосного доступа в стране требуется аналитик для ежемесячных прогнозов подписок пользователей для каждого из локальных рынков с целью предсказания требований к пропускной способности. Вы уже использовали эксперт построения моделей (Expert Modeler), чтобы создать модели и выполнить прогнозирование на три месяца вперед.

Теперь ваше хранилище данных обновлено фактическими данными за исходный период прогноза, поэтому было бы хорошо использовать эти данные для расширения горизонта прогнозирования на следующие три месяца.

В этом примере используется поток *broadband_apply_models.str*, содержащий ссылки на файл данных *broadband_2*. Эти файлы находятся в папке *Demos*, внутри папки, где установлен IBM SPSS Modeler. Доступ к ним можно получить из группы программ IBM SPSS Modeler в меню Пуск Windows. Файл *broadband_apply_models.str* находится в папке *streams*.

Получение потока

В этом примере вы будете воссоздавать узел временных рядов по модели временного ряда, сохраненной в первом примере. Не беспокойтесь, если вы забыли сохранить модель - ее можно получить в папке *Demos*.

1. Откройте поток *broadband_apply_models.str* из подпапки *streams* папки *Demos*.

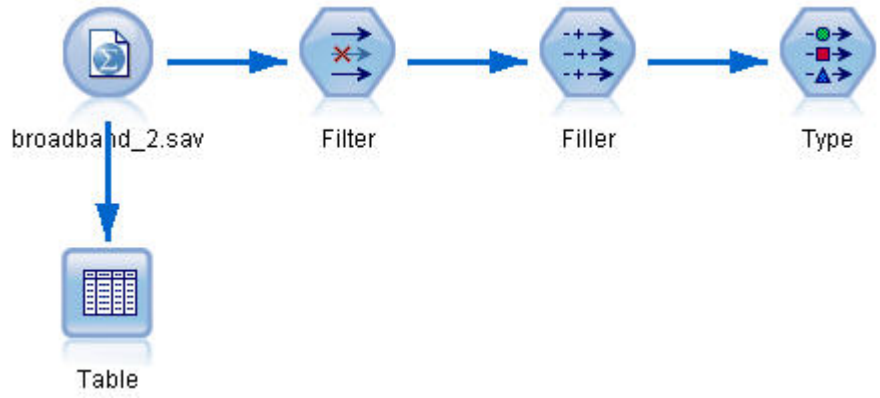
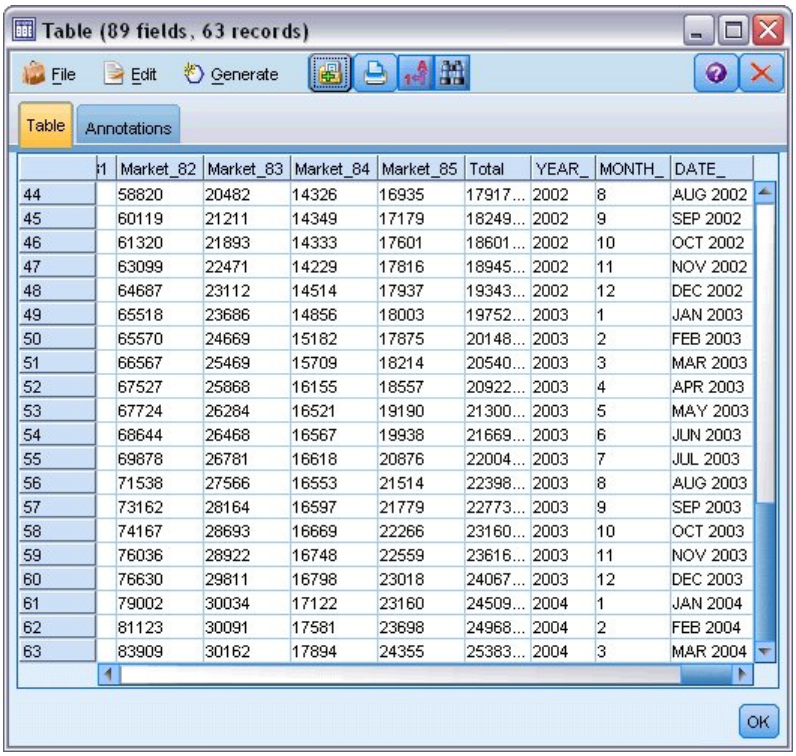


Рисунок 190. Открытие потока

Обновленные ежемесячные данные собраны в файле *broadband_1.sav*.

2. Присоедините узел Таблица к узлу источника Файл IBM SPSS Statistics, откройте узел Таблица и нажмите кнопку **Выполнить**.

Примечание: Файл данных обновлен данными фактических продаж за январь-март 2004 в строках с 61 по 63.



	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

Рисунок 191. Обновленные данные о продажах

Получение сохраненной модели

1. В меню IBM SPSS Modeler выберите **Вставить > Узел из файла** и выберите файл *TSmodel.nod* в папке *Demos* (или воспользуйтесь моделью временного ряда, сохраненной вами в первом примере временного ряда).

Этот файл содержит модели временного ряда из предыдущего примера. Операция вставки поместит соответствующий слепок модели временного ряда на холст.

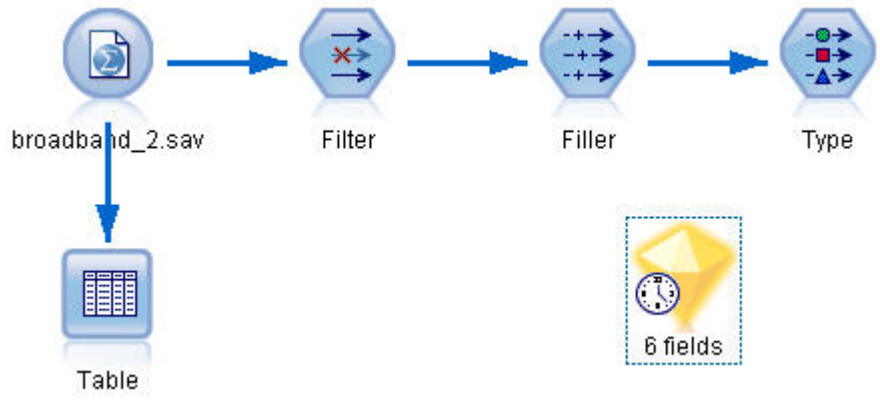


Рисунок 192. Добавление слепка модели

Генерирование узла моделирования

1. Откройте слепок модели временного ряда и выберите **Сгенерировать > Сгенерировать узел моделирования**.

В результате узел моделирования временных рядов будет помещен на холст.

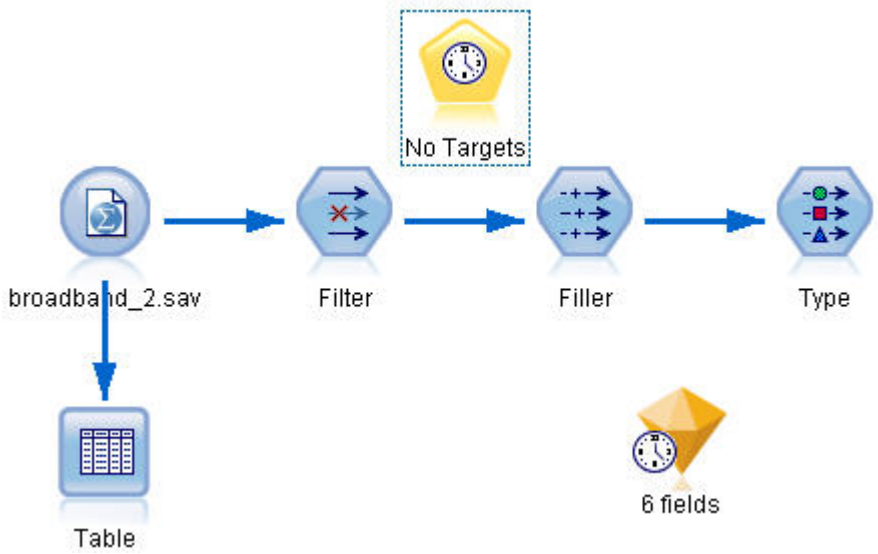


Рисунок 193. Генерирование узла моделирования по слепку модели

Создание новой модели

1. Закройте слепок модели временного ряда и удалите его с холста.

   Старая модель была основана на 60 строках данных. Необходимо сгенерировать новую модель на основе обновленных данных о продажах (63 строки).

2. Присоедините к потоку вновь сгенерированный узел построения временных рядов.

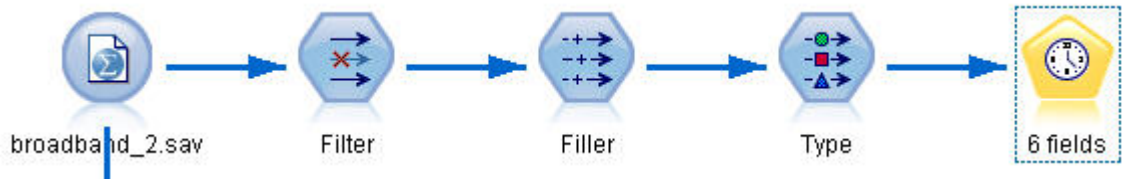


Рисунок 194. Присоединение узла моделирования к потоку

3. Откройте узел Временной ряд.
4. На вкладке **Опции модели** проверьте, что включена опция **Продолжить оценивание, используя существующие модели**.

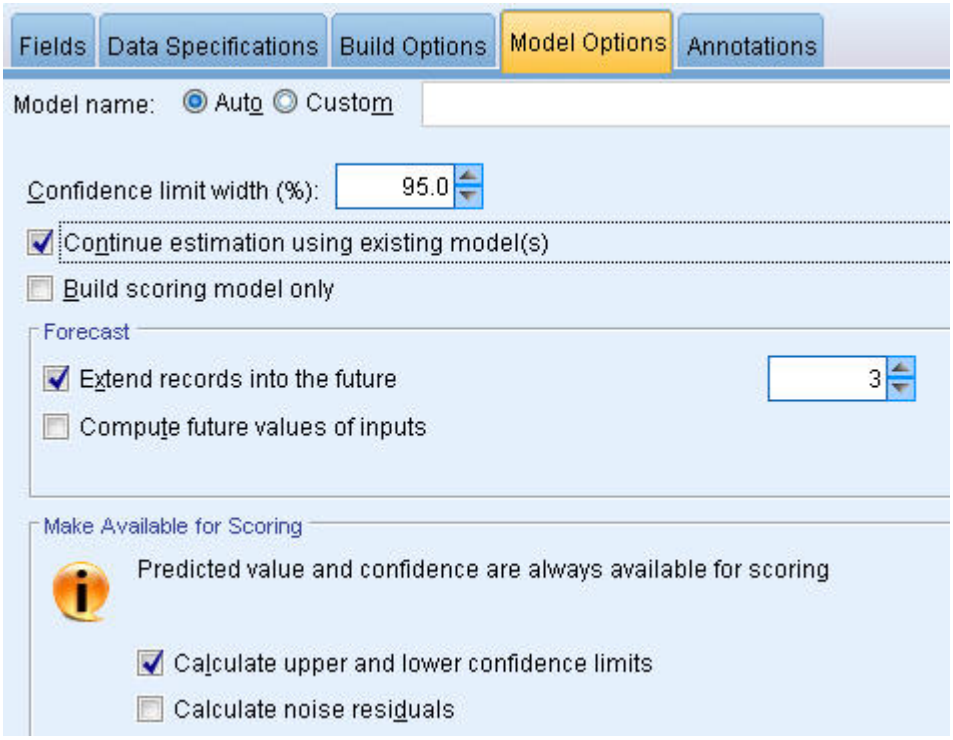


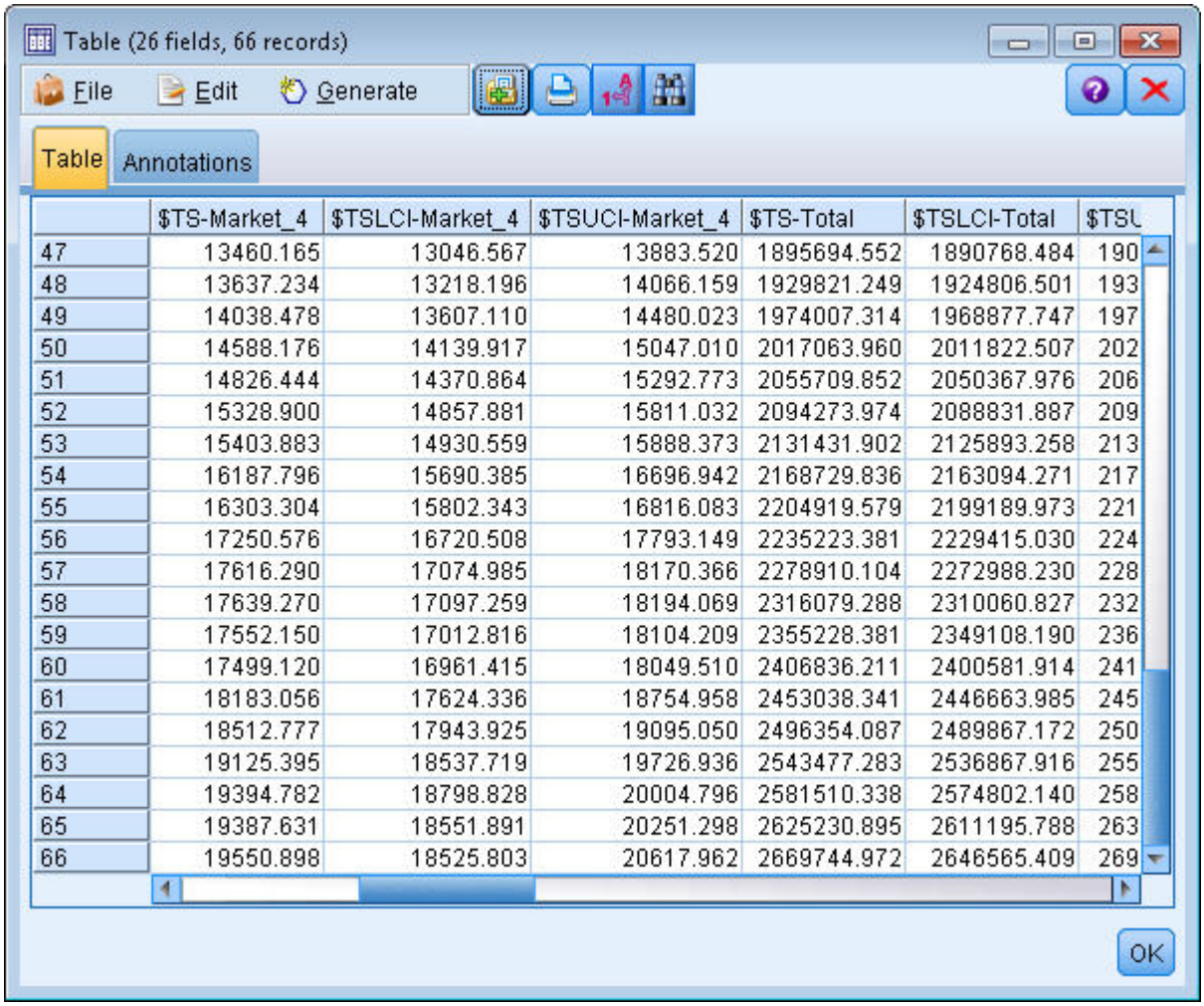
Рисунок 195. Повторное использование сохраненных параметров для модели временного ряда

5. Убедитесь, что для опции **Распространить записи на будущее** задано значение **3**.
6. Нажмите кнопку **Выполнить**, чтобы поместить слепок новой модели на холст и на палитру Модели.

Изучение новой модели

1. Присоедините узел Таблица к слепку новой модели временного ряда на холсте.
2. Откройте узел Таблица и щелкните по **Выполнить**.

   В новой модели прогноз по-прежнему выполняется на три месяца вперед, поскольку вы используете сохраненные параметры. Однако на этот раз интервал прогнозирования - с апреля по июнь (в строках с 64 по 66), поскольку теперь период оценки заканчивается в марте, а не в январе.



	$TS-Market_4	$TSLCI-Market_4	$TSUCI-Market_4	$TS-Total	$TSLCI-Total	$TSU
47	13460.165	13046.567	13883.520	1895694.552	1890768.484	190
48	13637.234	13218.196	14066.159	1929821.249	1924806.501	193
49	14038.478	13607.110	14480.023	1974007.314	1968877.747	197
50	14588.176	14139.917	15047.010	2017063.960	2011822.507	202
51	14826.444	14370.864	15292.773	2055709.852	2050367.976	206
52	15328.900	14857.881	15811.032	2094273.974	2088831.887	209
53	15403.883	14930.559	15888.373	2131431.902	2125893.258	213
54	16187.796	15690.385	16696.942	2168729.836	2163094.271	217
55	16303.304	15802.343	16816.083	2204919.579	2199189.973	221
56	17250.576	16720.508	17793.149	2235223.381	2229415.030	224
57	17616.290	17074.985	18170.366	2278910.104	2272988.230	228
58	17639.270	17097.259	18194.069	2316079.288	2310060.827	232
59	17552.150	17012.816	18104.209	2355228.381	2349108.190	236
60	17499.120	16961.415	18049.510	2406836.211	2400581.914	241
61	18183.056	17624.336	18754.958	2453038.341	2446663.985	245
62	18512.777	17943.925	19095.050	2496354.087	2489867.172	250
63	19125.395	18537.719	19726.936	2543477.283	2536867.916	255
64	19394.782	18798.828	20004.796	2581510.338	2574802.140	258
65	19387.631	18551.891	20251.298	2625230.895	2611195.788	263
66	19550.898	18525.803	20617.962	2669744.972	2646565.409	269

Рисунок 196. Таблица, содержащая новый прогноз

3. Присоедините узел Диаграмма зависимости от времени к слепку модели временного ряда.

   На этот раз мы будем использовать вывод временного графика, специально разработанный для моделей временных рядов.
4. На вкладке График задайте для **метка оси X** значение **Пользовательская** и выберите `Date_`.
5. Для параметра **График** выберите опцию **Выбранные модели временных рядов**.
6. В списке **Ряды** нажмите кнопку выбора полей, выберите поле `$TS-Market_4` и нажмите кнопку **OK**, чтобы добавить его в список.

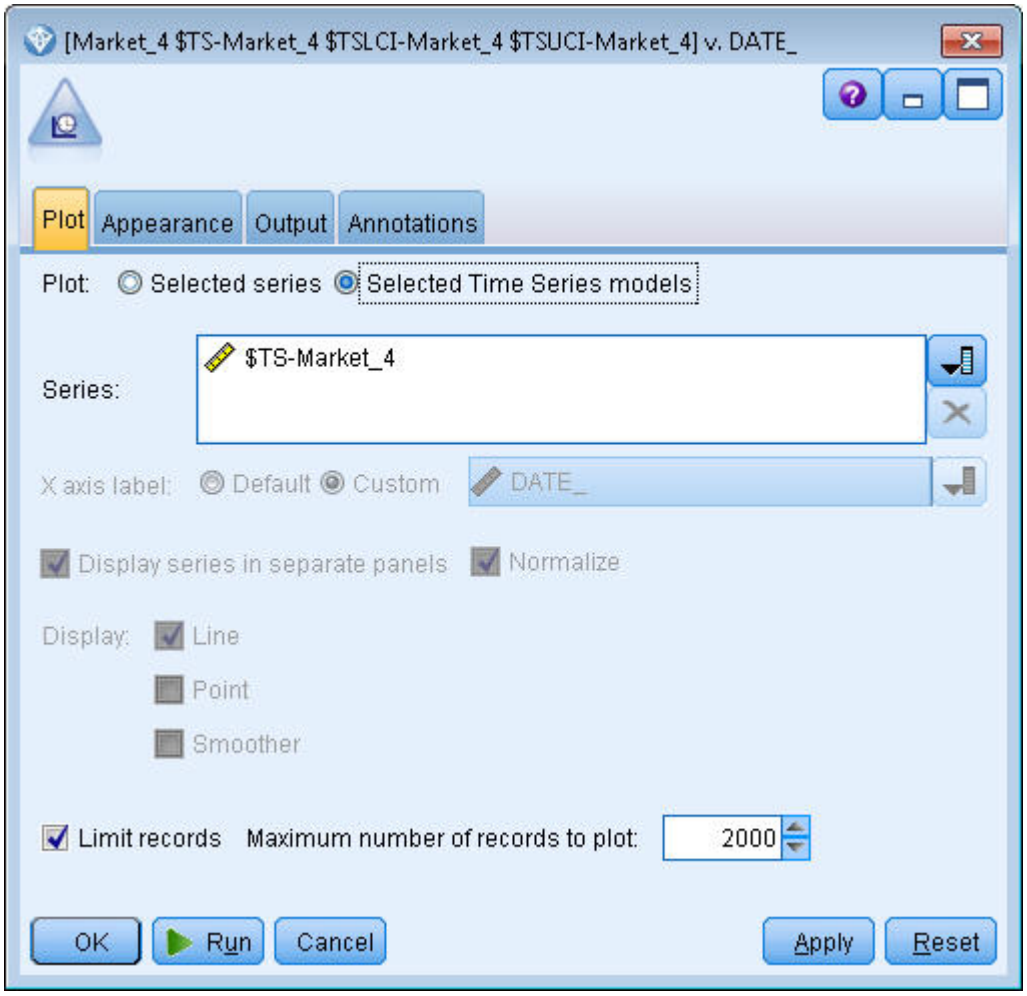


Рисунок 197. Указание полей для графика

7. Нажмите кнопку **Выполнить**.

Теперь у вас есть диаграмма, показывающая фактические продажи для рынка Market_4 вплоть до марта 2004 года вместе с прогнозом продаж (Предсказанные) и доверительным интервалом (показан голубым затенением) до июня 2004 года.

Как и в первом примере, значения прогноза на всем протяжении рассматриваемого периода времени вплотную приближаются к фактическим значениям, что вновь указывает на хорошее качество модели.

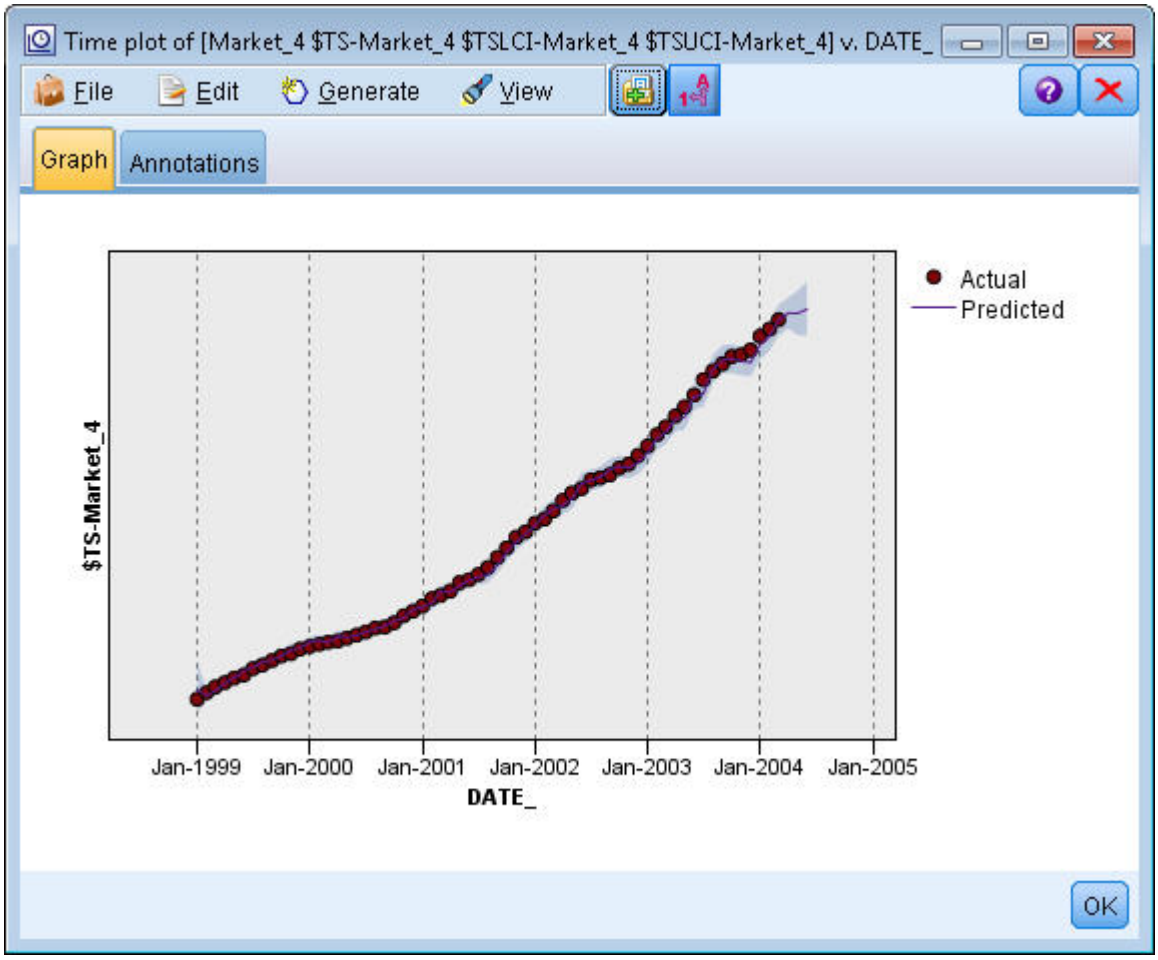


Рисунок 198. Прогноз, расширенный до июня

Итог

Вы научились применять сохраненные модели для расширения предыдущих прогнозов, когда становятся доступными новые текущие данные, без перепостроения моделей. Конечно, если есть основания предполагать, что модель изменилась, ее нужно построить заново.

Глава 15. Прогнозирование продаж по каталогу (временные ряды)

Компания, занимающаяся торговлей по каталогу, заинтересована в прогнозировании ежемесячных продаж линии мужской одежды на основании своих данных о продажах за последние 10 лет.

Этот пример использует поток *catalog_forecast.str*, в котором используется файл данных *catalog_seasfac.sav*. Эти файлы находятся в каталоге *Demos* любой установки IBM SPSS Modeler. Доступ к ним можно получить из группы программ IBM SPSS Modeler в меню Пуск Windows. Файл *catalog_forecast.str* находится в каталоге *streams*.

В прошлом примере было показано, как можно обеспечить эксперту по моделированию возможность выбрать наиболее подходящую модель для вашего временного ряда. Теперь можно посмотреть поближе на два способа, доступные при самостоятельном выборе модели - экспоненциальное сглаживание и ARIMA.

В решении о выборе подходящей модели поможет первоначальный вывод графика временного ряда. Визуальное исследование временного ряда часто может быть мощным средством, помогающим при выборе. В частности, вы должны спросить сами себя:

- Есть ли общая тенденция у этого ряда? Если есть, проявляется ли эта тенденция постоянно, или видно, что она затухает со временем?
- Показывает ли ряд сезонные изменения? Если да, выглядят ли сезонные флуктуации растущими со временем, или они кажутся постоянными в последовательные периоды времени?

Создание потока

1. Создайте новый поток и добавьте узел источника Файл статистики, указывающий на файл *catalog_seasfac.sav*.

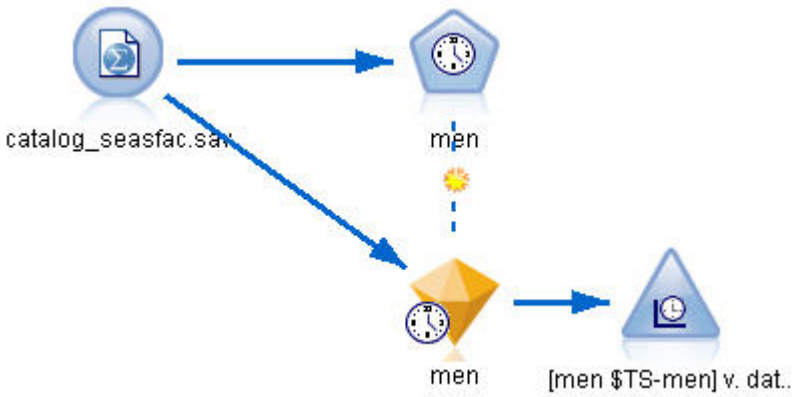


Рисунок 199. Прогноз продаж по каталогу

2. Откройте узел источника Файл IBM SPSS Statistics и выберите вкладку Типы.
3. Нажмите кнопку **Прочесть значения**, а затем кнопку **OK**.
4. Щелкните по столбцу **Роль** для поля `men` и задайте для этой роли значение **Назначение**.

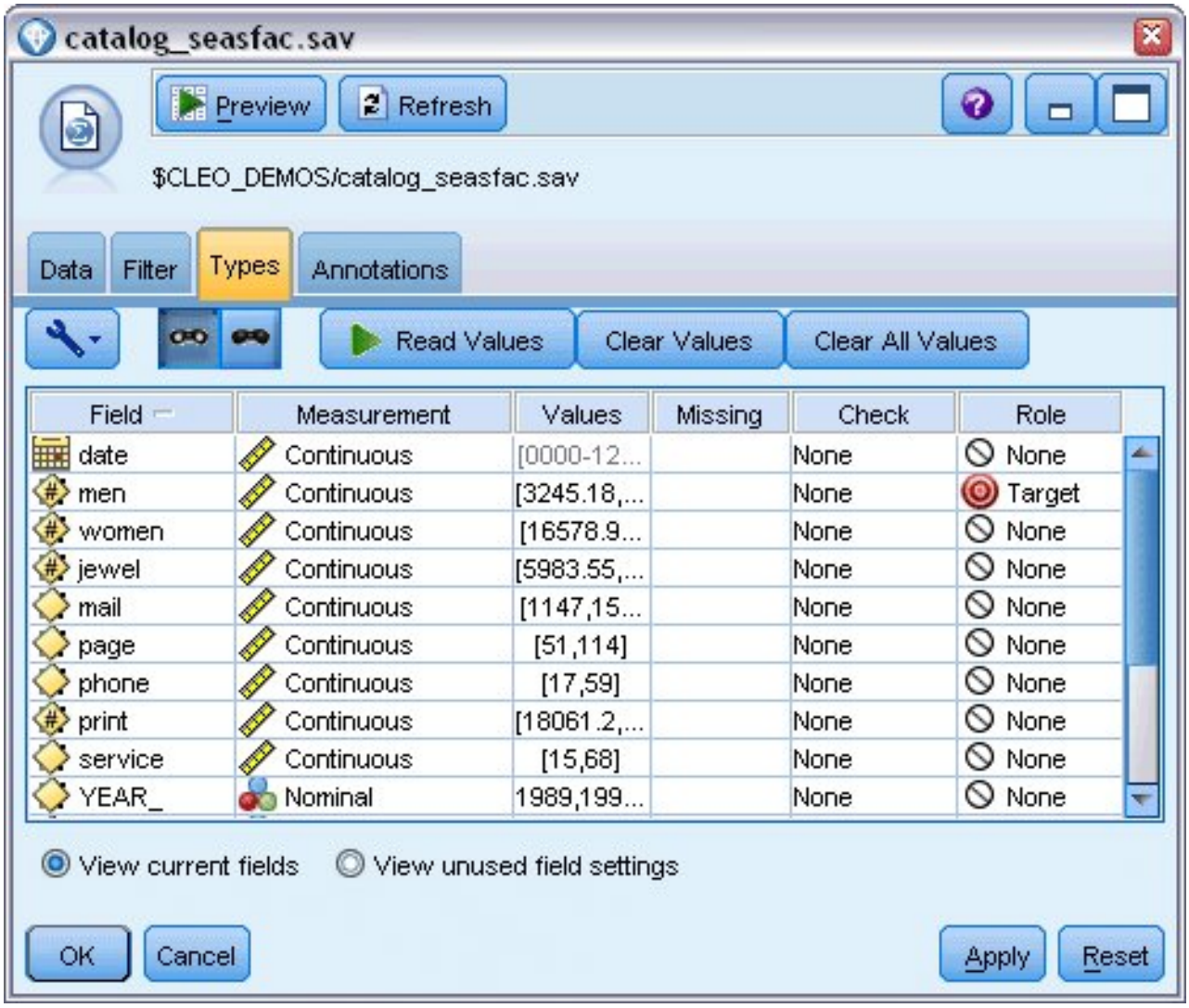


Рисунок 200. Указание поля назначения

5. Задайте для всех остальных полей роль **Нет** и нажмите кнопку **OK**.
6. Присоедините узел Диаграмма зависимости от времени к узлу источника Файл IBM SPSS Statistics.
7. Откройте узел График зависимости от времени и на вкладке График добавьте `men` в список **Ряды**.
8. Задайте для **метки оси X** значение **Пользовательская** и выберите `date`.
9. Снимите пометку с переключателя **Нормализовать**.

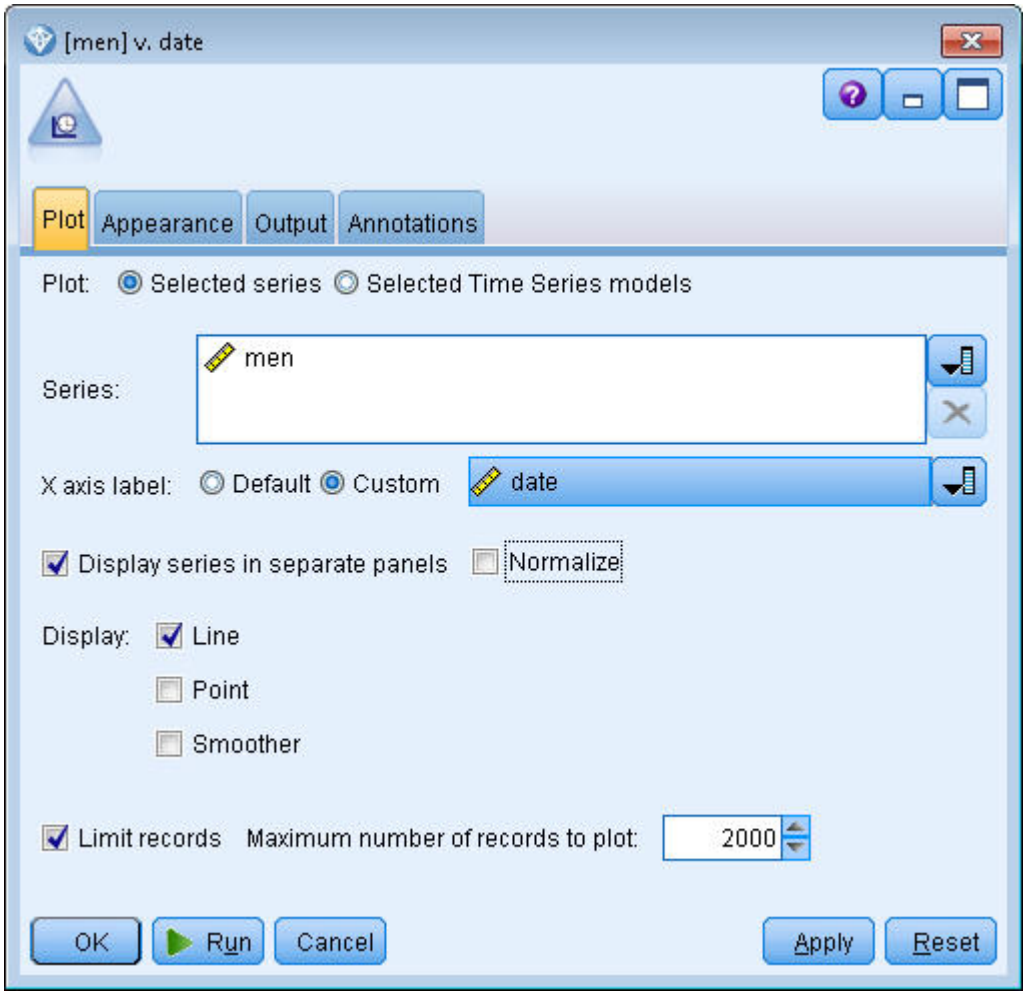


Рисунок 201. Графическое изображение временного ряда

10. Нажмите кнопку **Выполнить**.

Изучение данных

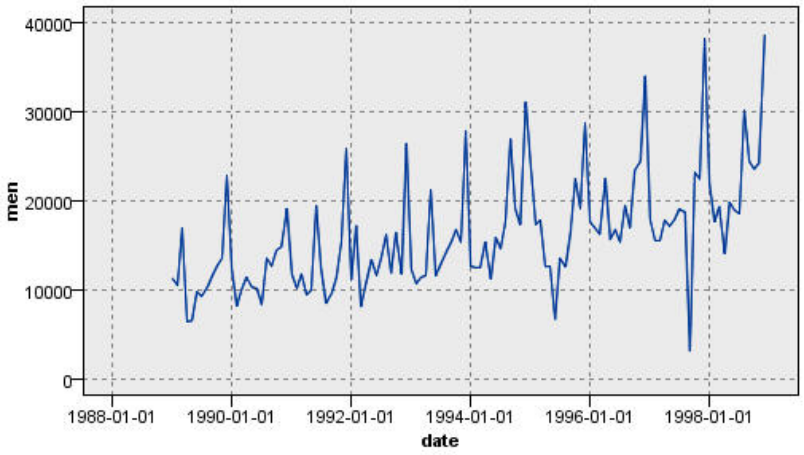


Рисунок 202. Фактические продажи мужской одежды

Этот ряд показывает общую тенденцию к росту, то есть значения в ряду увеличиваются со временем. Тенденция к повышению выглядит устойчивой, что указывает на линейный тренд.

Этот ряд демонстрирует также отчетливую сезонную структуру с пиками в декабре, как показывают вертикальные линии на диаграмме. Сезонные вариации продаж усиливаются на фоне общего тренда роста, что указывает на мультипликативную, а не аддитивную сезонность.

1. Нажмите кнопку **OK**, чтобы закрыть график.

Теперь, когда вы определили характеристики ряда, можно попытаться смоделировать его. Метод экспоненциального сглаживания полезен для прогнозирования рядов, проявляющих тренд и/или сезонность. Как мы видели, в ваших данных проявляются оба компонента.

Экспоненциальное сглаживание

Построение модели экспоненциального сглаживания с наилучшим согласием предполагает определение типа модели (нужно ли включить в модель тренд, сезонность или оба элемента) и затем получение параметров наилучшего согласия для выбранной модели.

График продаж мужской одежды в зависимости от времени предполагает модель с компонентом линейного тренда и компонентом мультипликативной сезонности. Это означает использование модели Винтера. Сначала, однако, нужно исследовать простейшую модель (без трендов и сезонности), а затем модель Хольта (включает линейный тренд без сезонности). Это поможет вам попрактиковаться в распознавании случаев, когда модель плохо согласуется с данными, что является существенным навыком для успешного построения моделей.

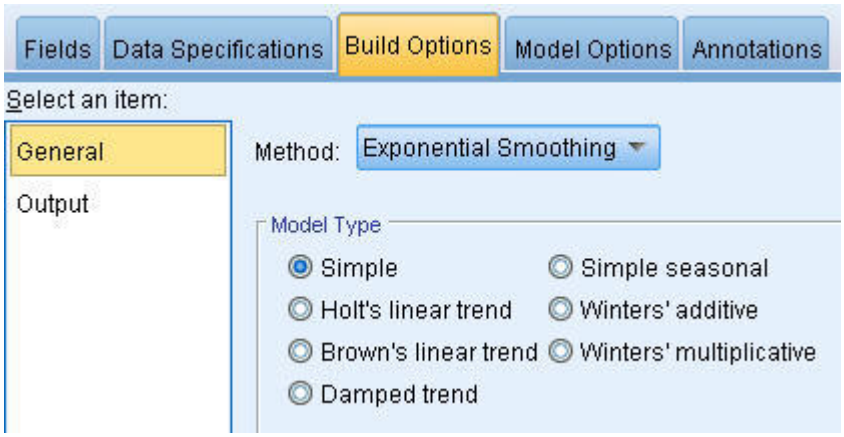


Рисунок 203. Задание экспоненциального сглаживания

Начнем с простейшей модели экспоненциального сглаживания.

1. Добавьте в поток узел Временные ряды и присоедините его к узлу источника.
2. На вкладке Спецификации данных в панели Наблюдения выберите `Дата` для поля **Дата/Время**.
3. Выберите `Месяцы` для поля **Интервал времени**.

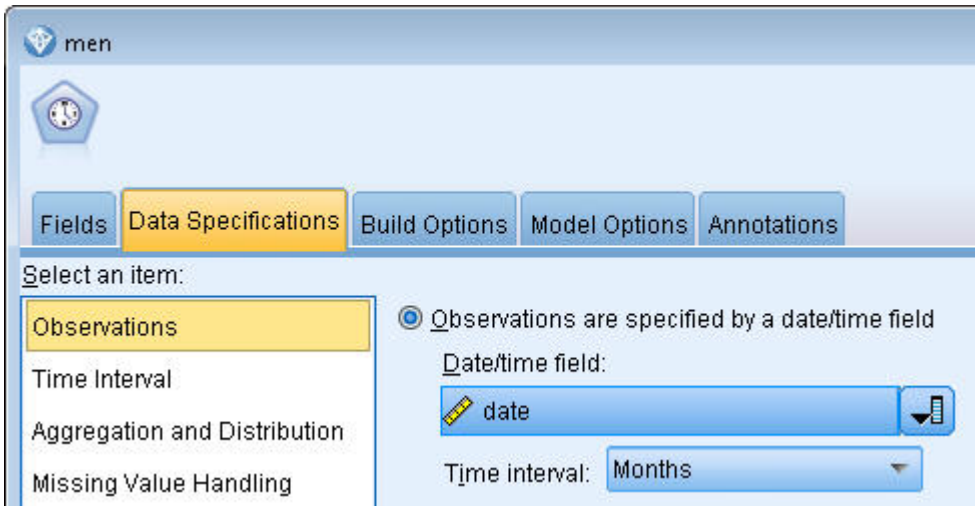


Рисунок 204. Задание временного интервала

4. На вкладке Опции построения в панели Общие задайте для **Метод** значение **Экспоненциальное сглаживание**.
5. Задайте для **Тип модели** значение **Простая**.

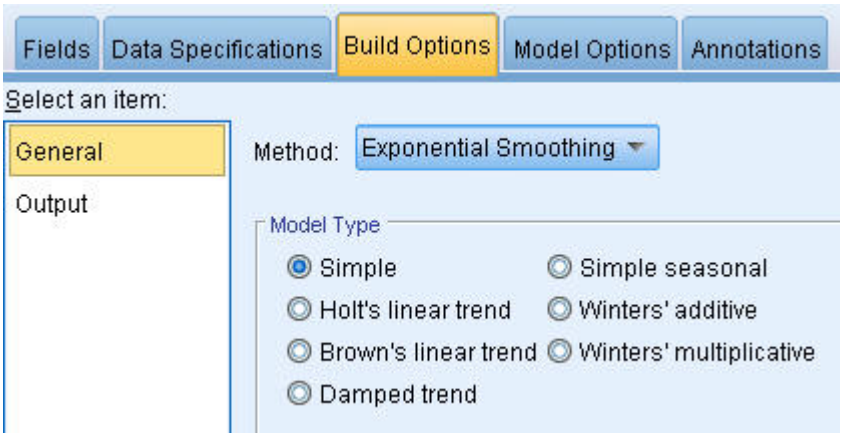


Рисунок 205. Задание метода построения модели

6. Нажмите кнопку **Выполнить**, чтобы создать слепок модели.

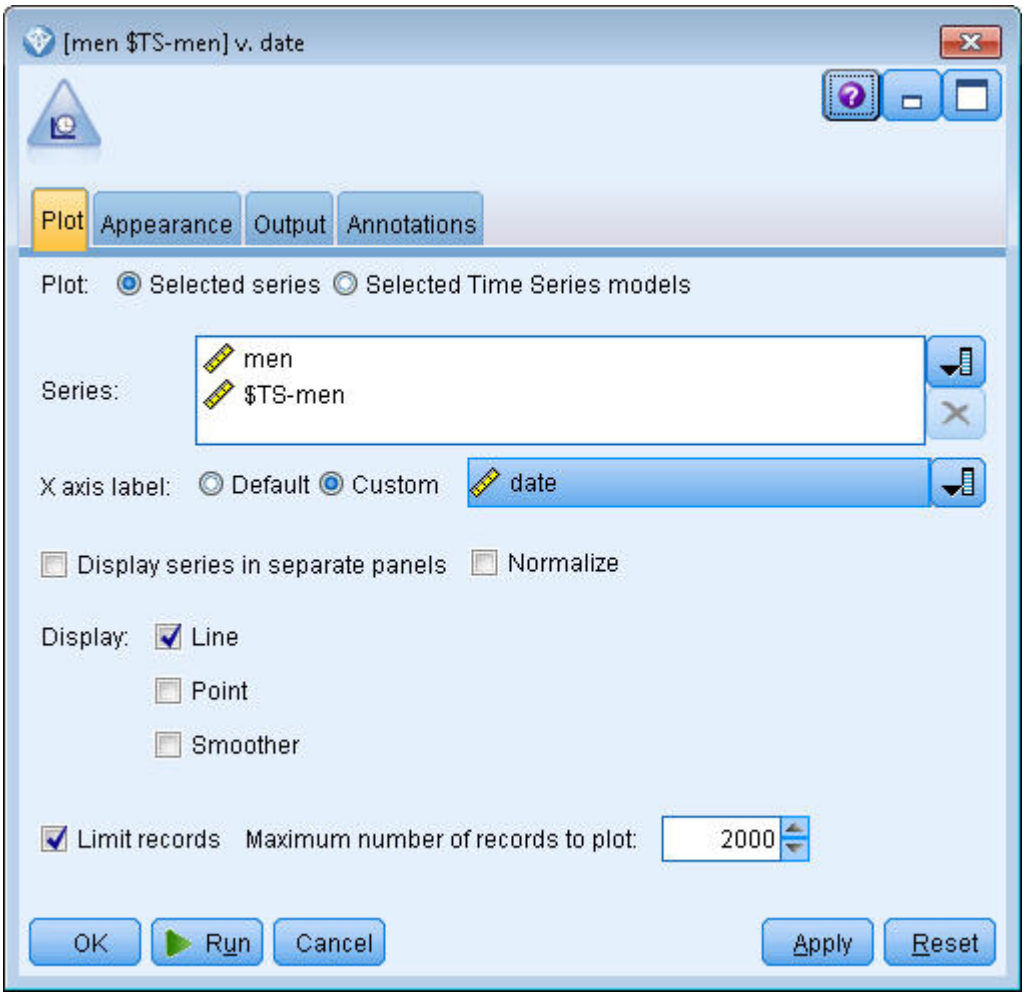


Рисунок 206. Графическое построение модели временного ряда

7. Присоедините к слепку модели узел График зависимости от времени.
8. На вкладке **График** добавьте поля `men` и `$TS-men` в список **Ряды**.
9. Задайте для **метка оси X** значение **Пользовательская** и выберите `date`.

10. Выключите переключатели **Показывать ряды в отдельных панелях** и **Нормализовать**.
11. Нажмите кнопку **Выполнить**.

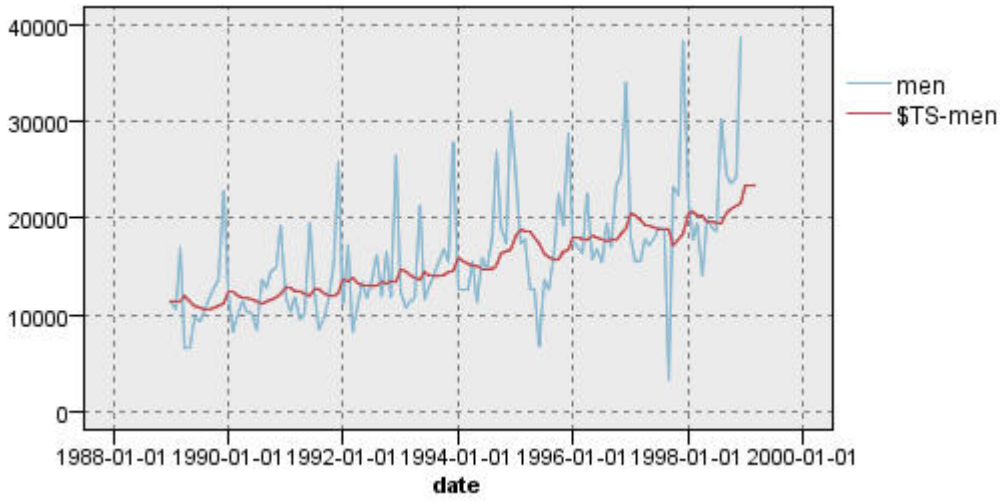


Рисунок 207. Модель простого экспоненциального сглаживания

График **men** представляет фактические данные, тогда как **$TS-men** отражает модель временного ряда.

Хотя простая модель действительно демонстрирует постепенный (и довольно сильный) положительный тренд, она никак не учитывает сезонность. Эту модель можно спокойно отбросить.

12. Нажмите кнопку **OK**, чтобы закрыть окно графика зависимости от времени.

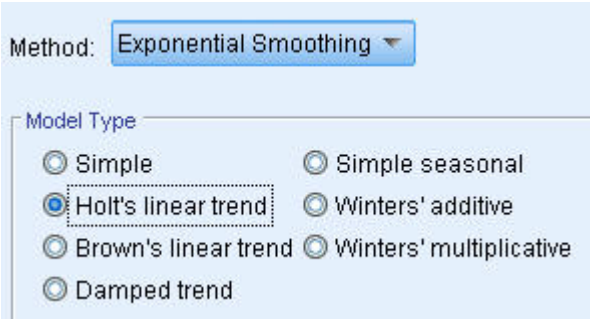


Рисунок 208. Выбор модели Хольта

Попробуем применить линейную модель Хольта. Эта модель способна лучше учитывать тенденции, хотя она также едва ли чувствительна к сезонности.

13. Повторно откройте узел Временной ряд.
14. На вкладке Опции построения в панели Общие оставьте выбранным **Экспоненциальное сглаживание** как **Метод** и выберите **Линейный тренд Хольта** как **Тип модели**.
15. Нажмите кнопку **Выполнить**, чтобы заново создать слепок модели.
16. Заново откройте узел графика временной зависимости и щелкните по **Выполнить**.

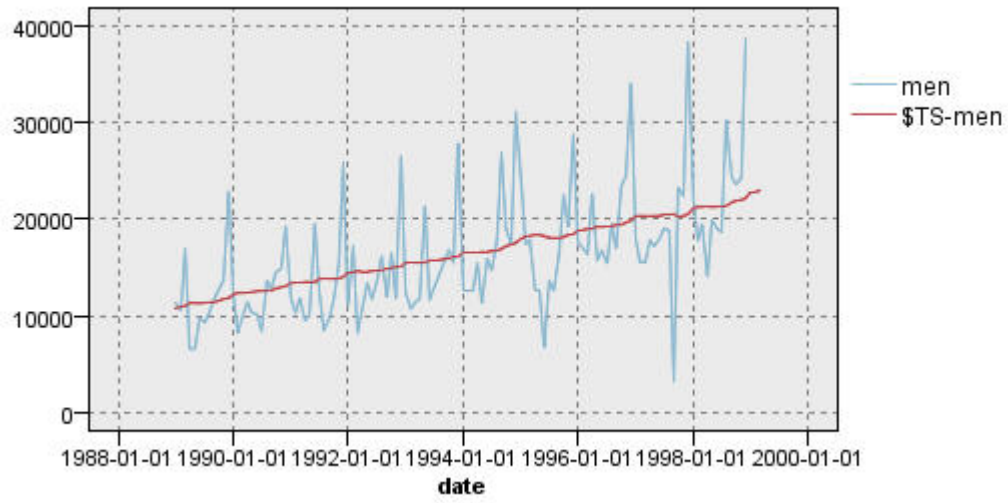


Рисунок 209. Модель линейного тренда Хольта

Модель Хольта демонстрирует более гладкий положительный тренд, чем простая модель, но она также не учитывает сезонность и поэтому может быть отклонена, как и предыдущая.

17. Закройте окно графика зависимости от времени.

    Вспомним, что исходный график продаж мужской одежды в зависимости от времени предполагает модель, объединяющую линейный тренд с мультипликативной сезонностью. Поэтому более подходящей в данном случае может быть модель Винтера.

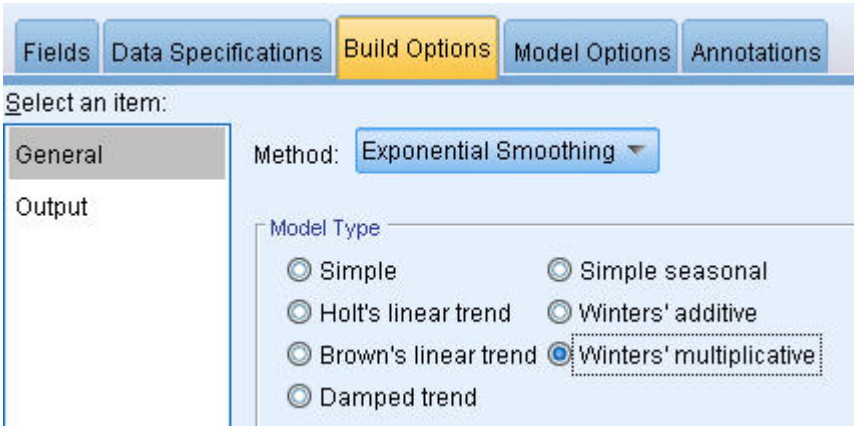


Рисунок 210. Выбор модели Винтера

18. Повторно откройте узел Временной ряд.
19. На вкладке Опции построения в панели Общие оставьте выбранным **Экспоненциальное сглаживание** как **Метод** и выберите **Мультипликативная Винтера** как **Тип модели**.
20. Нажмите кнопку **Выполнить**, чтобы заново создать слепок модели.
21. Откройте узел графика временной зависимости и щелкните по **Выполнить**.

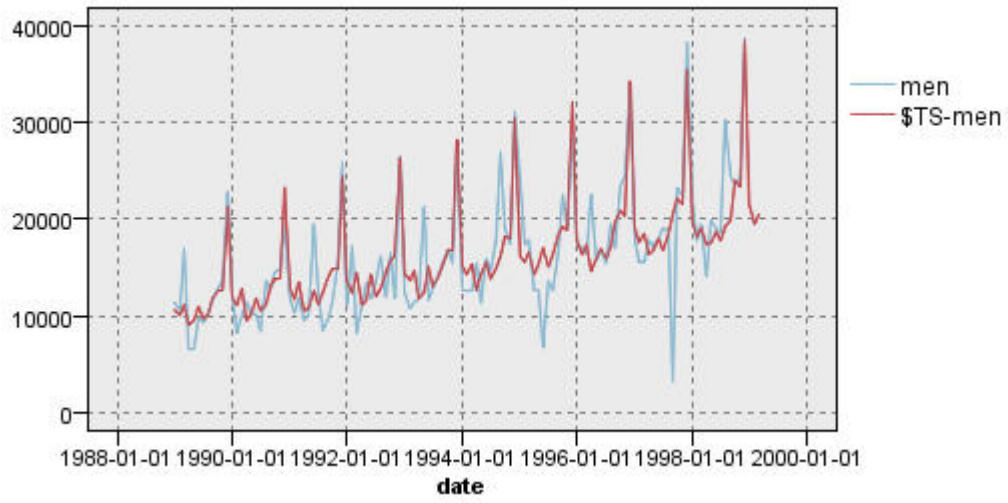


Рисунок 211. Мультипликативная модель Винтера

Этот результат выглядит лучше: модель отражает и тренд, и сезонность данных.

Набор данных охватывает десятилетний период и включает 10 ежегодных сезонных пиков, приходящихся на декабрь. Эти 10 пиков в предсказанных моделью результатах хорошо соответствуют 10 годовым пикам в реальных данных.

Однако в этих результатах недоучитываются ограничения процедуры экспоненциального сглаживания. Сочетание положительных и отрицательных пиков дает существенную структуру, которая не отражается в модели.

Если вас интересует прежде всего моделирование долговременных трендов с сезонными вариациями, метод экспоненциального сглаживания хорошо подойдет для этих целей. Для моделирования более сложных структур, подобных данной, может потребоваться использование процедуры ARIMA.

АРПСС

С помощью процедуры ARIMA можно создать авторегрессионную интегрированную модель скользящего среднего, пригодную для тонкой настройки моделирования временных рядов. Модели ARIMA предоставляют более сложные методы для моделирования трендовых и сезонных компонентов, чем модели экспоненциального сглаживания, и позволяют (в качестве дополнительного преимущества) включать в модель независимые (предикторные) переменные.

Продолжая пример с рассылочной фирмой, которой нужно разработать прогностическую модель, мы рассмотрели, как эта фирма собирает данные по месячным продажам мужской одежды, наряду с несколькими рядами данных, с помощью которых можно объяснить часть вариации в объеме продаж. Среди возможных предикторов - число рассылаемых по почте каталогов, число страниц в каталоге, число телефонных линий, открытых для заказа, расходы на печатную рекламу, а также число представителей по обслуживанию заказчиков.

Полезны ли какие-либо из этих предикторов для прогнозирования? Действительно ли модель с предикторами лучше модели без предикторов? С помощью процедуры ARIMA можно создать прогностическую модель с предикторами и оценить, есть ли существенная разница между прогностической способностью этой модели и модели экспоненциального сглаживания без предикторов.

С помощью метода ARIMA можно выполнить тонкую настройку модели, задав порядок авторегрессии, дифференцировки и скользящего среднего, а также сезонные эквиваленты этих компонентов. Определение

оптимальных значений для этих компонентов вручную может отнять много времени с широким использованием метода проб и ошибок, поэтому в данном примере Expert Modeler сам выберет для нас модель ARIMA.

Мы попытаемся построить модель лучшего качества, рассматривая некоторые из других переменных в наборе данных как переменные-предикторы. Наиболее полезными переменными для включения в модель в качестве предикторов представляются число разосланных по почте каталогов (`mail`), число страниц в каталоге, (`page`), число телефонных линий, открытых для заказа (`phone`), расходы на печатную рекламу (`print`) и число представителей по обслуживанию клиентов (`service`).

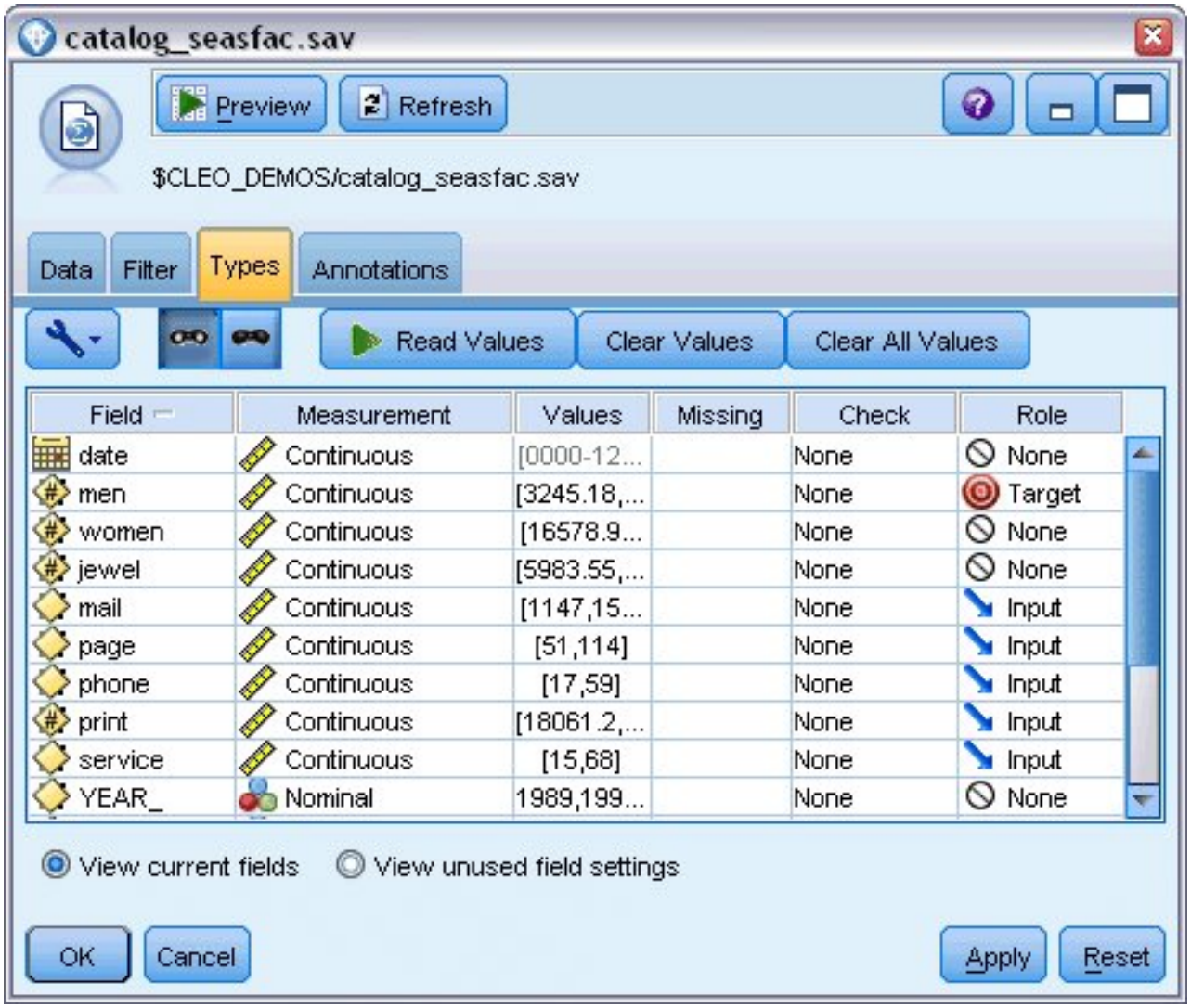


Рисунок 212. Задание полей предиктора

1. Откройте узел источника файла IBM SPSS Statistics.
2. На вкладке Типы задайте **Роль** для переменных `mail`, `page`, `phone`, `print` и `service`, указав значение **Входные**.
3. Убедитесь, что для роли `men` задано значение **Назначение** и что для всех остальных полей задано **Нет**.
4. Щелкните по **OK**.
5. Откройте узел Временной ряд.
6. На вкладке Опции построения на панели Общие задайте для **Метод** значение **Expert Modeler**.

7. Выберите опцию **Только модели ARIMA** и убедитесь, что включена опция **Включать в рассмотрение модели сезонности**.

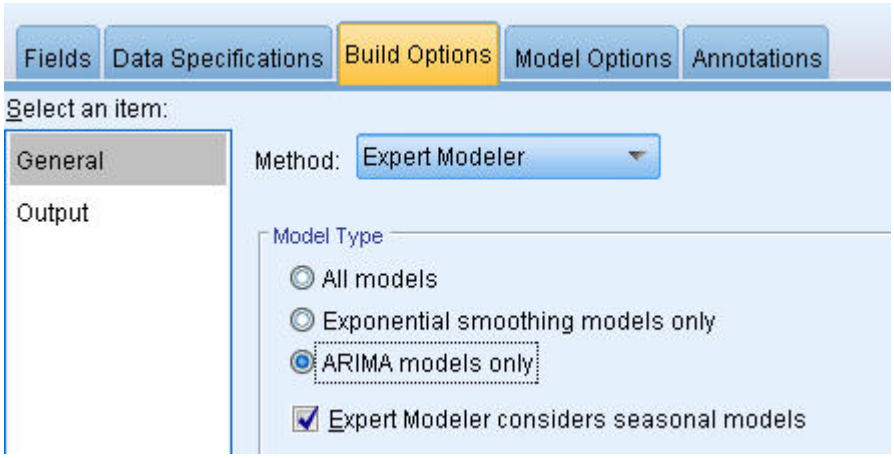


Рисунок 213. Как выбрать только модели ARIMA

8. Нажмите кнопку **Выполнить**, чтобы заново создать слепок модели.
9. Откройте слепок модели.

   В левом столбце вкладки Вывод выберите **Информация о модели**. Обратите внимание, что Expert Modeler выбрал только два из пяти указанных предикторов в качестве значимых для модели.

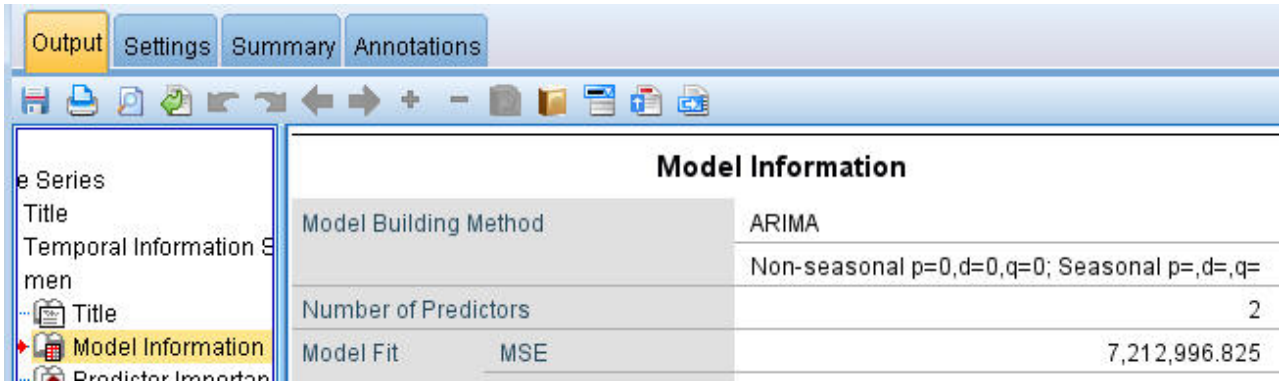


Рисунок 214. Expert Modeler выбирает два предиктора

10. Нажмите кнопку **OK**, чтобы закрыть слепок модели.
11. Откройте узел графика временной зависимости и щелкните по **Выполнить**.

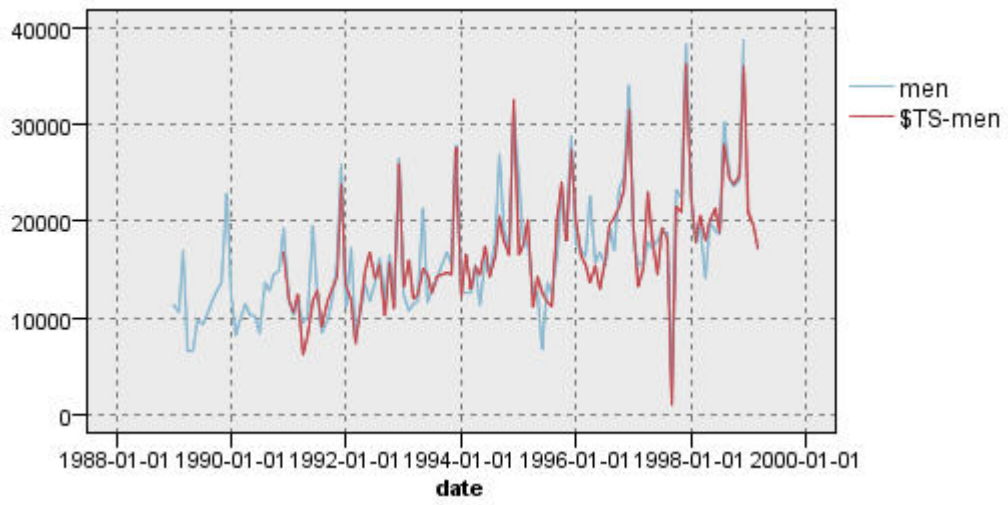


Рисунок 215. Модель ARIMA с заданными предикторами

Эта модель лучше предыдущей, поскольку в ней захватываются также большие отрицательные пики при сохранении наилучшего согласия.

Мы могли бы попытаться совершенствовать модель и дальше, но любые улучшения с этого момента, вероятно, будут минимальны. Поскольку мы установили, что модель ARIMA с предикторами предпочтительнее, будем использовать далее только что построенную модель. Для целей данного примера спрогнозируем продажи в наступающем году.

12. Нажмите кнопку **OK**, чтобы закрыть окно графика зависимости от времени.
13. Откройте узел Временной ряд и выберите вкладку Опции модели.
14. Включите переключатель **Распространить записи на будущее** и задайте для него значение 12.
15. Включите переключатель **Вычислить будущие значения входных данных**.
16. Нажмите кнопку **Выполнить**, чтобы заново создать слепок модели.
17. Откройте узел графика временной зависимости и щелкните по **Выполнить**.

Прогноз на 1999 год выглядит хорошо. Ожидается возврат к нормальным уровням продаж после декабрьского спада и устойчивая тенденция к их повышению во второй половине года, в целом объем продаж будет выше, чем за предыдущий год.

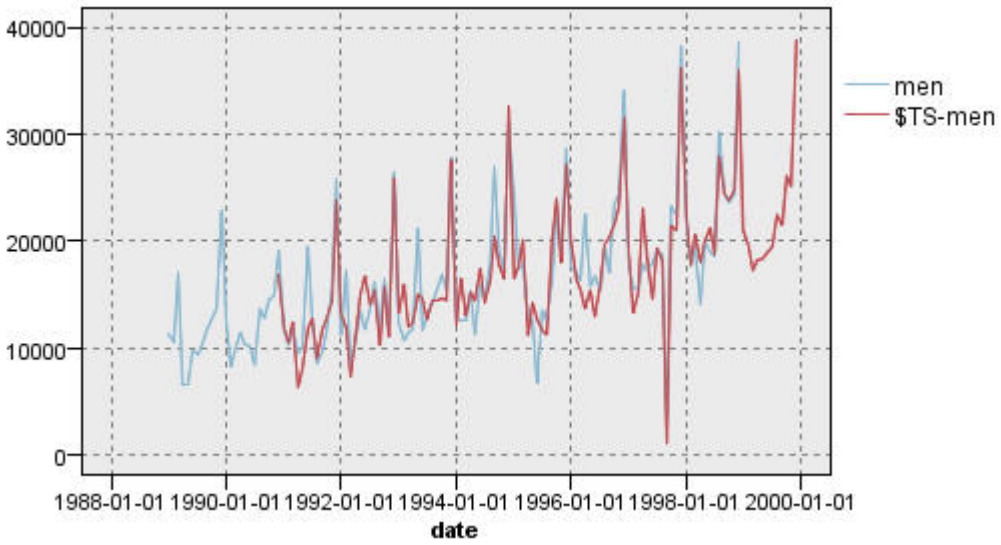


Рисунок 216. Прогноз продаж, расширенный до 12 месяцев

Итог

Вы успешно смоделировали сложный временной ряд, включающий в себя не только общую тенденцию роста, но и сезонные и другие изменения. Вы увидели также, как методом проб и ошибок можно всё ближе подходить к точной модели, которую вы затем используете для прогнозирования будущих продаж.

На практике вам может потребоваться повторно применить модель по мере обновления фактических данных о продажах, например, каждый месяц или каждый квартал, и генерировать измененные прогнозы. Дополнительную информацию смотрите в разделе “Повторное применение модели временных рядов” на стр. 171.

Глава 16. Внесение предложений покупателям (самообучение)

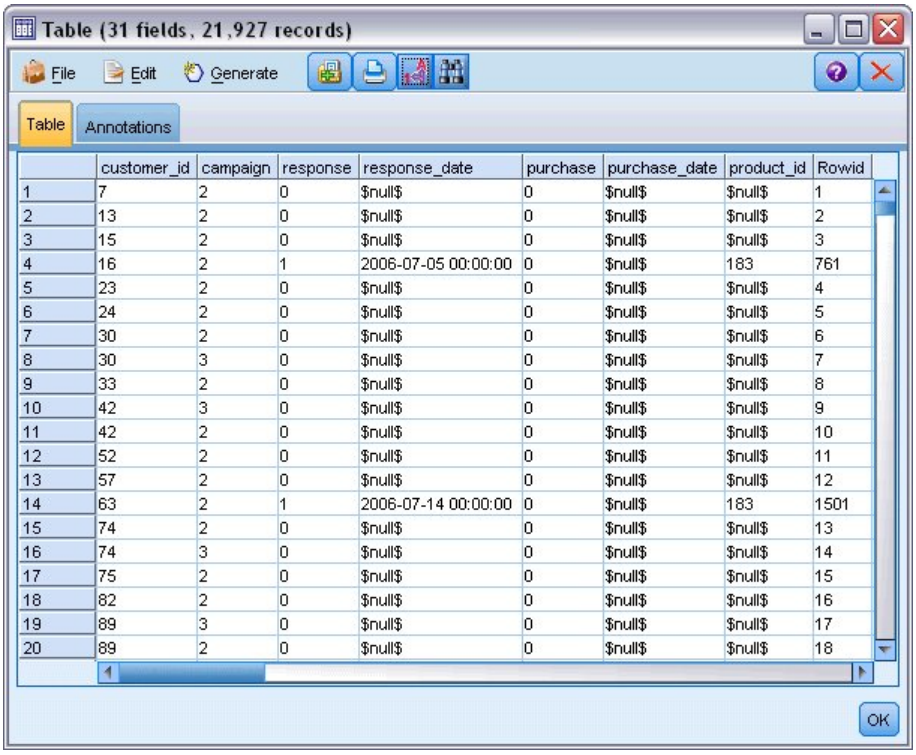
Узел модели откликов самообучения (Self-Learning Response Model, SLRM) генерирует и включает модель, позволяющую предсказать, какие предложения будут наиболее походить клиентам, а также предсказать вероятность принятия этих предложений. Модели такого сорта наиболее выгодны при управлении взаимосвязей с клиентами, например, в прикладных программах маркетинга и для колл-центров.

Этот пример основывается на работе вымышленного банка. Маркетинговый отдел хочет достичь более выгодных результатов в будущих рекламных кампаниях, подбирая правильные предложения финансовых услуг для каждого из клиентов. В частности, в этом примере используется модель откликов самообучения для идентификации характеристик клиентов, которые наиболее вероятно откликнутся положительно, на основе предыдущих рекламных кампаний и для продвижения наилучшего текущего предложения на основании результатов модели.

В этом примере используется поток *pm_selflearn.str*, ссылающийся на файлы данных *pm_customer_train1.sav*, *pm_customer_train2.sav* и *pm_customer_train3.sav*. Эти файлы находятся в папке *Demos*, внутри папки, где установлен IBM SPSS Modeler. Доступ к ним можно получить из группы программ IBM SPSS Modeler в меню Пуск Windows. Файл *pm_selflearn.str* находится в папке *streams*.

Существующие данные

У компании есть хронологические данные, отслеживающие предложения, сделанные для клиентам в прошлых кампаниях, а также данные об отклике. Эти данные содержат также демографическую и финансовую информацию, которая может использоваться, чтобы предсказать показатели откликов для разных клиентов.



	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

Рисунок 217. Ответы на предыдущие предложения

Построение потока

1. Добавьте узел источников файлов статистики, указывающий на файл *pm_customer_train1.sav*, расположенный в папке *Demos* вашего каталога установки IBM SPSS Modeler.

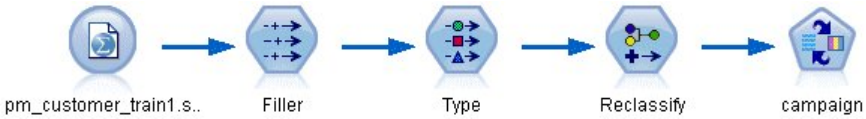


Рисунок 218. Поток примера SLRM

2. Добавьте узел заполнения и выберите для `компании` опцию Заполнить в поле.
3. Выберите тип замены **Всегда**.
4. В текстовом поле Заменить на введите `to_string(campaign)` и нажмите кнопку **OK**.

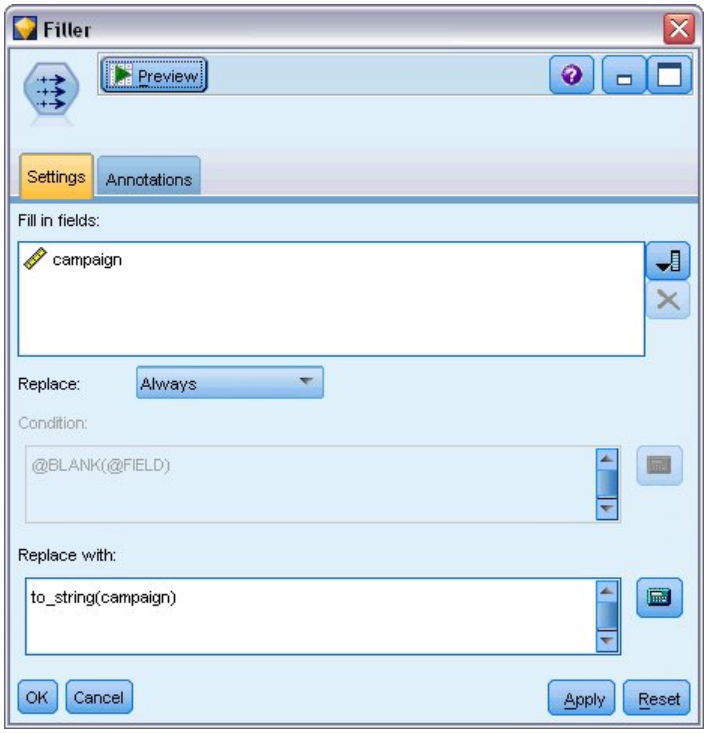


Рисунок 219. Вывод поля campaign

5. Добавьте узел Тип и задайте *Роли* значение **Нет** для полей *customer_id*, *response_date*, *purchase_date*, *product_id*, *Rowid* и *X_random* (ID клиента, дата ответа, дата покупки, ID продукта, ID строки и X_random).

Рисунок 220. Изменение параметров узла Тип

6. Задайте *Роли* значение **Назначение** для полей *campaign* и *response* (кампания и отклик). Это поля, на которых вы хотите основывать свои предсказания.

   Задайте для **Измерения** значение **Флаг** для поля *response*.

7. Нажмите кнопку **Прочесть значения**, а затем кнопку **OK**.

   Так как данные поля кампании показывают список чисел (1, 2, 3 и 4), вы можете переклассифицировать эти поля, чтобы их названия были более осмысленными.
8. Добавьте узел Переклассификация к узлу Тип.
9. В поле **Переклассифицировать в** выберите **Существующее поле**.
10. В списке **Переклассифицировать поле** выберите **campaign**.
11. Нажмите кнопку **Получить**; значения для кампании будут добавлены в столбец *Исходное значение*.
12. В столбце *Новое значение* введите следующие имена кампаний в первых четырех строках:
    - **Ипотека**
    - **Кредит на машину**
    - **Сбережения**
    - **Пенсия**
13. Щелкните по **OK**.

Рисунок 221. Переклассификация имен кампаний

14. Присоедините узел моделирования SLRM к узлу Переклассификация. На вкладке Поля выберите **campaign** (кампания) для поля назначения и **response** (отклик) для поля отклика назначения.

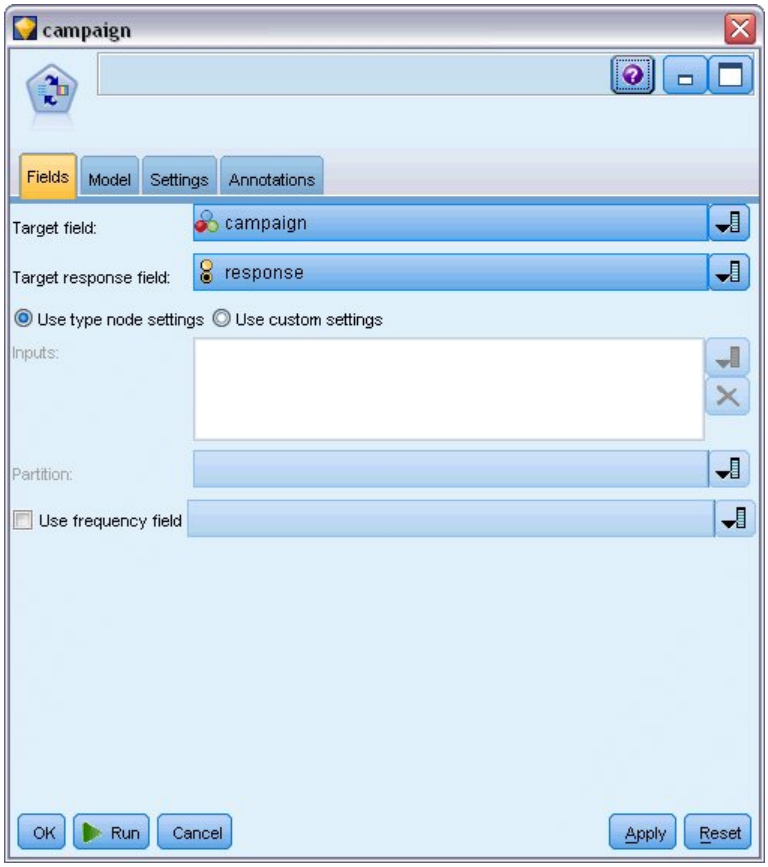


Рисунок 222. Выбор цели и целевого ответа

15. На вкладке Параметры в поле Максимальное число предсказаний в секунду уменьшите значение до 2.

    Это означает, что для каждого клиента будет идентифицировано два предложения с наибольшей вероятностью принятия.
16. Убедитесь, что выбрана опция **Учитывать надежность модели**, и нажмите кнопку **Выполнить**.

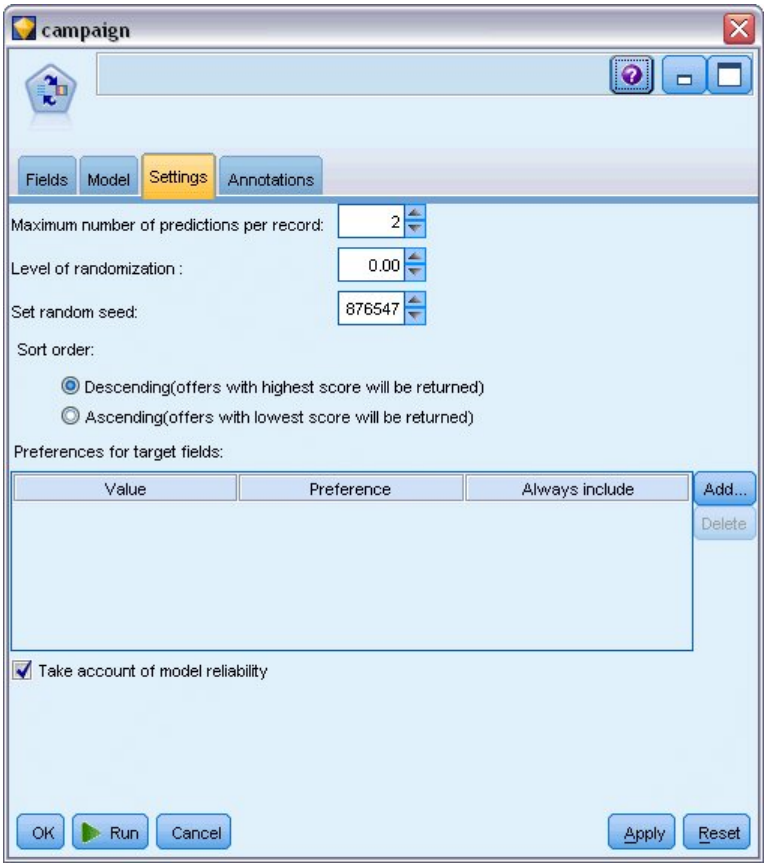


Рисунок 223. Параметры узла SLRM

Просмотр модели

1. Откройте слепок модели. Первоначально на вкладке Модель показывается оценка точности предсказаний для каждого предложения и относительная важность каждого предиктора для оценки модели.

   Для вывода корреляции каждого предиктора с переменной назначения выберите опцию **Связь с откликом** в списке **Представление** на правой панели.

2. Для переключения между каждым из четырех предложений, для которых существуют предсказания, выберите нужное предложение в списке **Представление** на левой панели.

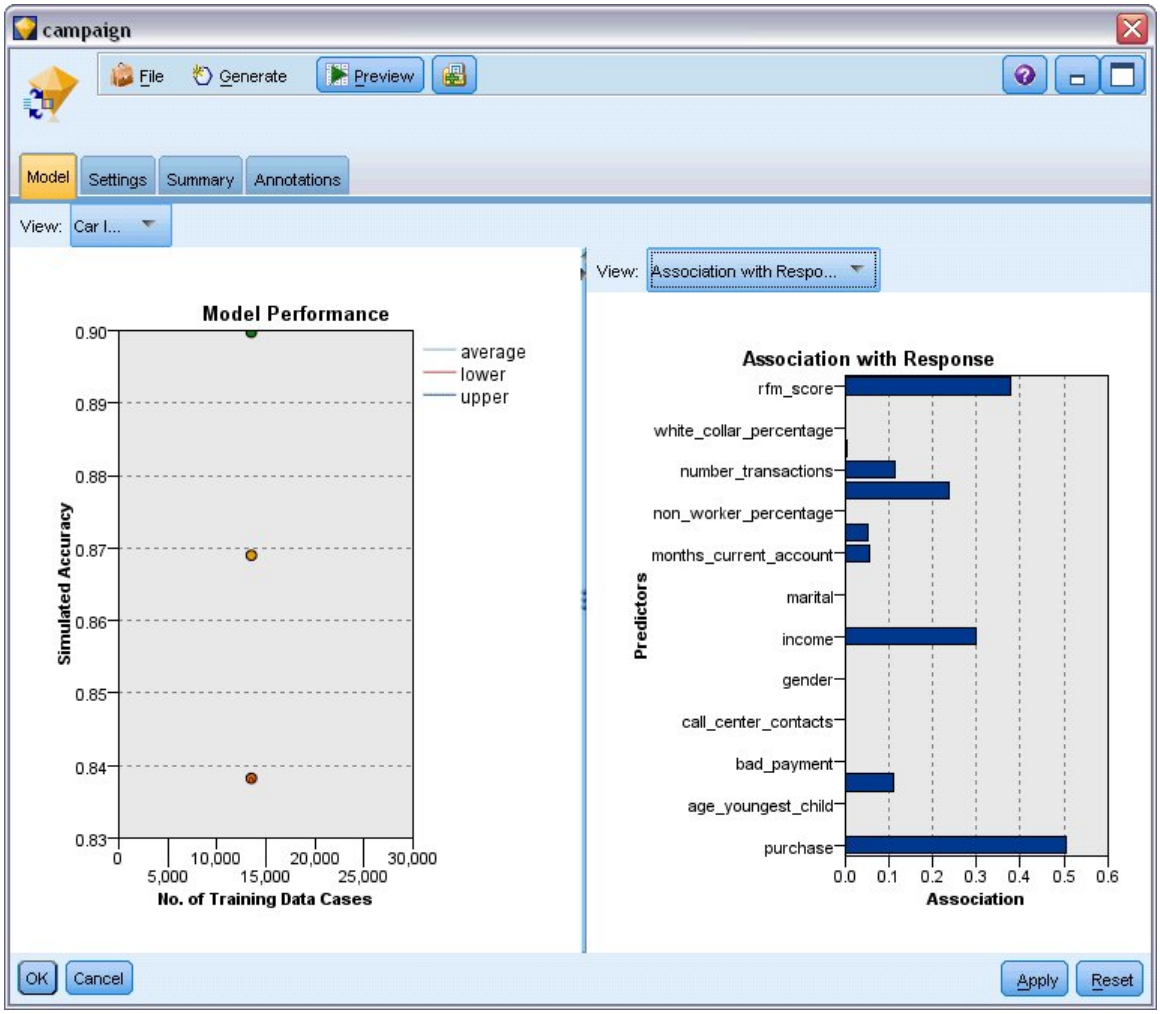


Рисунок 224. Слепок модели SLRM

3. Закройте окно слепка модели.
4. На холсте потока отсоедините узел источников файлов IBM SPSS Statistics, указывающий на файл *pm_customer_train1.sav*.
5. Добавьте узел источников файлов статистики, указывающий на файл *pm_customer_train2.sav*, расположенный в папке *Demos* вашего каталога установки IBM SPSS Modeler, и соедините его с узлом заполнения.

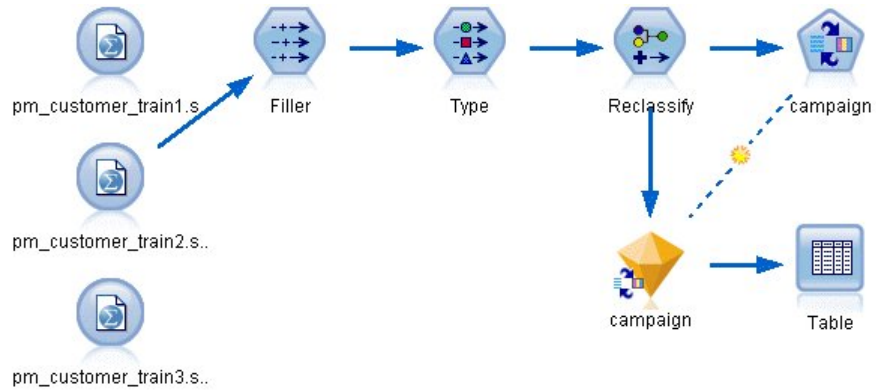


Рисунок 225. Присоединение второго источника данных к потоку SLRM

6. На вкладке Модель узла SLRM выберите опцию **Продолжить обучение существующей модели**.

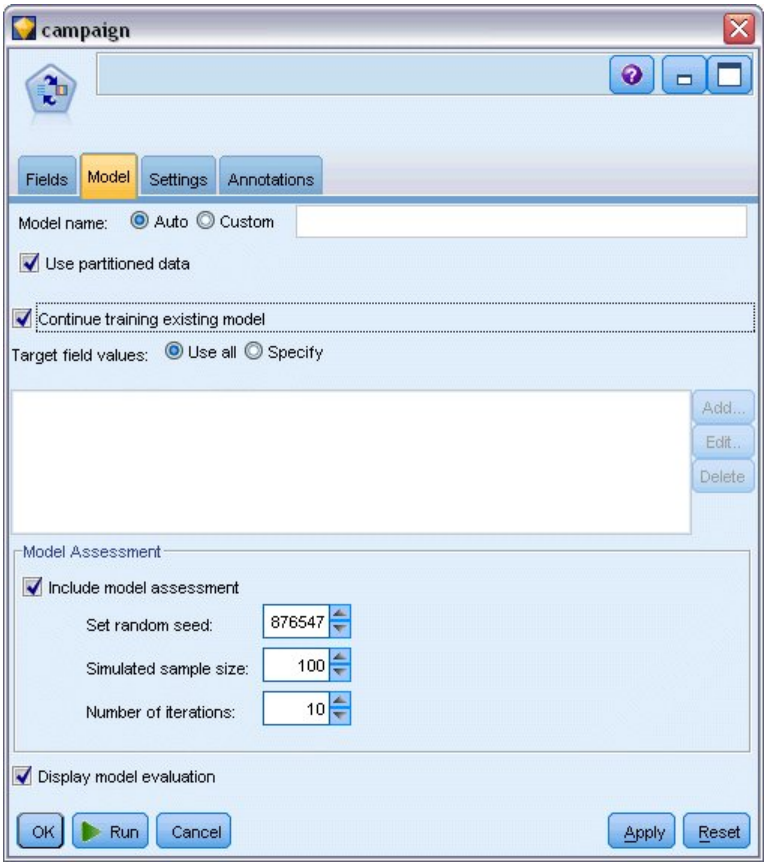


Рисунок 226. Продолжение обучения модели

7. Нажмите кнопку **Выполнить**, чтобы заново создать слепок модели. Чтобы посмотреть сведения о нем, дважды щелкните по слепку на холсте.

   Теперь на вкладке Модель будут показаны скорректированные оценки точности предсказаний для каждого предложения.

8. Добавьте узел источников файлов статистики, указывающий на файл *pm_customer_train3.sav*, расположенный в папке *Demos* вашего каталога установки IBM SPSS Modeler, и соедините его с узлом заполнения.

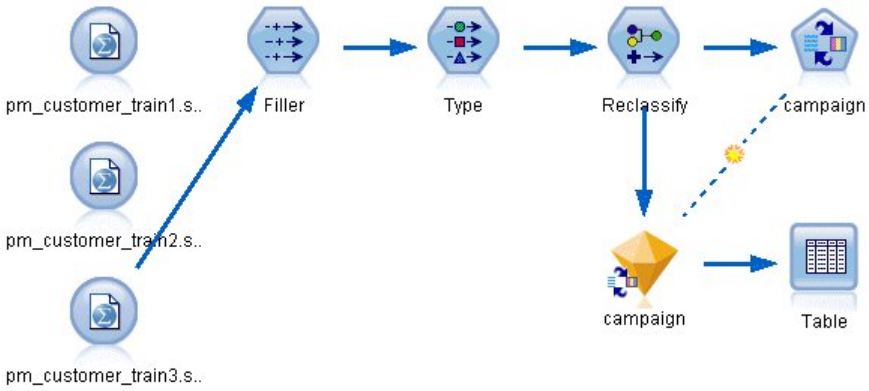


Рисунок 227. Присоединение третьего источника данных к потоку SLRM

9. Нажмите кнопку **Выполнить**, чтобы еще раз создать слепок модели. Чтобы посмотреть сведения о нем, дважды щелкните по слепку на холсте.
10. Теперь на вкладке Модель будут показаны окончательные оценки точности предсказаний для каждого предложения.

    Можно увидеть, что средняя точность после использования дополнительных источников данных немного упала (с 86,9% до 85,4%); однако размер этих флуктуаций минимален и может быть приписан малым отклонениям в доступных данных.

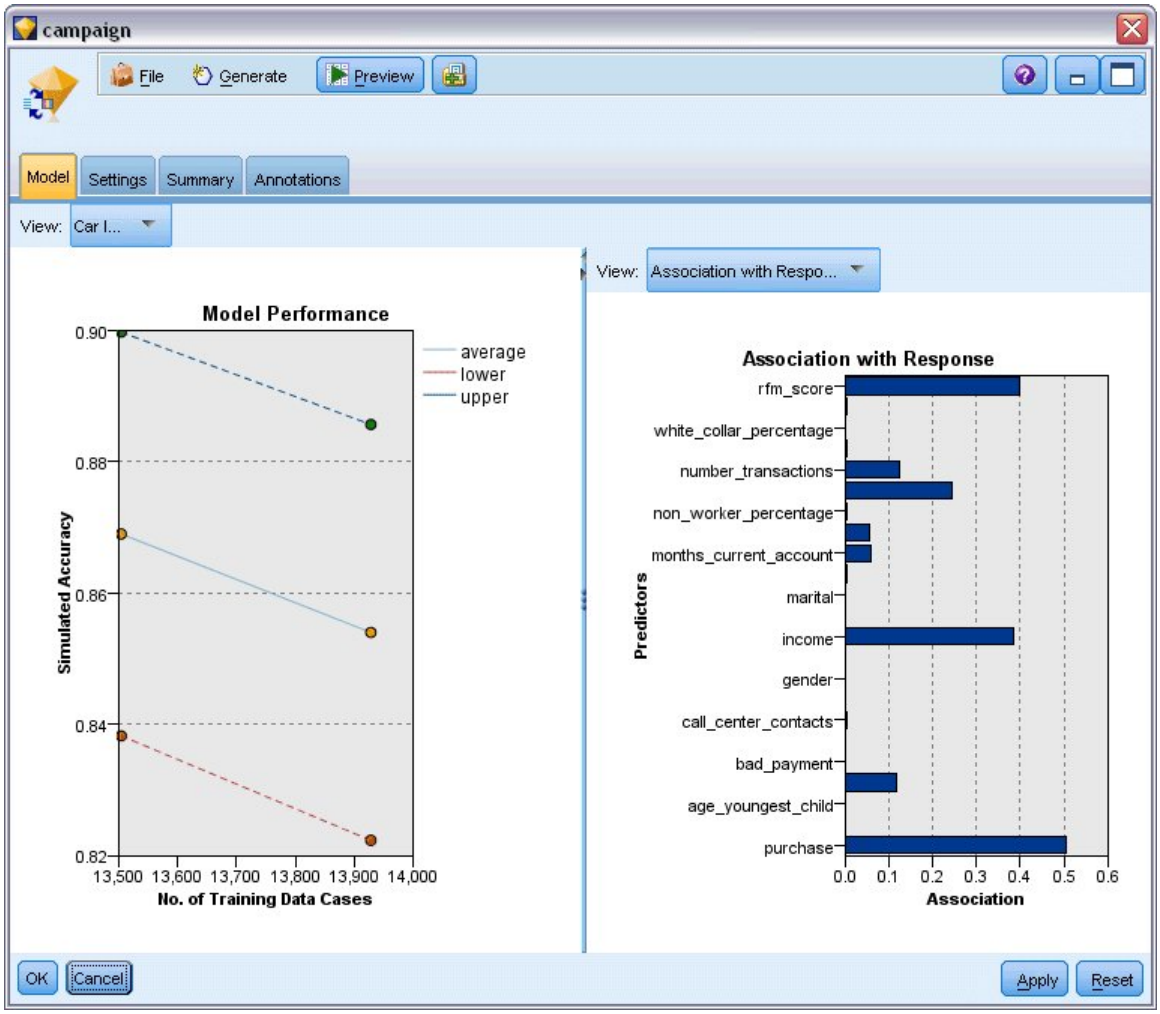
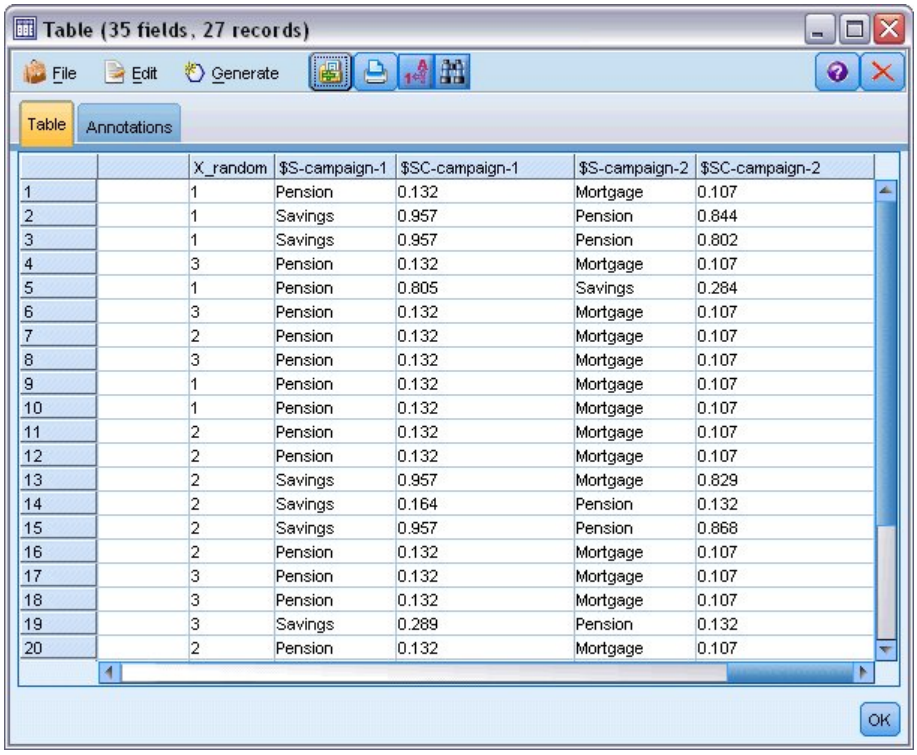


Рисунок 228. Обновленный слепок модели SLRM

11. Присоедините узел Таблица к последней (третьей) сгенерированной модели и выполните этот узел.
12. Прокрутите таблицу вправо. Предсказания покажут, какие предложения наиболее вероятно примут клиенты, и доверительную вероятность этого принятия в зависимости от подробностей каждого клиента.

Например, в первой строке таблицы показано, что существует вероятность только в 13,2% (отмечено значением 0.132 в столбце *$SC-campaign-1*), что клиент, ранее воспользовавшийся кредитом на автомобиль, примет предложение пенсионного счета, если оно будет сделано. Однако во второй и третьей строке показаны еще два клиента, также бравшие кредит на машину; в этих случаях есть доверительная вероятность 95,7%, что эти клиенты и другие клиенты с аналогичной историей откроют сберегательный счет в случае предложения, а также с вероятностью более 80% они примут предложение открыть пенсионный счет.



		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

Рисунок 229. Вывод модели - предсказанные предложения и достоверности

Объяснение математических основ методов моделирования, используемых в IBM SPSS Modeler, смотрите в публикации *IBM SPSS Modeler: Руководство по алгоритмам* (файл PDF, который входит в состав скачанного продукта.

Учтите, что эти результаты основаны только на обучающих данных. Чтобы оценить, насколько хорошо модель обобщается на другие данные реального мира, рекомендуется применить узел Разбиение, который будет содержать поднабор записей в целях проверки.

Глава 17. Предсказание неплательщиков по кредитам (байесовская сеть)

Байесовские сети позволяют построить вероятностную модель, которая, опираясь и на наблюдаемые зарегистрированные свидетельства, и на практические соображениями здравого смысла, дает оценку вероятностей тех или иных исходов, привлекая атрибуты, которые на первый взгляд не имеют к этому отношения.

Этот пример использует поток *bayes_bankloan.str*, в котором используется файл данных *bankloan.sav*. Эти файлы доступны в каталоге *Demos* любого каталога установки IBM SPSS Modeler, и доступ к ним можно получить из группы программ IBM SPSS Modeler в меню Windows Пуск. Файл *bayes_bankloan.str* находится в каталоге *streams*.

Допустим, например, что банк заинтересован в оценке вероятности невозврата кредита. Если для предсказания потенциальных клиентов, склонных к отказу от платежей по кредиту, можно использовать предыдущие данные по неплатежеспособности, этим клиентам повышенного риска можно отказать в кредите или предложить им другой продукт.

Суть этого примера - использование существующих данных об отказе от платежей для предсказания будущих неплательщиков; здесь рассматривается три разных типа моделей байесовской сети для определения, какая из них лучше подходит для предсказаний в этой ситуации.

Построение потока

1. Добавьте узел источников файла статистики, указывающий на файл *bankloan.sav* в папке *Demos*.

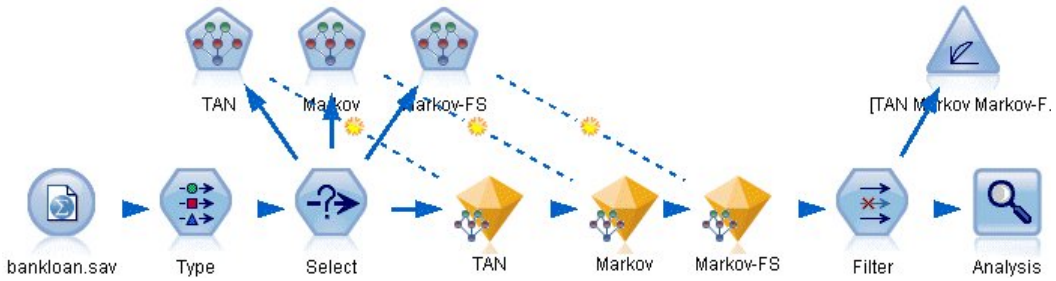


Рисунок 230. Поток примера байесовской сети

2. Добавьте узел Тип к узлу источника и задайте для поля **отказ от платежа** роль **Назначение**. Для всех остальных полей нужно задать роль **Ввод**.
3. Нажмите кнопку **Прочитать значения**, чтобы заполнить столбец *Значения*.

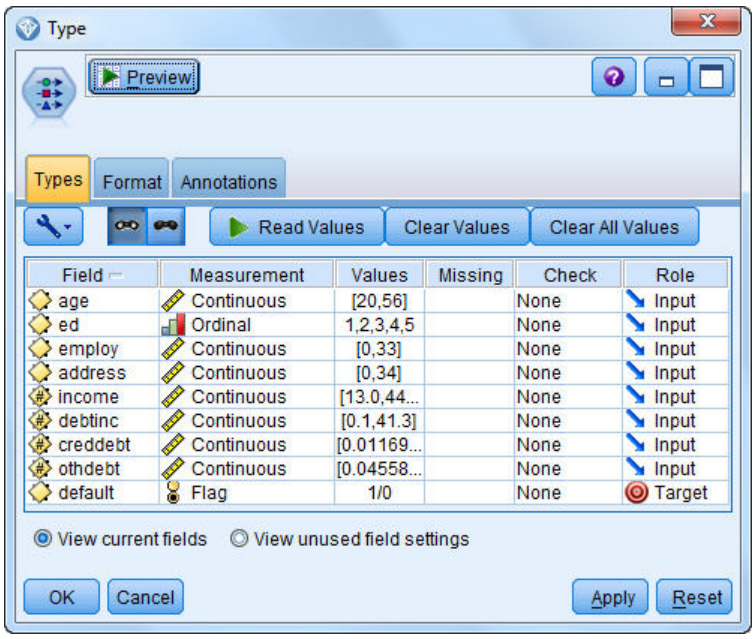


Рисунок 231. Выбор поля назначения

Наблюдения, для которых в поле назначения есть значение null, не используются при построении модели. Эти наблюдения можно исключить, чтобы они не использовались при оценке модели.

4. Добавьте узел Выбор к узлу Тип.
5. Для режима выберите значение **Отбрасывание**.
6. В поле Условие введите **default = '$null$'**.

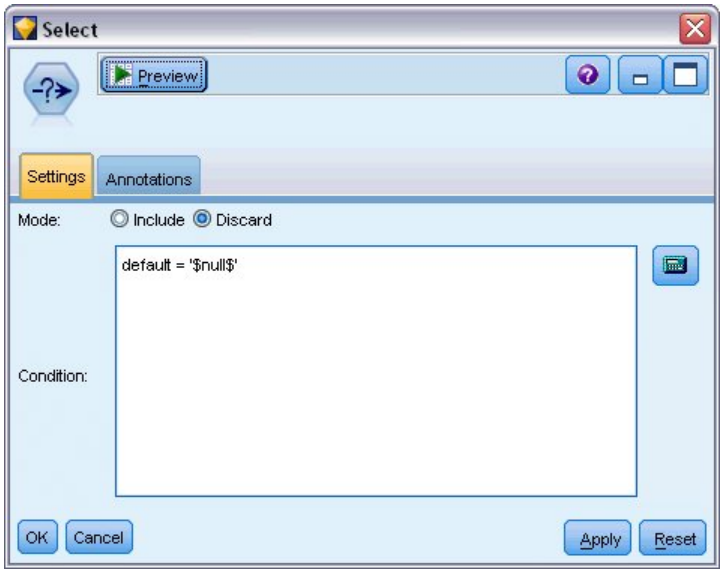


Рисунок 232. Отбрасывание полей назначения со значением null

Так как можно построить несколько разных типов байесовской сети, целесообразно сравнить их, чтобы понять, какая из этих моделей обеспечивает лучшие предсказания. Первая модель для создания - это усиленная деревом наивная байесовская модель (Tree Augmented Naïve Bayes, TAN).

7. Присоедините узел Байесовская сеть к узлу Выбор.

8. На вкладке Модель выберите для имени модели опцию **Пользовательское** и введите значение `TAN` в текстовом поле.
9. Для типа структуры выберите **TAN** и нажмите кнопку **OK**.

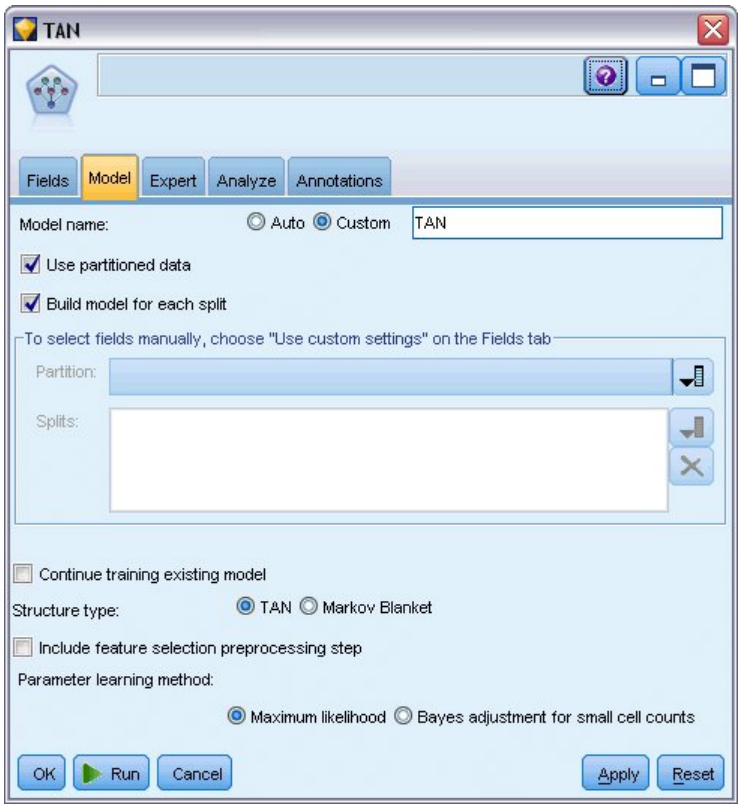


Рисунок 233. Создание усиленной деревом наивной байесовской модели

У второго типа модели для построения структура Марковское покрытие.

10. Присоедините второй узел байесовской сети к узлу Выбор.
11. На вкладке Модель выберите для имени модели опцию **Пользовательское** и введите значение `Марков` в текстовом поле.
12. Для типа структуры выберите **Марков** и нажмите кнопку **OK**.

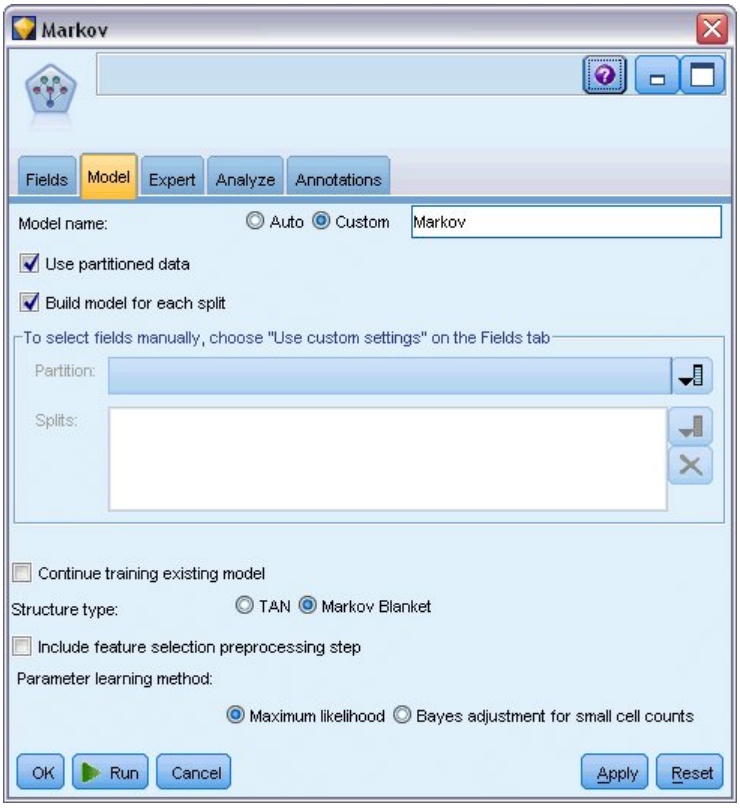


Рисунок 234. Создание модели Марковское покрытие

У третьего типа модели для построения структура Марковское покрытие, и также она использует предобработку выбора возможностей для выбора входных данных, которые существенно связаны с переменной назначения.

13. Присоедините третий узел байесовской сети к узлу Выбор.
14. На вкладке Модель выберите для имени модели опцию **Пользовательское** и введите значение `Марков-FS` в текстовом поле.
15. Для типа структуры выберите **Марковское покрытие**.
16. Выберите опцию **Включить шаг предобработки выбора возможностей** и нажмите кнопку **OK**.

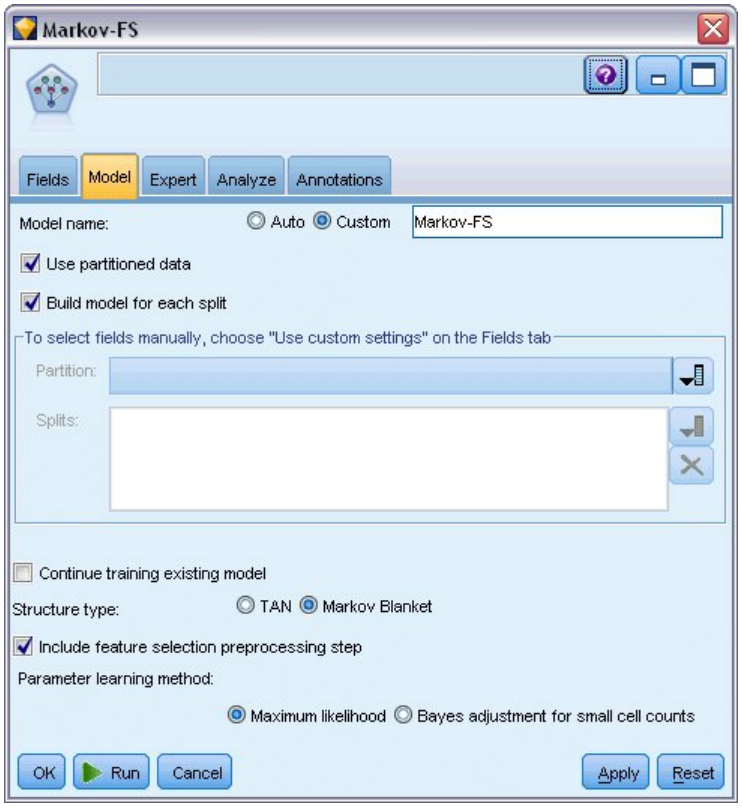


Рисунок 235. Создание модели марковского покрытия с предобработкой выбора возможностей

Просмотр модели

1. Запустите поток для создания слепков модели, которые будут добавлены в поток и на палитру Модели в правом верхнем углу. Для просмотра их подробностей дважды щелкните по любому из слепков модели в потоке.

   Вкладка Модель слепка модели разбита на две панели. На левой панели содержится сетевой граф узлов, показывающий взаимосвязь между назначением и наиболее важными предикторами, а также взаимосвязи между предикторами.

   Правая панель показывает или *Важность предикторов*, определяющую относительную важность каждого предиктора при оценке модели, или *Условные вероятности*, то есть условную вероятность для каждого значения узла и для каждой комбинации значений в родительских узлах.

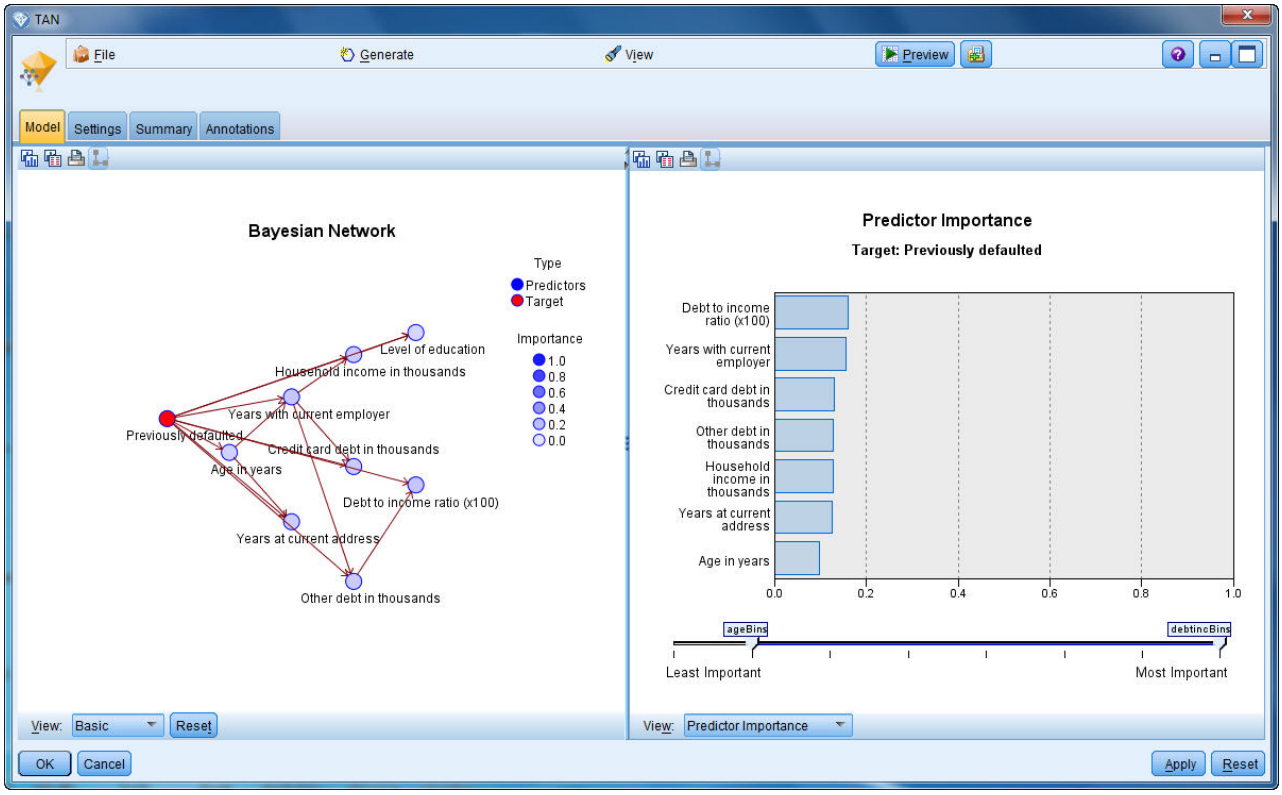


Рисунок 236. Просмотр усиленной деревом наивной байесовской модели

2. Присоедините слепок модели TAN к марковскому слепку (выберите опцию **Заменить** в диалоговом окне предупреждения).
3. Соедините марковский слепок со слепком Марков-FS (выберите опцию **Заменить** в диалоговом окне предупреждения).
4. Для простоты представления выравняйте три слепка с узлом Выбор.

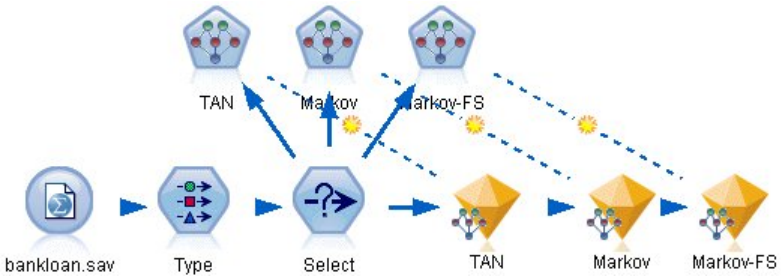


Рисунок 237. Выравнивание слепков в потоке

5. Чтобы для ясности переименовать данные вывода модели на диаграмме оценки, которую вы будете создавать, присоедините узел Фильтр к слепку модели Марков-FS.
6. В правом столбце *Поле* переименуйте $B-default на TAN, $B1-default на Марков и $B2-default на Марков-FS.

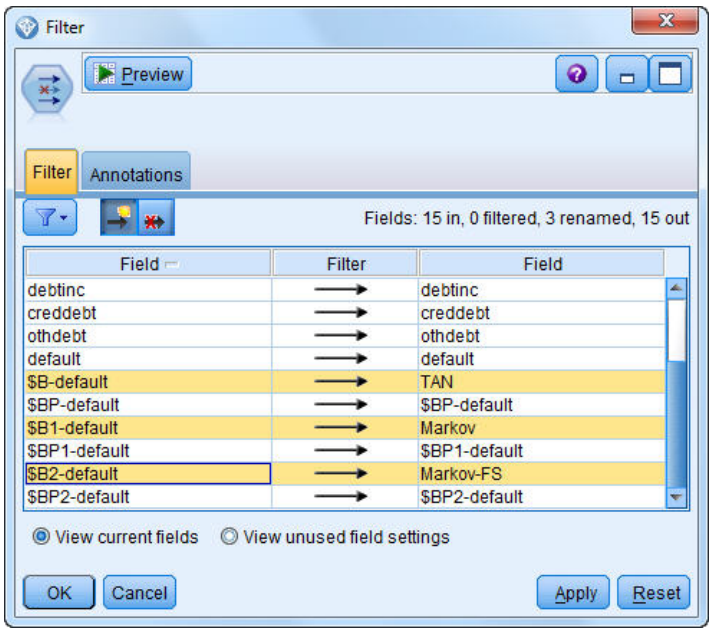


Рисунок 238. Переименование полей модели

Для сравнения предсказанной точности моделей можно построить диаграмму выигрыша.

7. Присоедините узел диаграммы оценки к узлу Фильтр и выполните узел этой диаграммы с его параметрами по умолчанию.

   На этой диаграмме видно, что все типы моделей приводят к аналогичному результату, но марковская модель несколько лучше.

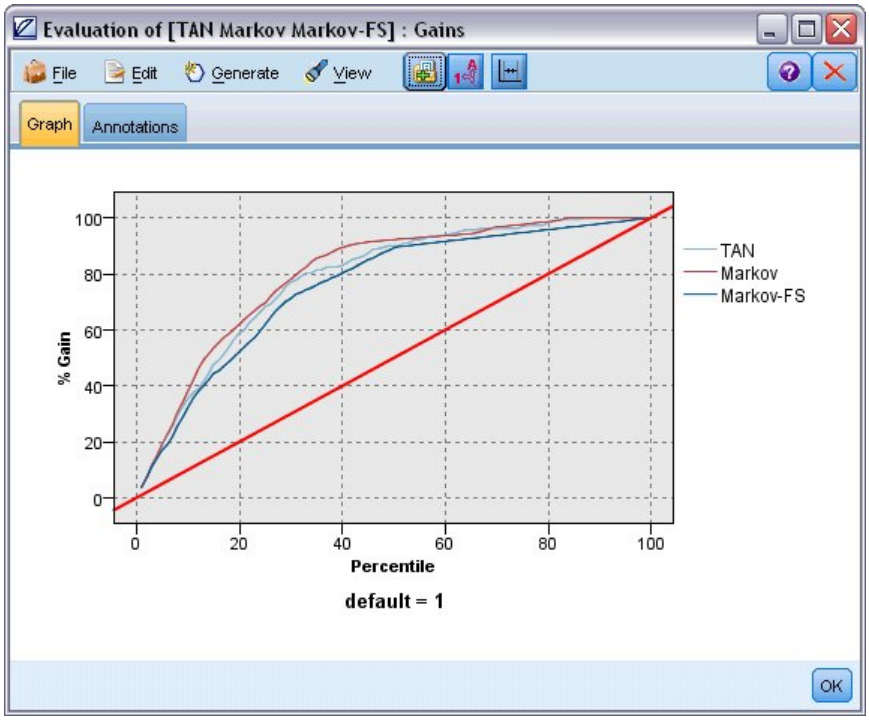


Рисунок 239. Оценка точности модели

Для проверки, насколько хорошо предсказывает каждая модель, можно было использовать узел Анализ вместо диаграммы Оценка. Он показывает точность в терминах процентной доли правильных и неправильных предсказаний.

8. Присоедините узел Анализ к узлу Фильтр и выполните узел Анализ с использованием его параметров по умолчанию.

Как и в случае диаграммы Оценка, здесь видно, что марковская модель несколько лучше в правильных предсказаниях, однако модель Марков-FS отстает от нее всего на несколько процентных пунктов. Это может означать, что предпочтительнее было бы использовать модель Марков-FS, так как для вычисления результатов она использует меньше входных полей, то есть требуется меньшее собрание данных и меньшее время на ввод и обработку.

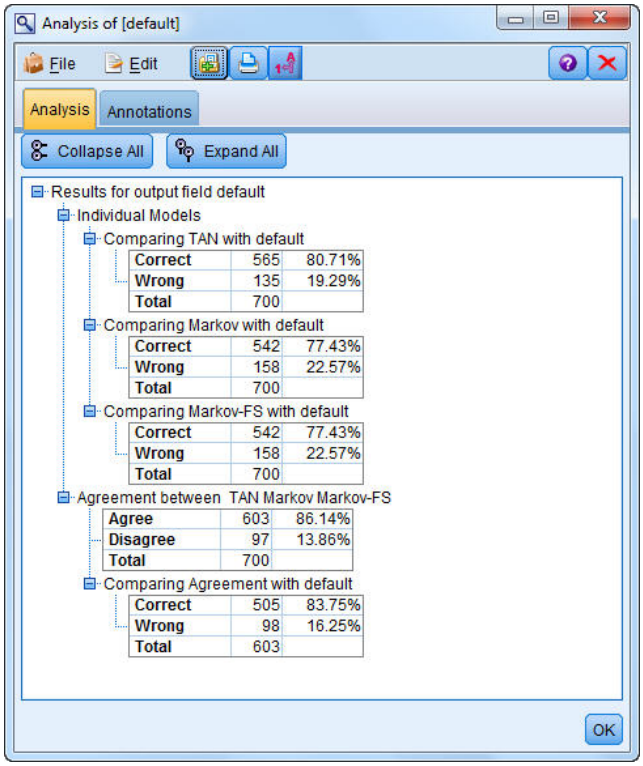


Рисунок 240. Анализ точности модели

Объяснение математических основ методов моделирования, используемых в IBM SPSS Modeler, смотрите в публикации *IBM SPSS Modeler: Руководство по алгоритмам* в каталоге *\Documentation* на установочном диске.

Учтите, что эти результаты основаны только на обучающих данных. Чтобы оценить, насколько хорошо модель обобщается на другие данные реального мира, рекомендуется применить узел Разбиение, который будет содержать поднабор записей в целях проверки.

Глава 18. Ежемесячное переобучение модели (байесовская сеть)

Байесовские сети позволяют построить вероятностную модель, которая, опираясь и на наблюдаемые зарегистрированные свидетельства, и на практические соображениями здравого смысла, дает оценку вероятностей тех или иных исходов, привлекая атрибуты, которые на первый взгляд не имеют к этому отношения.

В этом примере используется поток *bayes_churn_retrain.str*, ссылающийся на файлы данных *telco_Jan.sav* и *telco_Feb.sav*. Эти файлы доступны в каталоге *Demos* любого каталога установки IBM SPSS Modeler, и доступ к ним можно получить из группы программ IBM SPSS Modeler в меню Windows Пуск. Файл *bayes_churn_retrain.str* находится в каталоге *streams*.

Например, предположим, что провайдер телекоммуникационных услуг озабочен количеством клиентов, теряемых из-за конкурирующих компаний (отток клиентов). Если для предсказания, какие клиенты с наибольшей вероятностью откажутся от услуг в будущем, можно использовать хронологические данные клиентов, этим клиентам можно предложить поощрения или сделать другие спецпредложения, чтобы препятствовать их переходу к другим поставщикам услуг.

Суть этого примера - использование существующих данных об оттоке клиентов за месяц, чтобы предсказать, какие клиенты с наибольшей вероятностью откажутся от услуг в будущем, и добавление данных следующего месяца, чтобы уточнить и переобучить модель.

Построение потока

1. Добавьте узел источников файла статистики, указывающий на файл *telco_Jan.sav* в папке *Demos*.

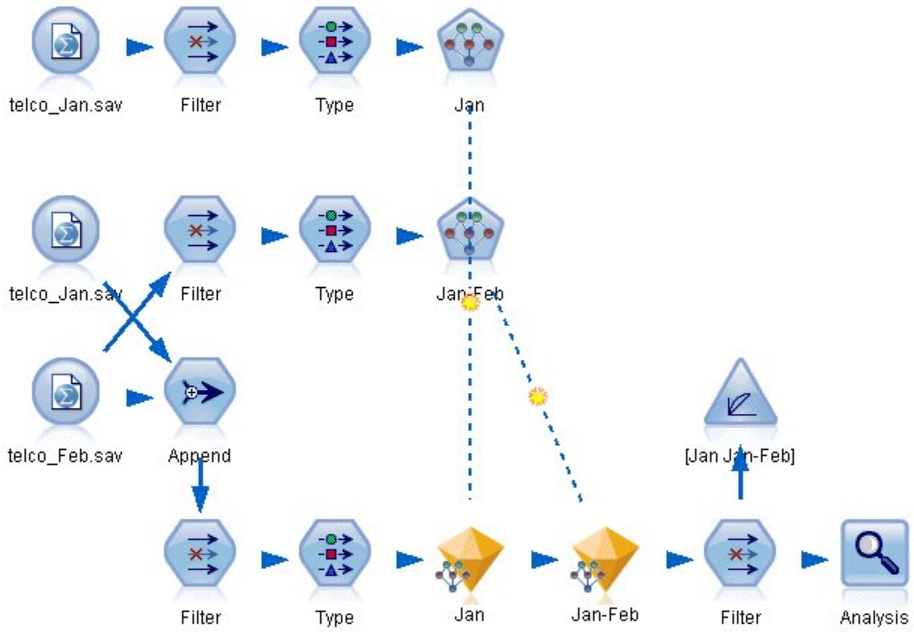


Рисунок 241. Поток примера байесовской сети

Предыдущий анализ показал, что нескольких полей данных мало значимы для предсказания оттока клиентов. Эти поля можно отфильтровать из вашего набора данных, чтобы повысить скорость обработки при построении и скоринге моделей.

2. Добавьте узел Фильтр к узлу Источник.
3. Исключите все поля, кроме полей *address*, *age*, *churn*, *custcat*, *ed*, *employ*, *gender*, *marital*, *reside*, *retire* и *tenure* (адрес, возраст, отток клиентов, категория клиента, образование, занятость, пол, семейное положение, место жительства, на пенсии и срок обслуживания).
4. Щелкните по **OK**.

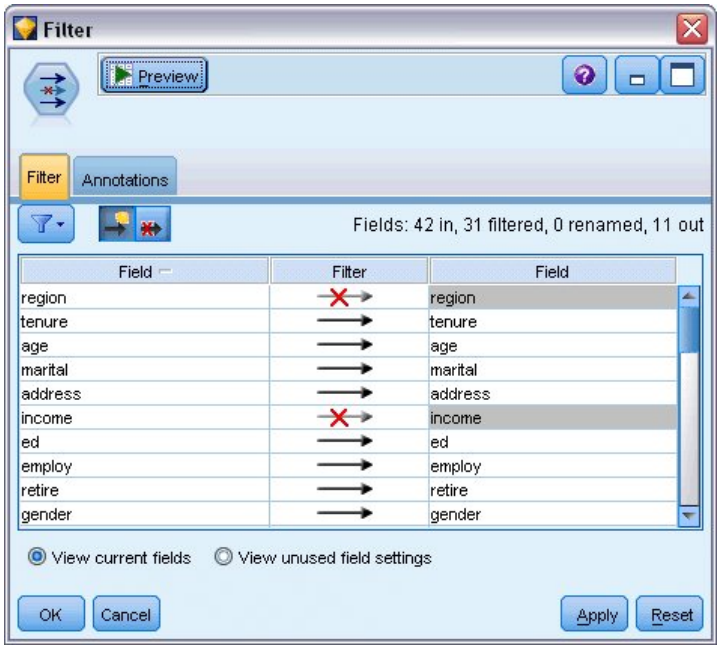


Рисунок 242. Фильтрация ненужных полей

5. Добавьте узел Тип к узлу Фильтр.
6. Откройте узел Тип и нажмите кнопку **Прочитать значения**, чтобы заполнить столбец *Значения*.
7. Чтобы узел оценки мог принять решение, какое из значений - это true, а какое - false, задайте для поля *churn* (отток клиентов) уровень измерения **Флаг** и роль **Назначение**. Щелкните по **OK**.

Рисунок 243. Выбор поля назначения

Можно построить несколько разных типов байесовской сети, однако для этого примера вы собираетесь построить укрепленную деревом наивную байесовскую модель (Tree Augmented Naïve Bayes, TAN). Это создает большую сеть и обеспечивает включение вами всех возможных связей между переменными данных, тем самым строя устойчивую начальную модель.

8. Присоедините узел Байесовская сеть к узлу Тип.
9. На вкладке Модель выберите для имени модели опцию **Пользовательское** и введите значение `Jan` в текстовом поле.
10. Для способа обучения параметров выберите **Байесовская корректировка для малых чисел в ячейках**.
11. Нажмите кнопку **Выполнить**. Слепок модели добавляется в поток и на палитру Модели в правом верхнем углу.

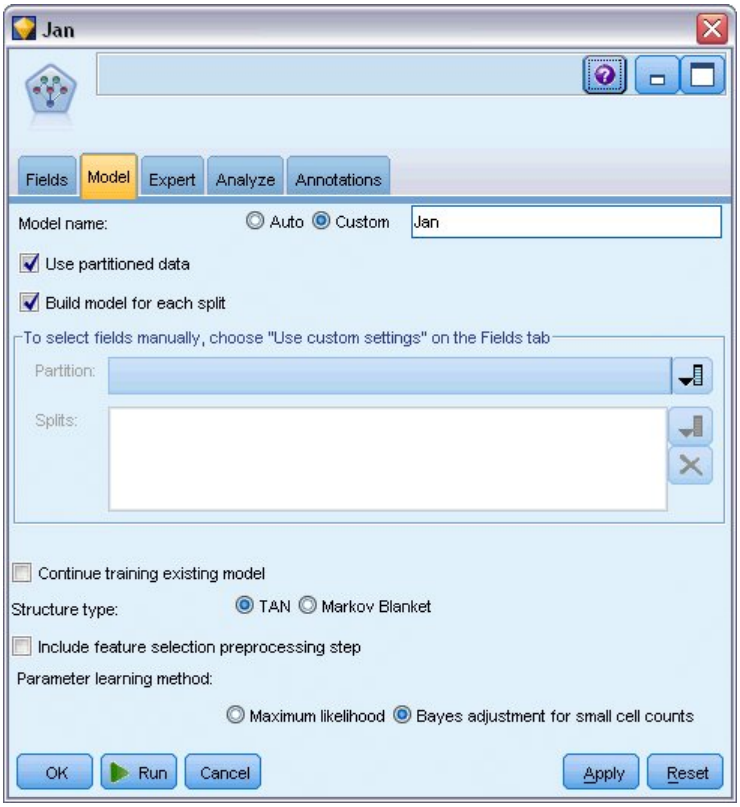


Рисунок 244. Создание усиленной деревом наивной байесовской модели

12. Добавьте узел источников файла статистики, указывающий на файл *telco_Feb.sav* в папке *Demos*.
13. Присоедините этот узел источников к узлу Фильтр (в диалоговом окне предупреждения выберите опцию **Заменить**, чтобы заменить соединение с предыдущим узлом источников).

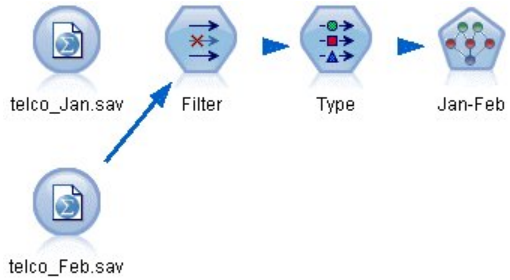


Рисунок 245. Добавление данных второго месяца

14. На вкладке Модель узла байесовской сети выберите для имени модели опцию **Пользовательское** и введите значение `Jan-Feb` в текстовом поле.
15. Выберите **Продолжить обучение существующей модели**.
16. Нажмите кнопку **Выполнить**. Этот слепок модели перезапишет существующий в потоке; он также добавляется на палитру Модели в правом верхнем углу.

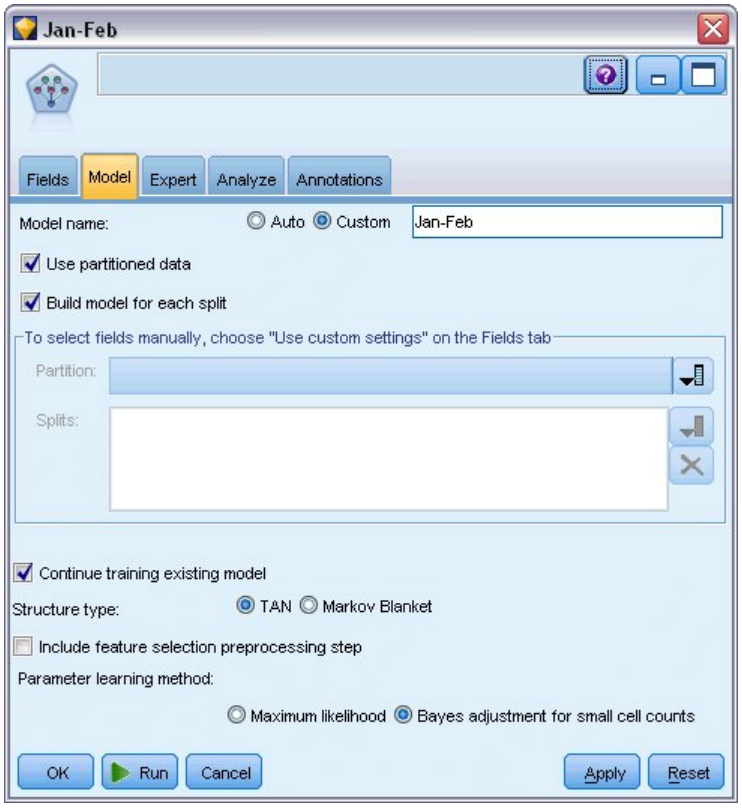


Рисунок 246. Повторное обучение модели

Оценка модели

Для сравнения моделей необходимо сочетать два набора данных.

1. Добавьте узел добавления и присоедините к нему узлы источника *telco_Jan.sav* и *telco_Feb.sav*.

Рисунок 247. Добавление двух источников данных

2. Скопируйте узлы Фильтр и Тип, более ранние по потоку, и поместите их на холст потока.
3. Присоедините узел добавления к вновь скопированному узлу Фильтр.

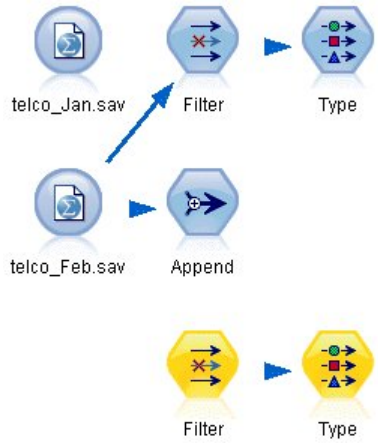


Рисунок 248. Вставка скопированных узлов в поток

Слепки для двух моделей байесовской сети расположены на палитре Модели в правом верхнем углу.

4. Дважды щелкните по слепку модели Jan, чтобы перенести его в поток, и присоедините этот слепок к вновь скопированному узлу Тип.
5. Присоедините слепок модели Jan-Feb, который уже в потоке, к слепку модели Jan.
6. Откройте слепок модели Jan.

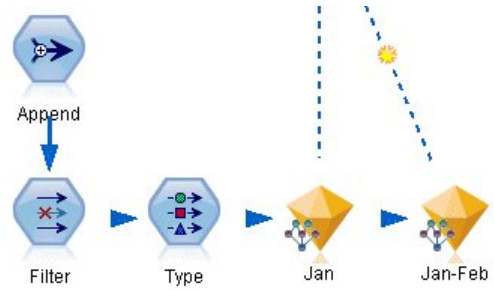


Рисунок 249. Добавление слепков в поток

Вкладка Модель слепка модели байесовской сети разделена на два столбца. В левом столбце содержится сетевой граф узлов, показывающий взаимосвязи поля назначения и его наиболее важных предикторов, а также взаимосвязи между предикторами.

В правом столбце показана или *Важность предикторов*, то есть определяется относительная важность каждого предиктора в оценке модели, или *Условные вероятности*, в которые входят значения условных вероятностей для всех значений узлов и для всех комбинаций значений в родительских узлах.

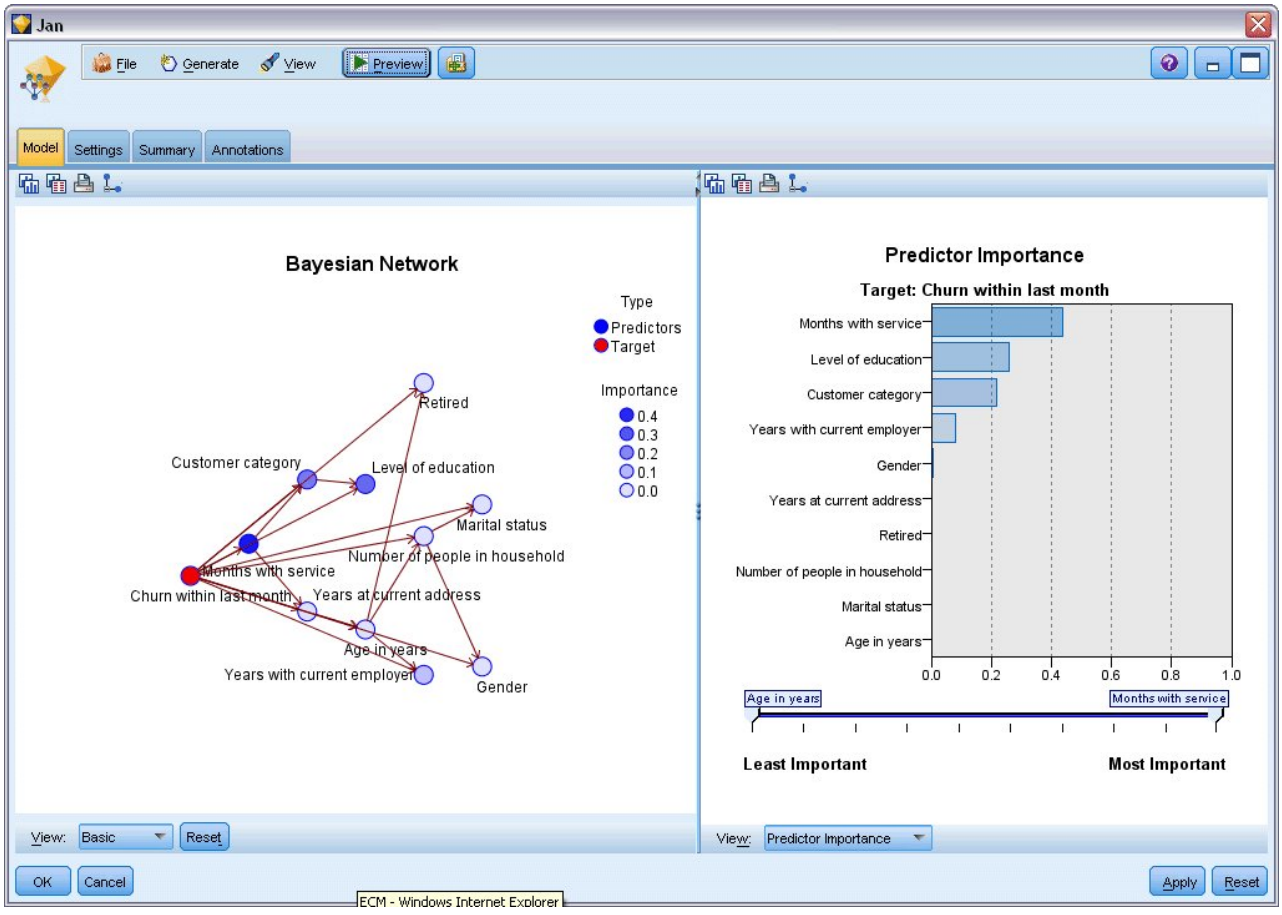


Рисунок 250. Модель байесовской сети, показывающая важность предикторов

Чтобы вывести условные вероятности для любого узла, щелкните по этому узлу в левом столбце. Правый столбец изменится, и в нем будут показаны нужные подробности.

Условные вероятности показываются для каждого из интервалов, на которые были разделены значения данных, относительно родительских и равноуровневых узлов данного узла.

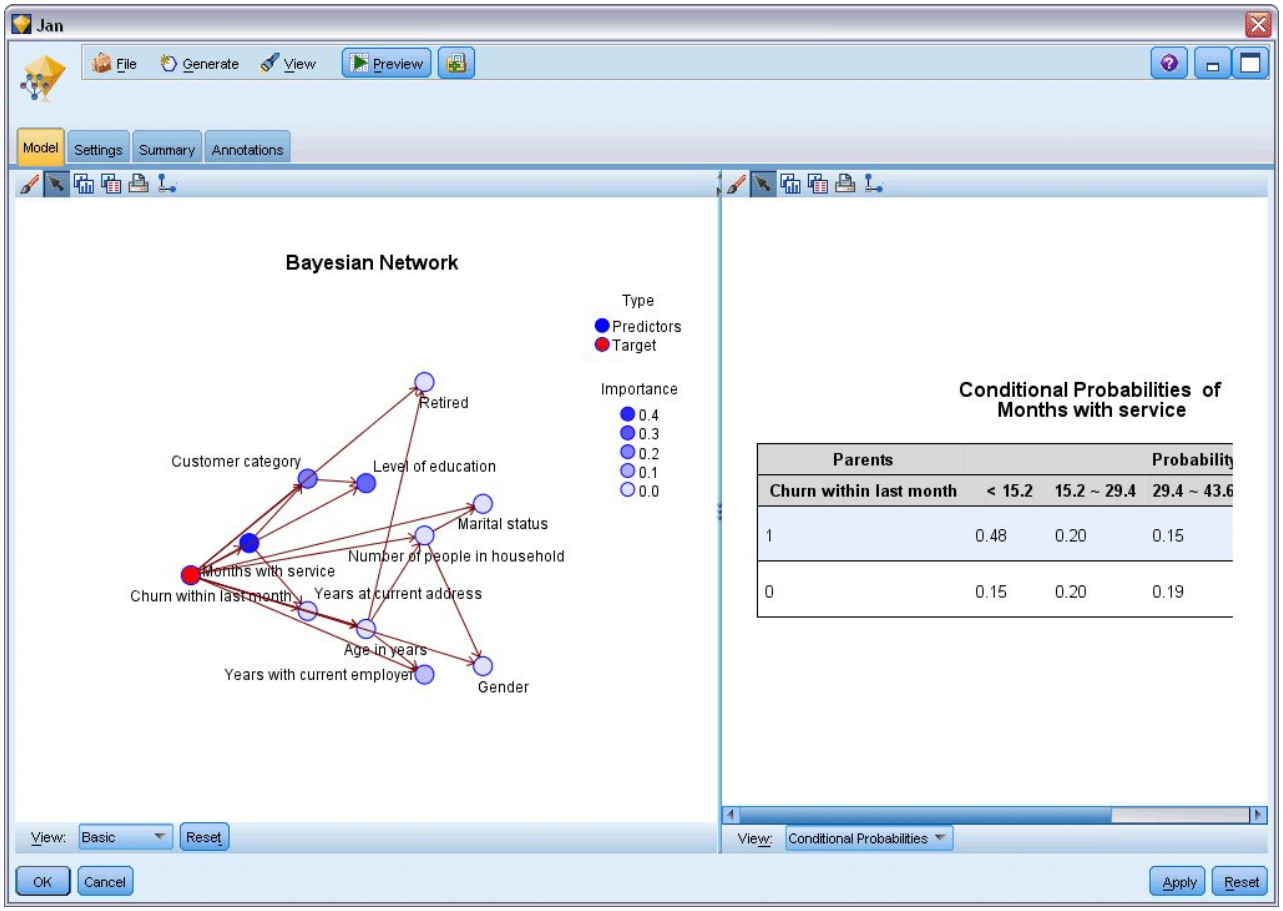


Рисунок 251. Модель байесовской сети, показывающая условные вероятности

7. Чтобы для ясности переименовать данные вывода модели, присоедините узел Фильтр к слепку модели Jan-Feb.
8. В правом столбце *Поле* переименуйте $B-churn в Jan, а $B1-churn - в Jan-Feb.

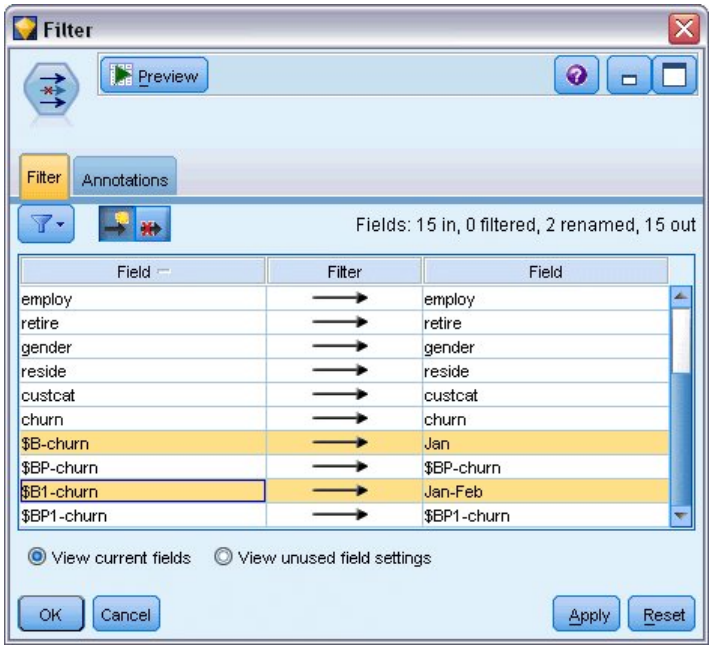


Рисунок 252. Переименование полей модели

Для проверки, насколько хорошо каждая модель предсказывает отток клиентов, используйте узел Анализ; точность будет показана в терминах процентной доли правильных и неправильных предсказаний.

9. Присоедините узел Анализ к узлу Фильтр.
10. Откройте узел Анализ и нажмите кнопку **Выполнить**.

    Видно, что у обеих моделей аналогичная степень точности при предсказании оттока клиентов.

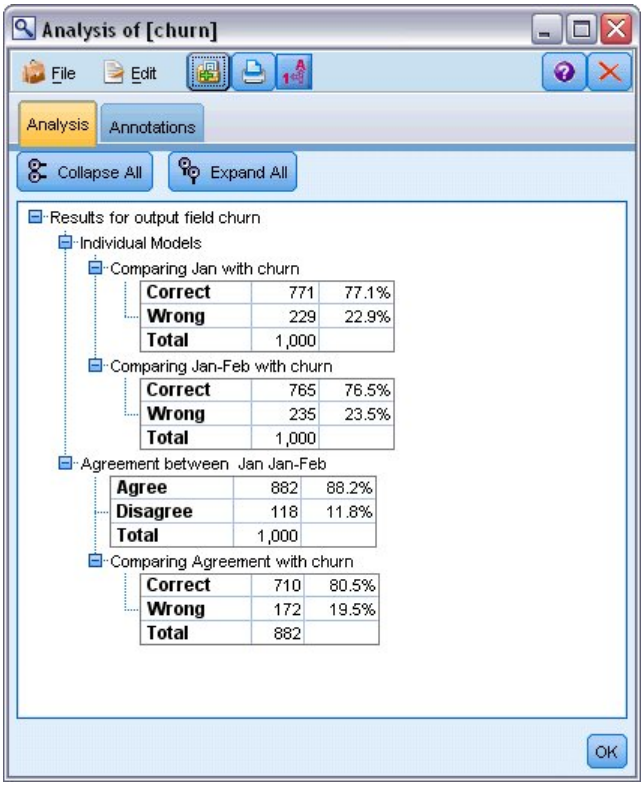


Рисунок 253. Анализ точности модели

Вместо узла Анализ можно использовать диаграмму оценки, чтобы сравнивать предсказанную точность моделей при помощи построения диаграммы выигрыша.

11. Присоедините диаграмму оценки к узлу Фильтр

и выполните узел этой диаграммы с его параметрами по умолчанию.

Как и на узле Анализ, эта диаграмма показывает, что модели всех типов приводят примерно к одинаковому результату; однако повторно обученная модель, использующая данные двух месяцев, немного лучше, так как для ее предсказаний выше уровень доверительной вероятности.

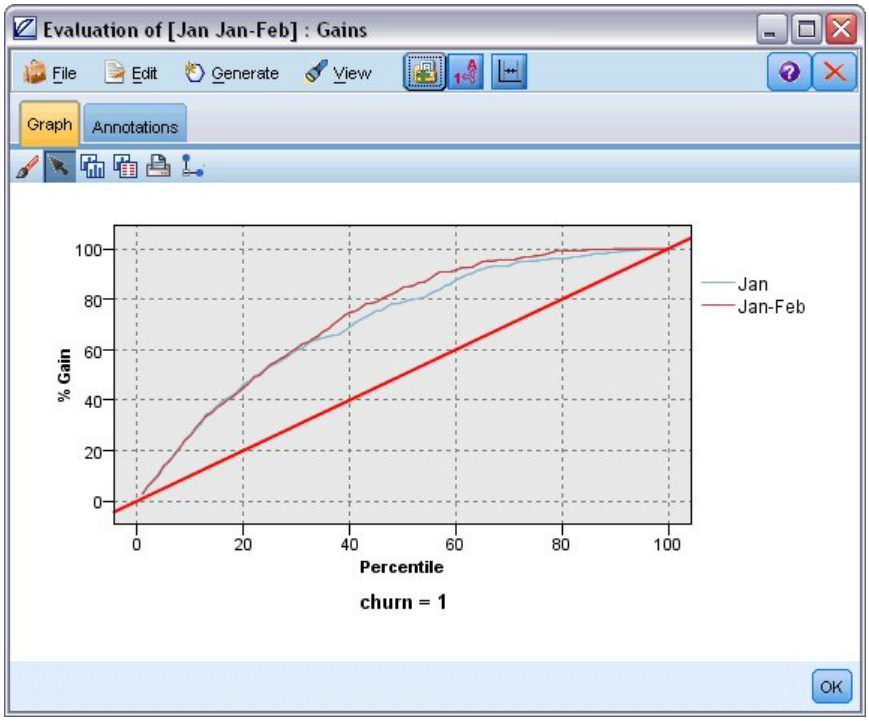


Рисунок 254. Оценка точности модели

Объяснение математических основ методов моделирования, используемых в IBM SPSS Modeler, смотрите в публикации *IBM SPSS Modeler: Руководство по алгоритмам* в каталоге *\Documentation* на установочном диске.

Учтите, что эти результаты основаны только на обучающих данных. Чтобы оценить, насколько хорошо модель обобщается на другие данные реального мира, рекомендуется применить узел Разбиение, который будет содержать поднабор записей в целях проверки.

Глава 19. Рекламная кампания для розничной продажи (нейросеть/C&RT)

В этом примере рассматриваются данные, описывающие линейки товаров в розничной торговле и влияние рекламной кампании на продажи. (Это фиктивные данные.) Ваша цель в этом примере - предсказать влияние будущих рекламных кампаний на продажи. Аналогично примеру с мониторингом условий, этот процесс исследования данных состоит из фаз исследования, подготовки данных, обучения и проверки.

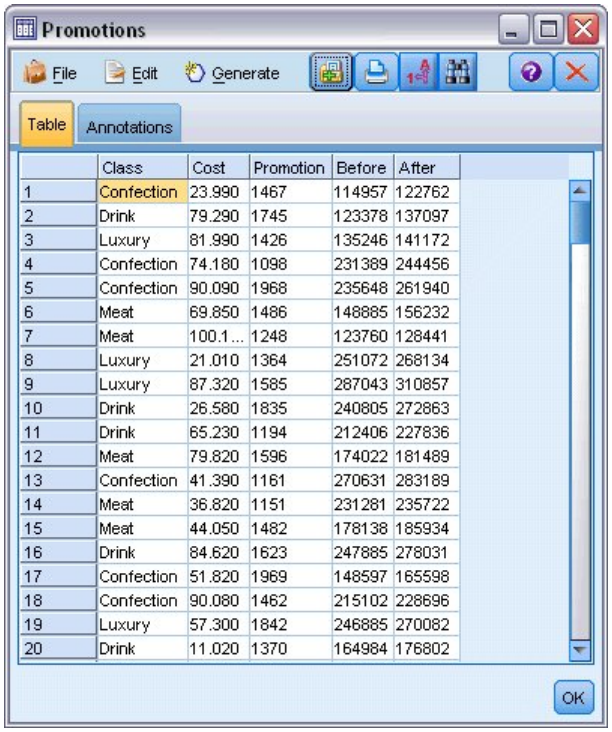
В этом примере используются потоки с именами *goodsplot.str* и *goodslearn.str*, связанные с файлами данных *GOODS1n* и *GOODS2n*. Эти файлы находятся в каталоге *Demos* любой установки IBM SPSS Modeler. Доступ к ним можно получить из группы программ IBM SPSS Modeler в меню Пуск Windows. Поток *goodsplot.str* находится в папке *streams*, а файл *goodslearn.str* - в каталоге *streams*.

Изучение данных

Каждая запись содержит:

- *Класс*. Тип продукта.
- *Стоимость*. Цена за единицу.
- *Рекламная кампания*. Показатель затрат на конкретную рекламную кампанию.
- *До*. Доход до рекламной кампании.
- *После*. Доход после рекламной кампании.

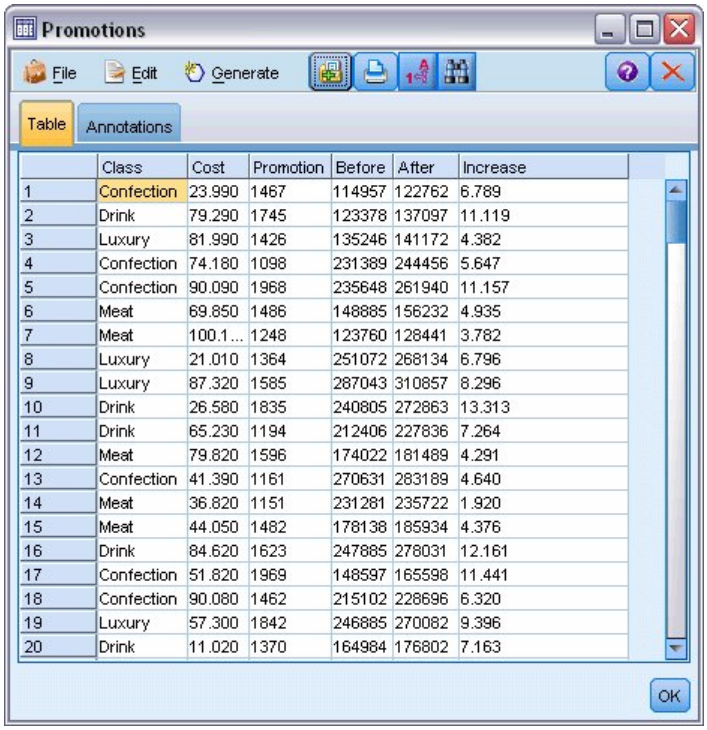
Поток *goodsplot.str* содержит простой поток для вывода данных в таблицу. Два поля доходов (*До* и *После*) выражены в абсолютных величинах; однако более полезным кажется график увеличения доходов после рекламной кампании (предположительно, в ее результате).



	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

Рисунок 255. Влияние рекламной кампании на продажи товаров

goodsplot.str содержит также узел для извлечения этого значения, выраженного как процентная доля дохода до рекламной кампании в поле *Повышение*, и выводит таблицу, показывающую это поле.



	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

Рисунок 256. Повышение доходов после рекламной кампании

Кроме этого, этот поток выводит гистограмму повышения и его диаграмму рассеяния в зависимости от затрат на кампанию с наложением категории участвующего товара.

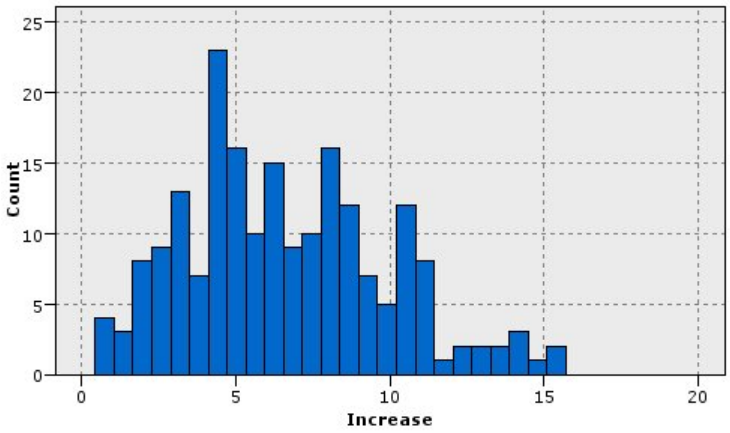


Рисунок 257. Гистограмма повышения доходов

Диаграмма рассеяния показывает, что для каждого класса товаров наблюдается почти линейная зависимость роста доходов от затрат на рекламную кампанию. Поэтому весьма вероятно, что дерево решений или нейронная сеть могли бы весьма точно предсказать повышение доходов по другим доступным полям.

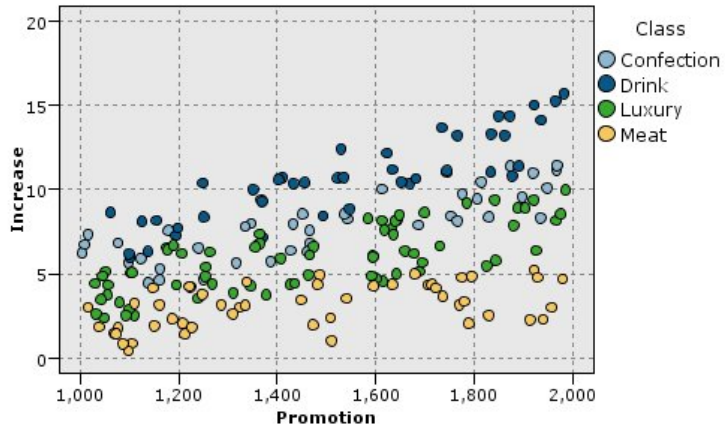


Рисунок 258. Повышение доходов в зависимости от затрат на рекламную кампанию

Обучение и проверка данных

Поток `goodslearn.str` обучает нейронную сеть и дерево решений для выполнения предсказаний роста доходов.

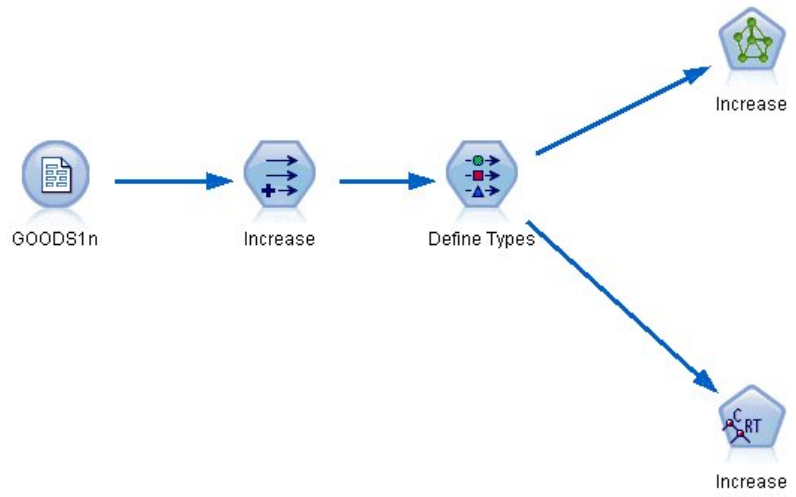


Рисунок 259. Поток моделирования goodslearn.str

После выполнения узлов моделей и генерирования фактических моделей можно проверить результаты процесса обучения. Для этого дерево решений и нейросеть соединяется подряд с узлом Тип и с новым узлом Анализ, входной файл данных изменяется на GOODS2n и выполняется узел Анализ. По выходным данным этого узла, в частности по линейной корреляции между предсказанным ростом и действительными данными, вы обнаружите, что обученные системы весьма успешно предсказывают повышение дохода.

Дальнейшее исследование можно сфокусировать на тех наблюдениях, для которых обученные системы делают относительно большие ошибки; эти случаи можно идентифицировать, откладывая на графике предсказанный рост в зависимости от фактического. Выбросы на этом графике можно выбрать, используя интерактивные средства работы с графикой в SPSS Modeler, и по их свойствам можно настроить описания данных или процесс обучения для повышения точности.

Глава 20. Мониторинг условий (нейронная сеть/C5.0)

Этот пример относится к компьютерной информации о состоянии мониторинга условий и к проблеме распознавания и предсказания аварийных состояний. Данные созданы в результате фиктивного моделирования и состоят из нескольких последовательных временных рядов измерений. Каждая запись - это отчет о снимке для компьютера со следующими элементами:

- *Время*. Целое число.
- *Мощность*. Целое число.
- *Температура*. Целое число.
- *Давление*. 0 для нормального давления, 1 для текущего предупреждения о давлении.
- *Рабочее время*. Время с момента последнего техобслуживания.
- *Состояние*. Для нормального состояние 0, при ошибке изменяется на код ошибки (101, 202 или 303).
- *Выходные данные*. Код ошибки, проявляющейся в этом временном ряду, или 0, если ошибок нет. (Эти коды доступны только при использовании возможностей оценки прошедших событий.)

В этом примере используются потоки с именами *condplot.str* и *condlearn.str*, связанные с файлами данных *COND1n* и *COND2n*. Эти файлы находятся в каталоге *Demos* любой установки IBM SPSS Modeler. Доступ к ним можно получить из группы программ IBM SPSS Modeler в меню Пуск Windows. Файлы *condplot.str* и *condlearn.str* находятся в каталоге *streams*.

Для каждого временного ряда существует ряд записей за период нормального функционирования, после которого следует период, приводящий к ошибке, как показано в следующей таблице:

Время	Мощность	Температура	Давление	Рабочее время	Статус	Результат
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
53	938	251	0	209	0	101
54	936	251	0	209	0	101
			...			
208	644	251	0	209	0	101

Время	Мощность	Температура	Давление	Рабочее время	Статус	Результат
209	640	251	0	209	101	101

Следующий процесс - общий для большинства проектов исследования данных:

- Проверьте данные, чтобы определить, какие атрибуты могут быть релевантными для предсказания или распознавания представляющих интерес состояний.
- Сохраните эти атрибуты (если они уже существуют) или вычислите и добавить их к данным, если это необходимо.
- Используйте полученные данные для обучения правил и нейронных сетей.
- Проверьте обученные системы, используя независимые проверочные данные.

Изучение данных

Файл *condplot.str* иллюстрирует первую часть этого процесса. Он содержит поток, по которому строится несколько диаграмм. Если во временных рядах температуры или мощности есть видимые паттерны, можно различить условия угрожающих ошибок или, возможно, предсказывать их возникновение. И для температуры, и для мощности поток ниже изображает значения временных рядов с тремя разными кодами ошибок на отдельных графиках, что дает в сумме шесть графиков. Узлы выбора разделяют данные, связанные с тремя различными кодами ошибок.

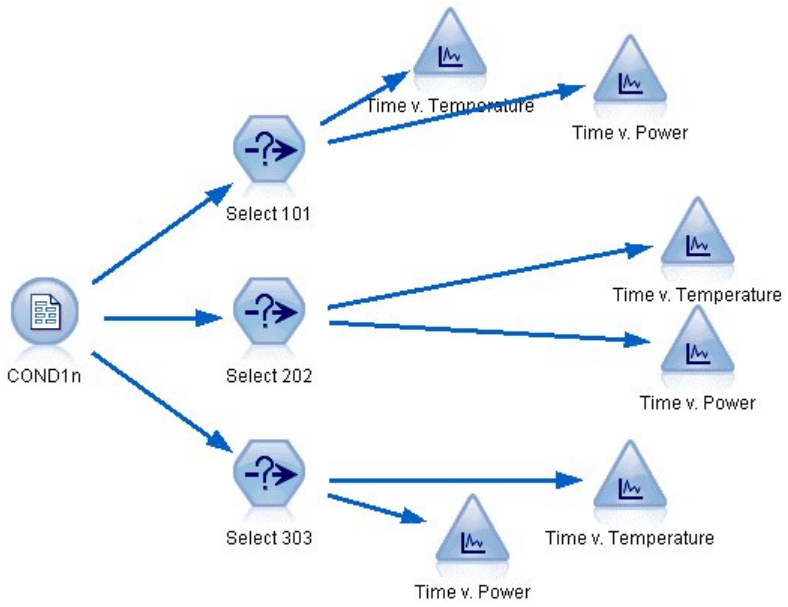


Рисунок 260. Поток Condplot

На этом рисунке показаны результаты данного потока.

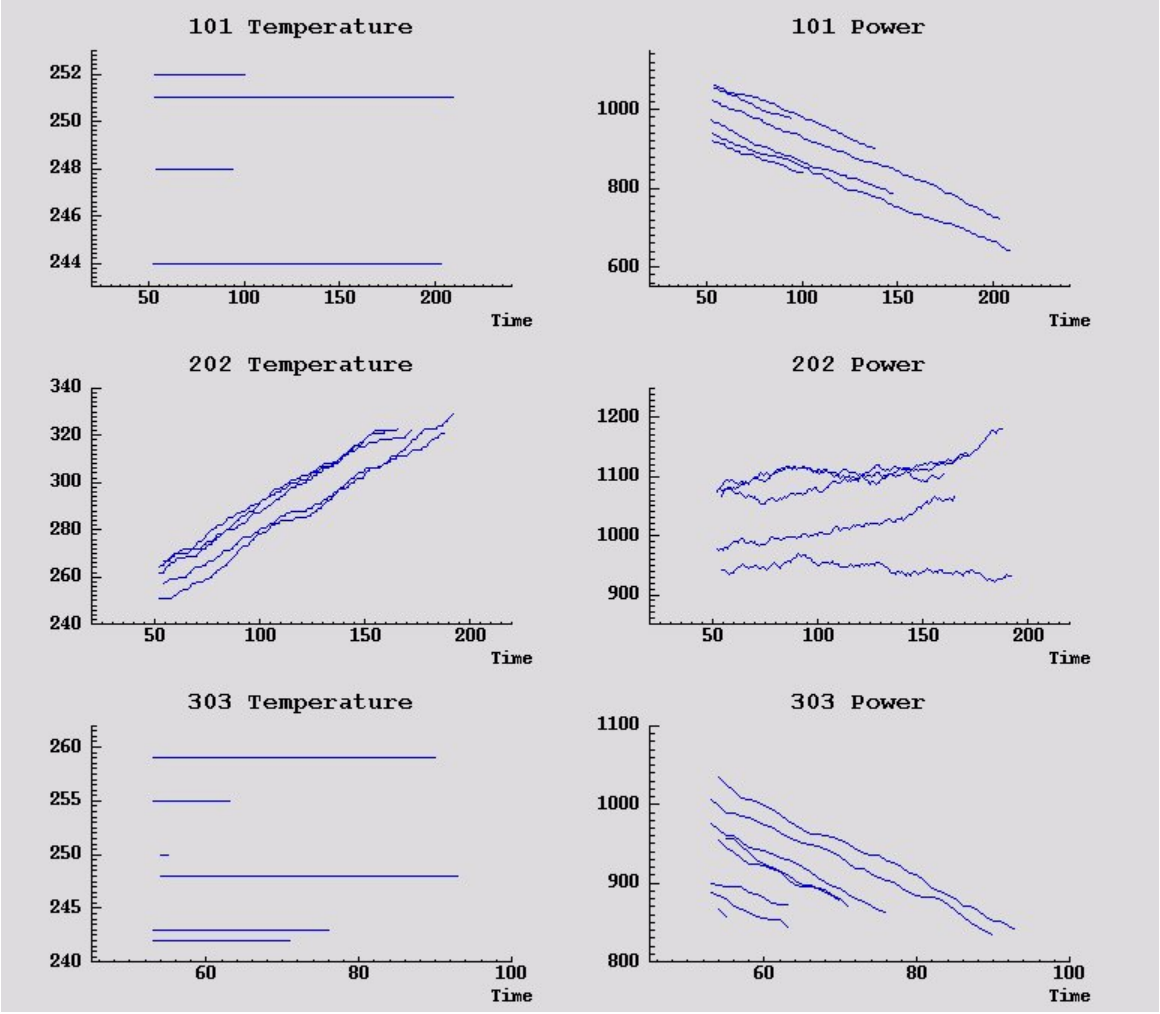


Рисунок 261. Изменение температуры и давления со временем

На этих графиках ясно показаны паттерны, отличающие ошибку 202 от ошибок 101 и 303. Ошибка 202 связаны с ростом температуры и флуктуациями мощности, а другие ошибки - нет. Однако разделение паттернов для ошибок 101 и 303 менее очевидно. Обеим ошибкам сопутствует стабильная температура и падение мощности, но для ошибок 303 падение мощности кажется более резким.

На основании этих графиков выявляется, что наличие изменений для температуры и мощности и скорость этих изменений, а также наличие и степень флуктуаций могут быть релевантными показателями для предсказания и распознавания сбоев. Поэтому эти атрибуты нужно добавить к данным, прежде чем применять системы обучения.

Подготовка данных

На основании результатов исследования данных поток *condlearn.str* получает надежные данные и обучается предсказывать ошибки.

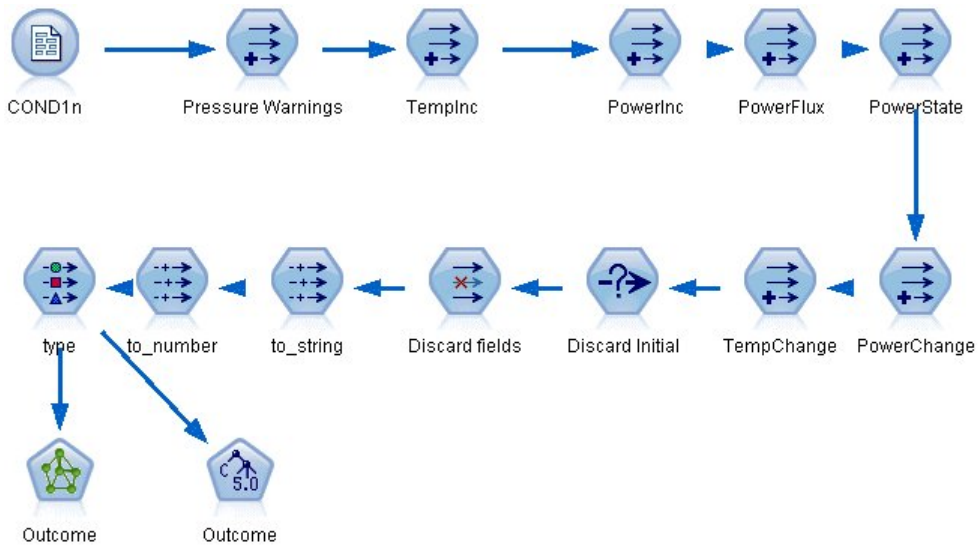


Рисунок 262. Поток Condlearn

Этот поток использует несколько узлов извлечения, чтобы подготовить данные для моделирования.

- **Узел файла переменных**. Читает файл данных *COND1n*.
- **Получить предупреждения о давлении**. Подсчитывает число отдельных предупреждений о давлении. Сбрасывает показания, когда время возвращается к значению 0.
- **Получить TempInc**. Вычисляет текущую скорость изменения температуры, используя `@DIFF1`.
- **Получить PowerInc**. Вычисляет текущую скорость изменения мощности, используя `@DIFF1`.
- **Получить PowerFlux**. Флаг, его значение true, если мощность изменяется в разных направлениях в этой и в предыдущей записи, то есть для пика или провала мощности.
- **Получить PowerState**. Состояние с начальным значением *Стабильное*; оно переключается на *Флуктуирующее*, когда обнаружены два последовательных экстремума мощности. Переключается обратно на *Стабильное* только в случае отсутствия экстремумов мощности в течение пяти интервалов времени или когда сбрасывается значение *Время*.
- **PowerChange**. Среднее значение *PowerInc* за пять последних интервалов времени.
- **TempChange**. Среднее значение *TempInc* за пять последних интервалов времени.
- **Отбросить начальные значения (выбрать)**. Отбрасывает первую запись каждого временного ряда, чтобы исключить большие ложные скачки *Давления* и *Температуры* на границах.
- **Отбросить поля**. Обрезает записи до полей *Рабочее время*, *Состояние*, *Выходные данные*, *Предупреждения о давлении*, *PowerState*, *PowerChange* и *TempChange*.
- **Тип**. Определяет роль поля *Выходные данные* как **Назначение** (поле, которое будет предсказываться). Кроме этого, определяет уровень измерения поля *Выходные данные* как **Номинальное**, поля *Предупреждения о давлении* - как **Количественное**, а поля *PowerState* - как **Флаг**.

Обучение

Запуск потока в *condlearn.str* обучает нейронную сеть и готовит правило C5.0. Обучение сети может потребовать некоторого времени, но его можно прервать и раньше, чтобы сохранить сеть, которая даст достоверные результаты. После завершения обучения вкладка Модели в правом верхнем углу окна менеджеров будет мигать, оповещая о создании двух новых слепков: один для представления нейронной сети и один для правила.

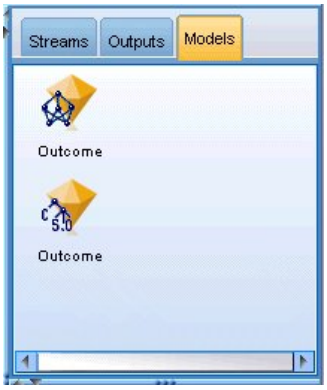


Рисунок 263. Управление моделями при помощи слепков моделей

Слепки модели добавляются также в существующий поток, позволяя проверять систему или экспортировать результаты модели. В этом примере мы проверим результаты модели.

Проверка

Слепки модели добавляются в поток, причем оба из них соединяются с узлом Тип.

1. Переставьте слепки, как показано, чтобы узел Тип соединялся со слепком нейронной сети, который в свою очередь соединялся бы со слепком C5.0.
2. Присоедините узел Анализ к слепку C5.0.
3. Измените исходный узел источника для чтения файла *COND2n* (вместо *COND1n*), так как *COND2n* содержит невидимые проверочные данные.

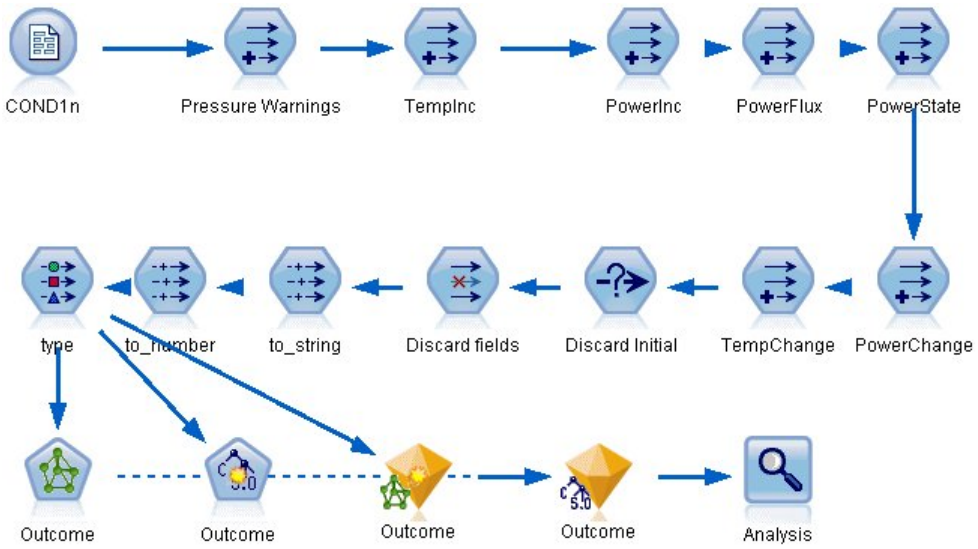


Рисунок 264. Проверка обученной сети

4. Откройте узел Анализ и нажмите кнопку Выполнить.

В результате появятся рисунки, отображающие точность обученной нейросети и правила.

Глава 21. Классификация клиентов в сфере телекоммуникаций (Дискриминантный анализ)

Дискриминантный анализ - это статистический метод для классификации записей на основании значений входных полей. Она аналогична линейной регрессии, но логистическая регрессия использует категориальные поля назначения вместо числовых.

Например, предположим, что провайдер связи сегментировал базу своих клиентов по шаблонам использования сервисов, категоризуя клиентов в четыре группы. Если демографические данные можно использовать для прогноза состава группы, то можно настроить предложения по отдельным возможным заказчикам.

Этот пример использует поток *telco_custcat_discriminant.str*, в котором используется файл данных *telco.sav*. Эти файлы находятся в каталоге *Demos* любой установки IBM SPSS Modeler. Доступ к ним можно получить из группы программ IBM SPSS Modeler в меню Пуск Windows. Файл *telco_custcat_discriminant.str* находится в каталоге *streams*.

Этот пример фокусируется на использовании демографических данных для предсказания паттернов использования. У поля назначения *категория клиента* есть четыре возможных значения, соответствующих четырем группам клиентов:

Значение	**Метка**
1	Базовое обслуживание
2	Интернет-обслуживание
3	Дополнительное обслуживание
4	Полное обслуживание

Создание потока

1. Сначала задайте в свойствах потока вывод меток переменных и значений. Выберите в меню:
   Файл > Свойства потока... > Опции > Общие
2. Убедитесь, что выбрано **Показать метки полей и значений в выводе** и нажмите кнопку **OK**.

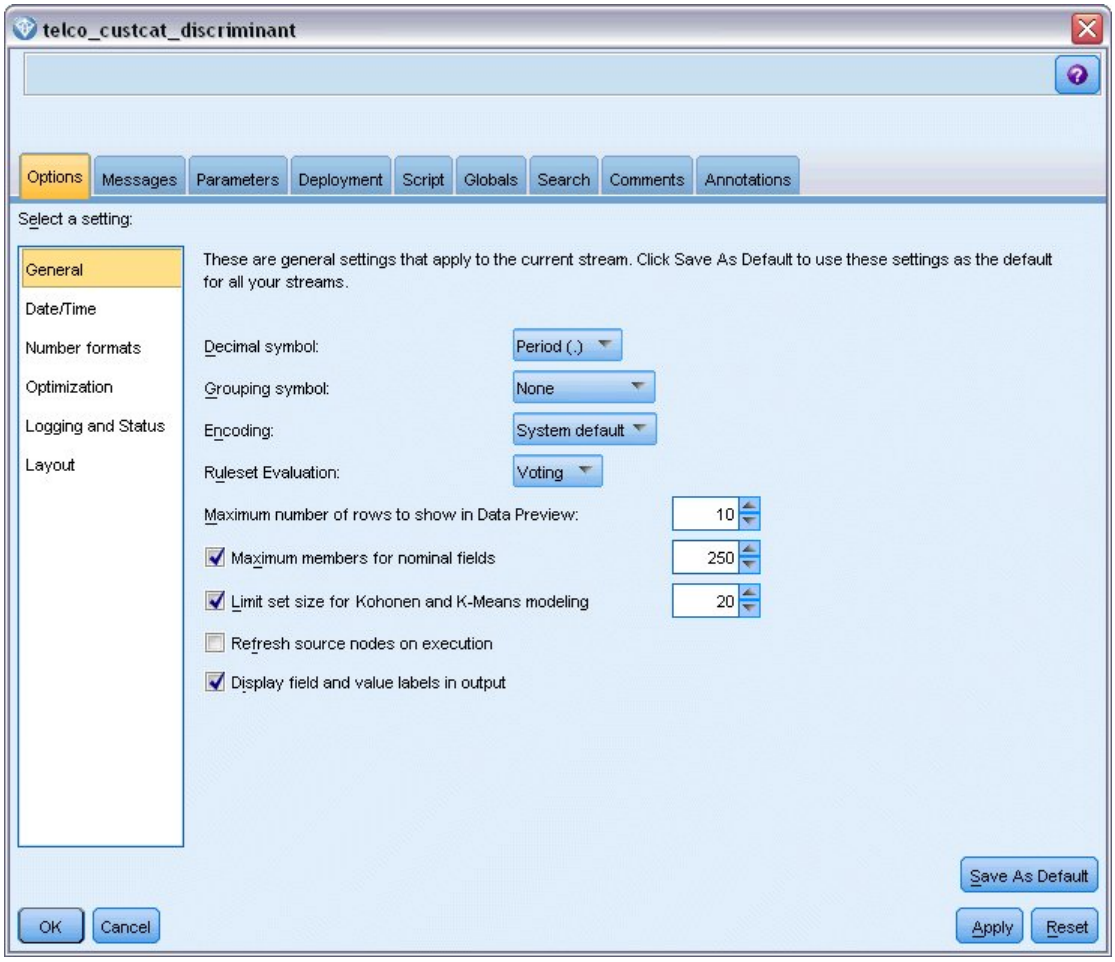


Рисунок 265. Свойства потока

3. Добавьте узел источников файла статистики, указывающий на файл *telco.sav* в папке *Demos*.

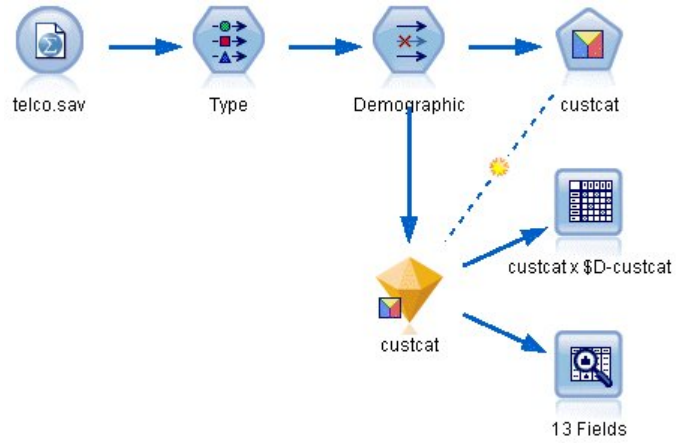


Рисунок 266. Пример потока для классификации клиентов при помощи дискриминантного анализа

a. Добавьте узел Тип и щелкните по **Прочитать значения**, чтобы убедиться, что все типы измерений заданы правильно. Например, большинство полей со значениями 0 и 1 можно считать флагами.

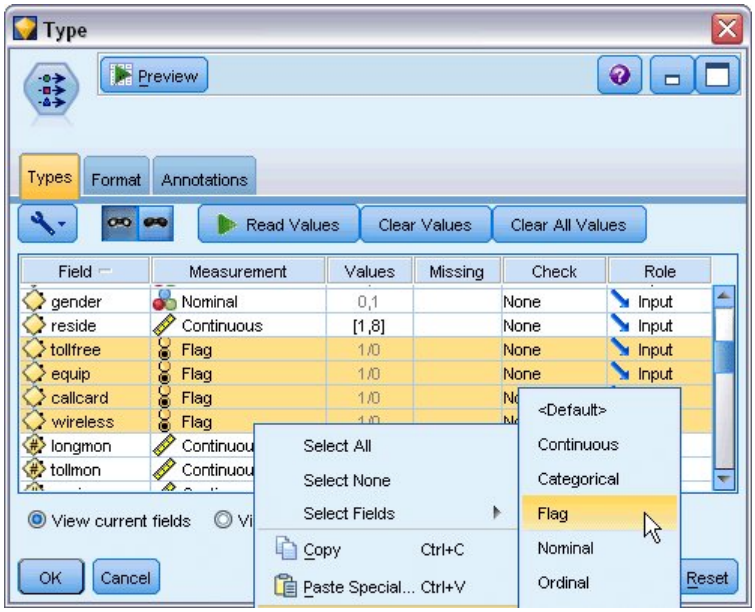


Рисунок 267. Задание уровня измерения для нескольких полей

Совет: чтобы изменить свойства нескольких полей со сходными значениями (например, 0/1), щелкните по заголовку столбца *Значения*, и когда поля будут отсортированы по значению, удерживайте нажатой клавишу Shift и выделите мышью или клавишами со стрелками все поля, которые нужно изменить. Затем можно щелкнуть правой кнопкой мыши по выбранному, чтобы изменить уровень измерения или другие атрибуты выбранных полей.

Заметим, что поле *gender* (пол) правильнее рассматривать не как флаг, а как поле с набором из двух значений, поэтому оставьте Тип измерения для этого поля **Номинальный**.

b. Для поля *custcat* (категория клиента) задайте роль **Назначение**. Для всех остальных полей нужно задать роль **Ввод**.

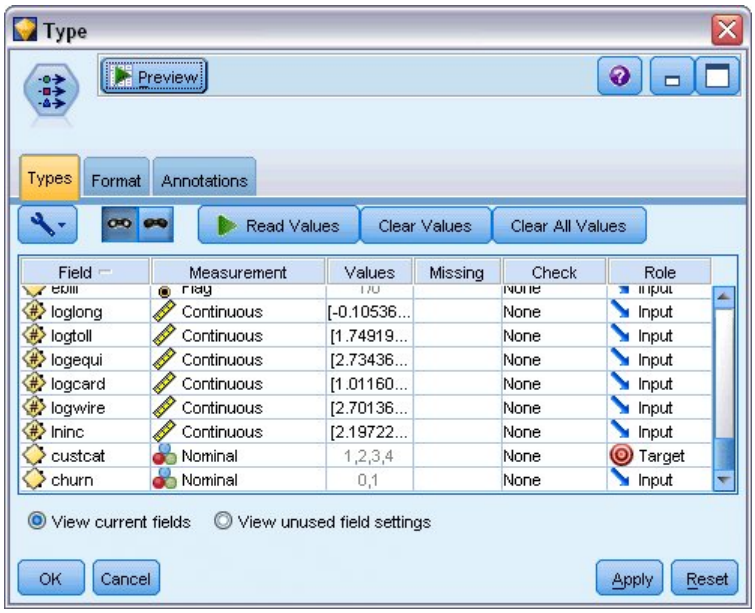


Рисунок 268. Задание роли поля

Поскольку предмет этого примера - демографические показатели, используйте узел Фильтр, чтобы оставить только нужные поля (*region*, *age*, *marital*, *address*, *income*, *ed*, *employ*, *retire*, *gender*, *reside* и *custcat* - регион, возраст, семейное положение, адрес, доход, образование, занятость, на пенсии, пол, место жительства и категория клиента). Остальные поля в ходе данного анализа можно исключить.

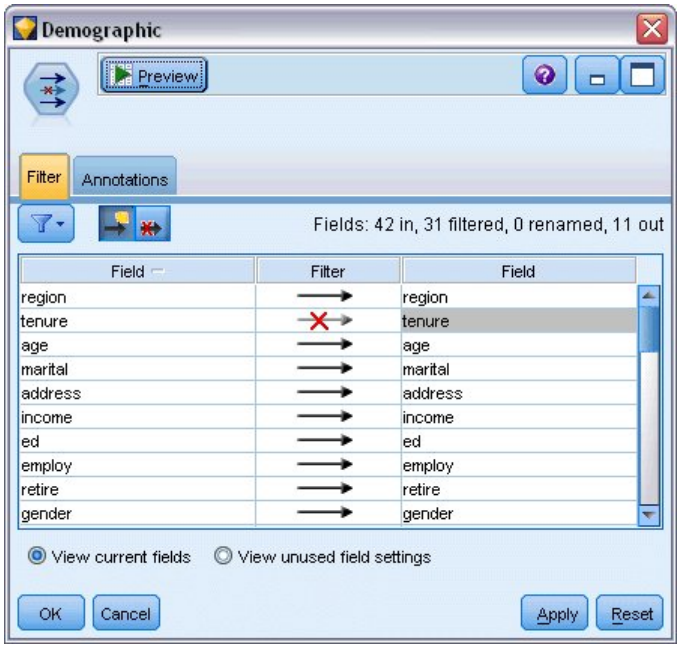


Рисунок 269. Фильтрация демографических полей

(Другой вариант - не исключать эти поля, а задать для них роль **Нет**, или выбрать нужные поля в узле моделирования.)

4. На узле Дискриминантный перейдите на вкладку Модель и выберите **Пошаговый** метод.

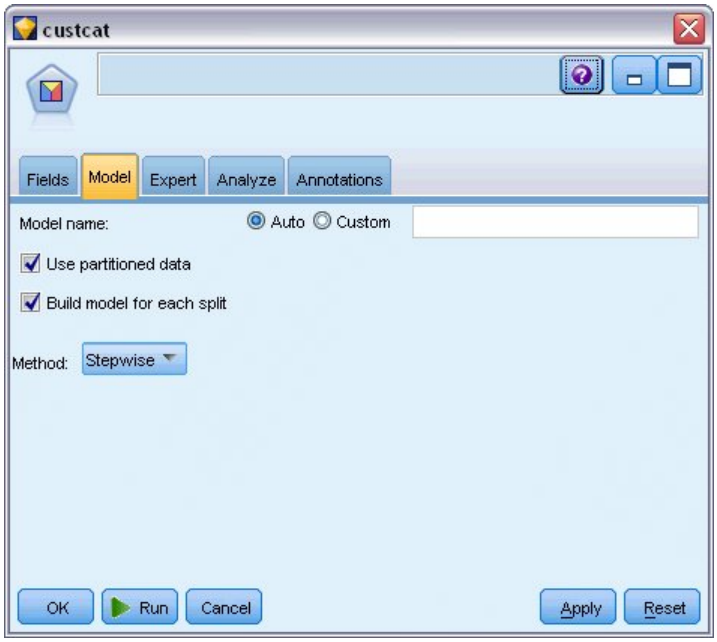


Рисунок 270. Выбор опций модели

5. На вкладке Дополнительно задайте режим **Дополнительно** и нажмите кнопку **Вывод**.

6. В диалоговом окне Расширенный вывод выберите **Сводная таблица**, **Территориальная карта** и **Сводка шагов**, затем нажмите кнопку **OK**.

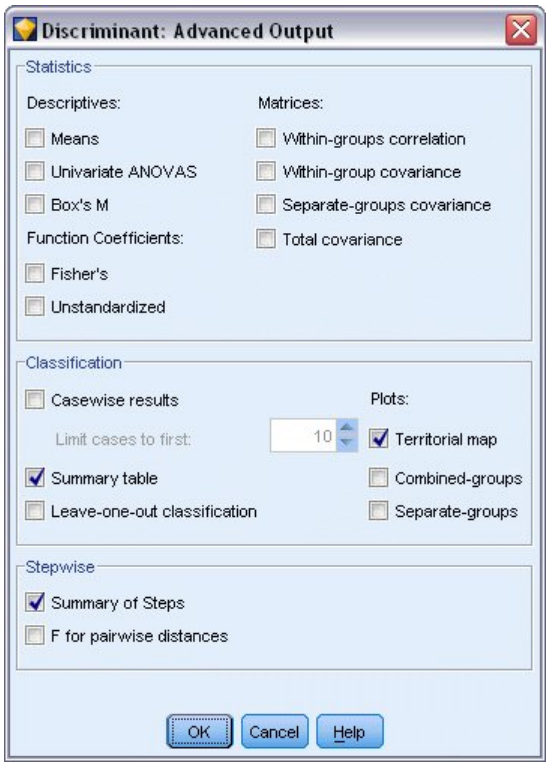


Рисунок 271. Выбор параметров вывода

Изучение модели

1. Нажмите кнопку **Выполнить**, чтобы создать модель, которая будет добавлена в поток и на палитру Модели в верхнем правом углу. Для просмотра сведений о ней щелкните дважды по слепку модели в потоке.

   На вкладке Сводка (среди прочих элементов) будет показано назначение и полный список входных данных (полей предикторов), переданных на рассмотрение.

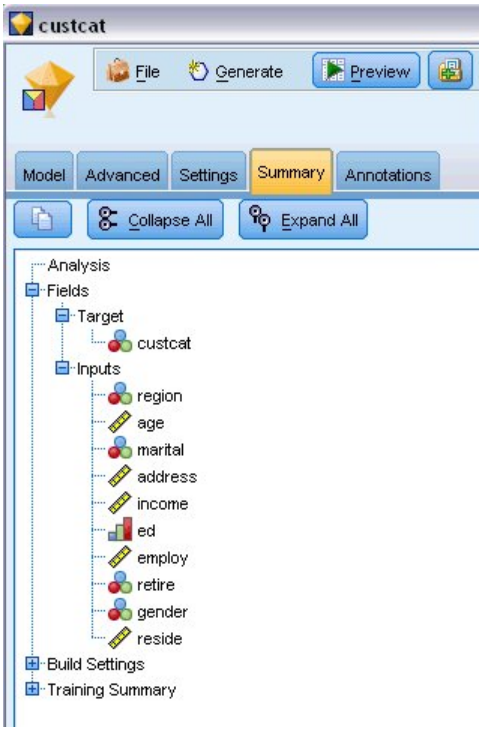


Рисунок 272. Сводка модели, иллюстрирующая целевые и входные поля

Для просмотра подробных результатов дискриминантного анализа:

2. Щелкните по вкладке Дополнительно.
3. Нажмите кнопку "Запустить во внешнем браузере" (под вкладкой Модель), чтобы посмотреть результаты в вашем браузере.

Анализ вывода дискриминантного анализа для классификации телекоммуникационных заказчиков

Пошаговый дискриминантный анализ

Step		Tolerance	Min. Tolerance	F to Enter	Wilks' Lambda
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
	Retired	1.000	1.000	3.005	.991
	Years with current employer	1.000	1.000	16.976	.951
	Gender	1.000	1.000	.373	.999
	Number of people in household	1.000	1.000	3.976	.988

Рисунок 273. Переменные вне анализа, шаг 0

Если у вас есть много предикторов, может оказаться полезным пошаговый метод, автоматически выбирающий "лучшие" переменные для использования в модели. Пошаговый метод начинает выполняться с модели, в состав которой никакие предикторы не входят. На каждом шаге в модель добавляется наибольшее

значение *F-включение*, превышающее критерии включения в модель (по умолчанию 3,84).

Step		Tolerance	Min. Tolerance	F to Enter	Wilks' Lambda
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

Рисунок 274. Переменные вне анализа, шаг 3

У всех переменных, не участвующих в анализе, на последнем шаге значения *F-включение* меньше 3,84, поэтому они больше ничего не добавляют.

Step		Tolerance	F to Remove	Wilks' Lambda
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

Рисунок 275. Переменные для анализа

В приведенной таблице показана статистика для переменных, участвующих в анализе на каждом шаге. *Толерантность* - это доля дисперсии переменной, не объясненной другими независимыми переменными уравнения. Переменная со слишком низкой толерантностью вносит в модель мало информации и может создать вычислительные проблемы.

Значения *F-исключение* полезны для описания того, что случится, если из текущей модели будет удалена переменная (при условии, что другие переменные останутся). *F-исключение* для ввода переменной - то же, что и *F-включение* на предыдущем шаге (показано в таблице Переменные вне анализа).

Предостережение относительно пошаговых методов

Пошаговые методы удобны, но для них существуют ограничения. Следует помнить о том что, поскольку пошаговые методы выбирают модели исключительно на основе преимуществ статистики, возможен выбор предикторов, у которых нет никакой *практической значимости*. Если у вас есть некоторый опыт работы с данными и предположения о том, какие предикторы будут важны, следует использовать это знание и отказаться от пошаговых методов. Однако если у вас множество предикторов, а вы не знаете, с чего начать, выполнение пошагового анализа и настройка выбранной модели будет лучше, чем если модели не будет вовсе.

Проверка подгонки модели

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.198	80.2	80.2	.407
2	.048	19.4	99.6	.214
3	.001	.4	100.0	.031

Рисунок 276. Собственные значения

Почти вся объясняемая моделью дисперсия обуславливается первыми двумя дискриминантными функциями. Подгонка этих функций выполняется автоматически, но третью функцию из-за ее очень малого собственного значения можно спокойно игнорировать.

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 3	.796	227.345	9	.000
2 through 3	.953	47.486	4	.000
3	.999	.929	1	.335

Рисунок 277. Лямбда Уилкса

Лямбда Уилкса подтверждает, что полезны только первые две функции. Для каждого набора функций она проверяет гипотезы о том, что средние указанных функций по всем группам равны. У проверки функции 3 уровень значимости больше 10,0, поэтому эта функция даёт малый вклад в модель.

Матрица структуры

	Function		
	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*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

*. Largest absolute correlation between each variable and any discriminant function

a. This variable not used in the analysis.

Рисунок 278. Матрица структуры

При наличии нескольких дискриминантных функций с помощью звездочки (*) помечается наибольшая абсолютная корреляция каждой переменной с одной из канонических функций. В каждой функции эти помеченные переменные упорядочиваются затем по размеру корреляции.

- Переменная *Уровень образования* наиболее сильно коррелируется с первой функцией, и это единственная переменная, максимально сильно коррелируемая с этой функцией.

- Переменные *Стаж в компании*, *Возраст в годах*, *Семейный доход в тысячах*, *Проживание по текущему адресу*, *Пенсионер* и *Пол* наиболее сильно коррелируются со второй функцией, хотя переменные *Пол* и *Пенсионер* коррелируются слабее, чем другие. С помощью остальных переменных эта функция помечается как функция "стабильности".
- *Число людей в домашнем хозяйстве* и *Семейное положение* наиболее сильно коррелируют с третьей дискриминантной функцией, но эта функция бесполезна, следовательно, это практически бесполезные предикторы.

Территориальная карта

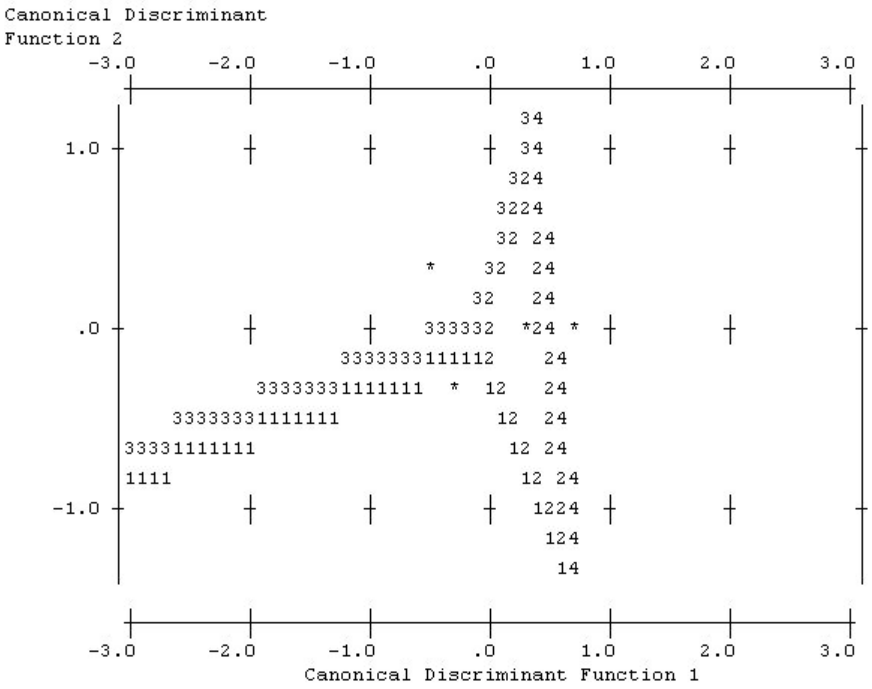


Рисунок 279. Территориальная карта

Территориальная карта помогает изучать взаимосвязи между группами и дискриминантными функциями. В сочетании с результатами матрицы структуры она дает графическую интерпретацию взаимосвязи между предикторами и группами. Первая функция, показанная на вертикальной оси, отделяет группу 4 (клиентов *Полное обслуживание*) от остальных. Поскольку *Уровень образования* сильно положительно коррелирует с первой функцией, это дает основание полагать, что клиенты группы *Полное обслуживание* в целом имеют наибольший уровень образования. Вторая функция разделяет группы 1 и 3 (клиентов *Базовое обслуживание* и *Дополнительное обслуживание*). Стаж и возраст клиентов группы *Дополнительное обслуживание*, как правило, больше, чем у клиентов группы *Базовое обслуживание*. Заказчики группы *Интернет-обслуживание* не отделяются в значительной мере от остальных, хотя карта предполагает, что у них, как правило, хорошее образование при среднем опыте работы.

В целом, близость центроидов групп, помеченных звездочкой (*), к территориальным линиям говорит о том, что разделение между всеми группами не очень сильное.

На карту нанесены только первые две дискриминантные функции, но поскольку третья функция оказалась достаточно несущественной, эта территориальная карта предлагает исчерпывающее представление дискриминантной модели.

Результаты классификации

			Predicted Group Membership				
		Customer category	Basic service	E-service	Plus service	Total service	Total
Original	Count	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% of original grouped cases correctly classified.

Рисунок 280. Результаты классификации

Из показателя лямбда Уилкса вы знаете, что ваша модель дает результаты лучше, чем приблизительная оценка, но чтобы определить, насколько лучше, нужно обратиться к результатам классификации. С учетом данных наблюдений "пустая" модель (то есть модель без предикторов) будет классифицировать всех клиентов в модальную группу *Дополнительное обслуживание*. Поэтому пустая модель будет правильной в 281/1000 = 28,1% случаев. Ваша модель дает на 11,4% больше, то есть 39,5% клиентов. В частности, ваша модель получает лучшие результаты при идентификации клиентов *Полное обслуживание*. Однако она выполняет исключительно плохо задание по классификации клиентов *Интернет-обслуживание*. Для разделения этих клиентов, возможно, потребуется найти другой предиктор.

Итог

Вы создали дискриминантную модель, классифицирующую клиентов в одну из заранее заданных групп "использования услуг" на основе демографической информации от каждого клиента. При помощи матрицы и территориальной карты вы определили, какие переменные наиболее полезны для сегментирования базы ваших клиентов. И наконец, результаты этой классификации показывают, что при классификации клиентов группы *Интернет-обслуживание* полученная модель работает неважно. Требуется провести дополнительные исследования, чтобы определить другую предикторную переменную, лучше классифицирующую этих клиентов, но в зависимости от того, что вы стремитесь предсказать, полученной модели может оказаться вполне достаточно для ваших потребностей. Например, модель может быть достаточно точной, если вы не занимаетесь идентификацией клиентов группы *Интернет-обслуживание*. Возможно, это тот случай, когда обслуживание Интернет-обслуживание служит всего лишь цели привлечь клиентов, принося небольшую прибыль. Если, например, ваша самая высокая прибыть на инвестиции поступает от клиентов группы *Дополнительное обслуживание* или *Полное обслуживание*, полученная модель может дать вам необходимую информацию.

Учтите, что эти результаты основаны только на обучающих данных. Чтобы оценить, насколько хорошо модель обобщается на другие данные, можно применить узел Разбиение, который будет содержать поднабор записей в целях проверки.

Объяснение математических основ методов моделирования, используемых в IBM SPSS Modeler, смотрите в публикации IBM SPSS Modeler: Руководство по алгоритмам. Оно доступно в каталоге *\Documentation* установочного диска.

Глава 22. Анализ данных выживания, цензурированных по интервалам (обобщенные линейные модели)

При анализе данных выживания с цензурированием по интервалам (то есть если точное время исследуемого события неизвестно, но только известно, что оно должно было произойти в указанный интервал) применение модели Кокса к опасностям событий в интервалах приводит к модели дополняющей лог-лог регрессии.

Частичная информация из исследования, предназначенного для сравнения эффективности двух методов лечения для предотвращения рецидива язвы собирается в файле *ulcer_recurrence.sav*. Этот набор данных был представлен и проанализирован в другой работе [1]. С помощью обобщенных линейных моделей можно повторить результаты для моделей дополняющей лог-лог регрессии.

Этот пример использует поток *ulcer_genlin.str*, ссылающийся на файл данных *ulcer_recurrence.sav*. Файл данных находится в папке *Demos*, а файл потока - в подпапке *streams*.

Создание потока

1. Добавьте узел источников файла статистики, указывающий на файл *ulcer_recurrence.sav* в папке *Demos*.

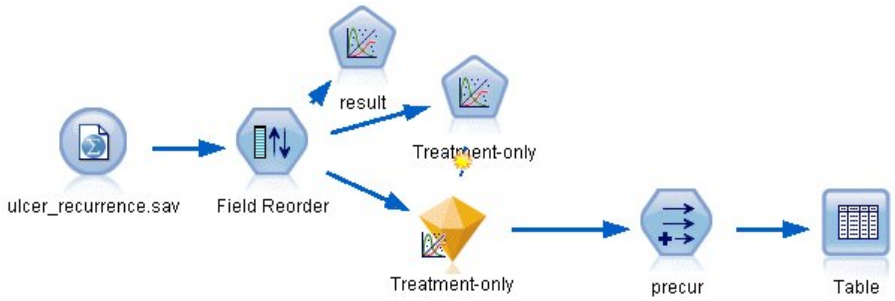


Рисунок 281. Поток примера для прогноза рецидива язвы

2. На вкладке Фильтр узла источника отфильтруйте поля *id* и *time*.

1. Collett, D. 2003. *Modelling survival data in medical research*, 2 ed. Boca Raton: Chapman & Hall/CRC.

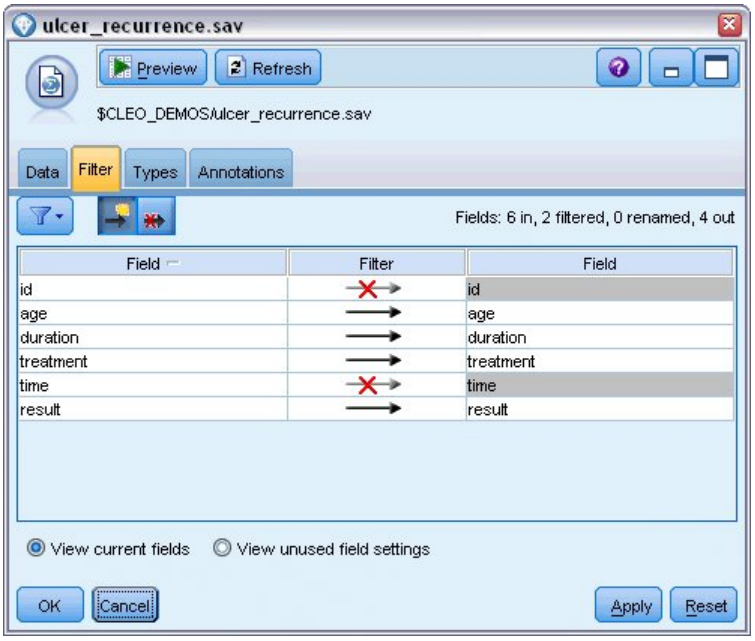


Рисунок 282. Отфильтровать нежелательные поля

3. На вкладке Типы узла источника задайте для поля *result* роль **Назначение** и задайте его шкалу измерения как **Флаговую**. Результат 1 означает рецидив язвы. Для всех остальных полей нужно задать роль **Ввод**.
4. Нажмите кнопку **Чтение данных**, чтобы создать экземпляр данных.

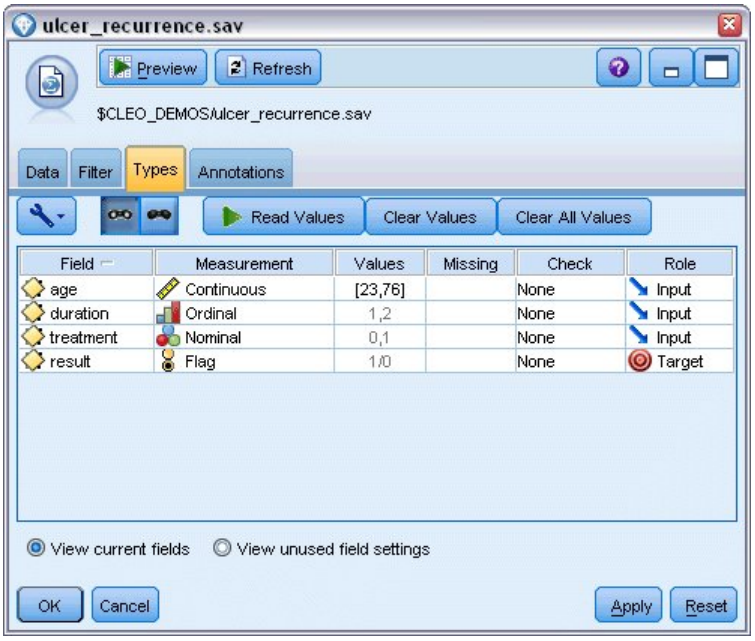


Рисунок 283. Задание роли поля

5. Добавьте узел переупорядочения полей и в качестве порядка входных данных задайте *duration (продолжительность)*, *treatment (лечение)* и *age (возраст)*. Он определит порядок ввода полей в модель и поможет попытаться повторить результаты Коллета.

Рисунок 284. Переупорядочение полей для их ввода в модель нужным образом

6. Присоедините к узлу источника узел Обобщенная линейная регрессия; на узле обобщенной линейной регрессии щелкните по вкладке **Модель**.
7. В качестве эталонной категории для назначения выберите **Первая (с наименьшим значением)**. Это означает, что вторая категория будет представлять собой исследуемое событие, а ее влияние на модель - заключаться в интерпретации оценок параметров. Непрерывный предиктор с положительным коэффициентом означает повышенную вероятность рецидива при увеличении значений предиктора; категории номинального предиктора с более крупными коэффициентами означают повышенную вероятность рецидива относительно других категорий набора.

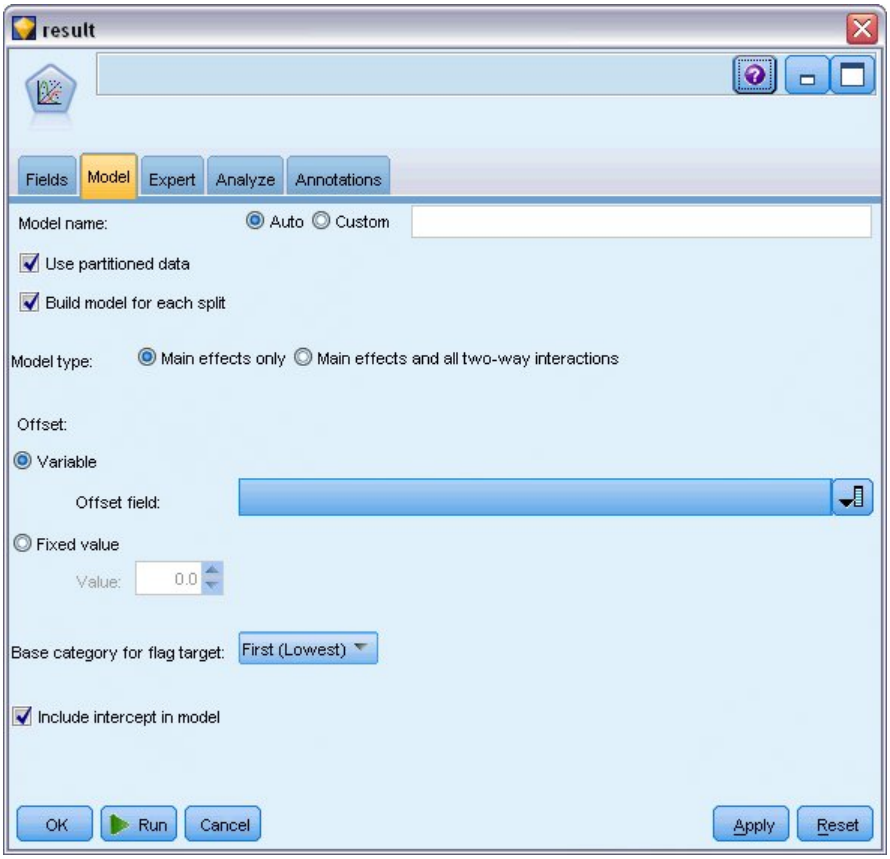


Рисунок 285. Выбор опций модели

8. Щелкните по вкладке **Эксперт** и выберите **Эксперт**, чтобы активировать экспертные опции моделирования.
9. В качестве распределения выберите **Биномиальное**, а в качестве функции связи - **Дважды логарифмическая**.
10. В качестве метода для оценки параметра масштаба выберите **Фиксированное значение** и оставьте значение по умолчанию 1,0.
11. Выберите **По убыванию** в качестве порядка категорий для коэффициентов. Это значит, что первая категория каждого фактора будет его эталонной категорией; влияние этого выбора на модель заключается в интерпретации оценок параметров.

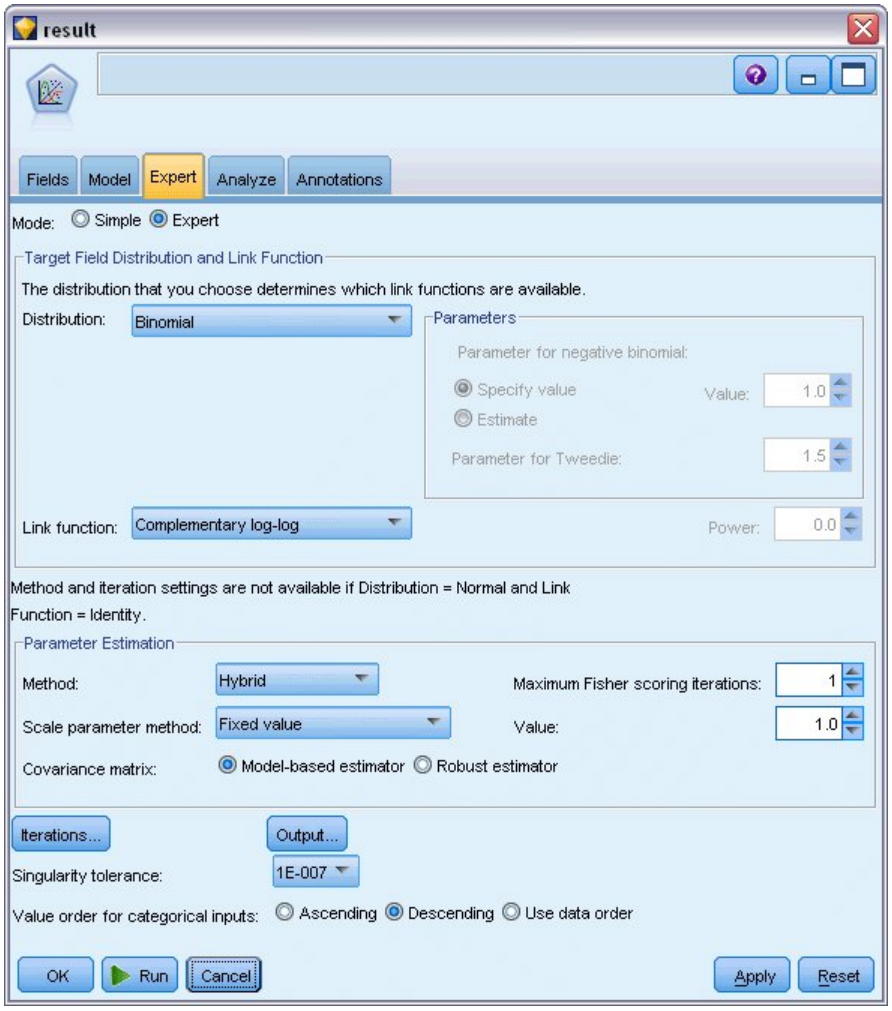


Рисунок 286. Выбор дополнительных опций

12. Запустите поток, чтобы создать слепок модели, который будет добавлен на холст потока, а также на палитру моделей в верхнем правом углу. Для просмотра подробностей модели щелкните правой кнопкой мыши по слепку и выберите **Изменить** или **Просмотреть**.

Проверки эффектов модели

Source	Type III Wald Chi-Square	df	Sig.
(Intercept)	.536	1	.464
duration	.003	1	.958
treatment	.382	1	.537
age	.358	1	.550

Dependent Variable: Result
Model: (Intercept), duration, treatment, age

Рисунок 287. Тестирование эффектов моделей для модели главных эффектов

Никакие из эффектов приведенной модели статистически не значимы; однако любые наблюдаемые различия в эффектах лечения представляют собой клинический интерес, поэтому мы выполним подгонку упрощенной модели, членом которой будет только лечение.

Подгонка модели Только лечение

1. На вкладке Поля узла обобщенной линейной модели щелкните по **Использовать пользовательские параметры**.
2. В качестве назначения выберите *result*.
3. В качестве монопольного поля ввода выберите *treatment*.

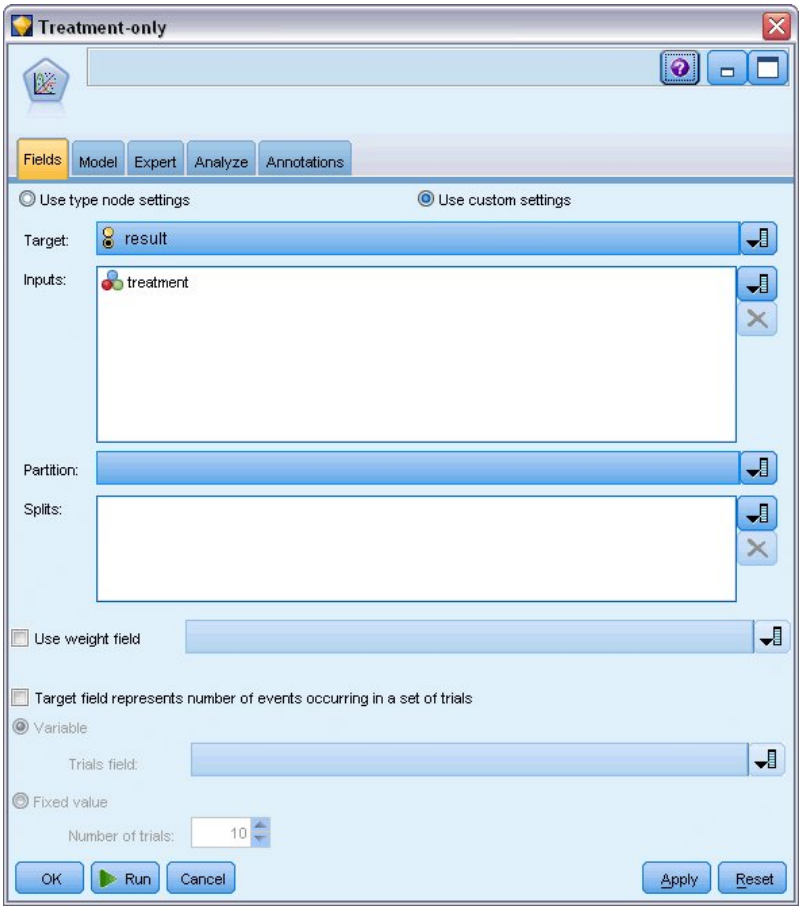


Рисунок 288. Выбор опции полей

4. Запустите поток и откройте полученный слепок модели.

В слепке модели выберите вкладку **Дополнительно** и прокрутите ее вниз до конца.

Оценки параметров

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-1.442	.5012	-2.425	-.460	8.282	1	.004
[treatment=1]	.378	.6288	-.855	1.610	.361	1	.548
[treatment=0]	0[a]	.	.	.	.	.	.
(Scale)	1[b]						

Dependent Variable: Result
Model: (Intercept), treatment

a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

Рисунок 289. Оценки параметров для модели только обработки

Эффект лечения (различие в линейном предикторе между двумя уровнями лечения, то есть коэффициент для *[treatment=1]*) по прежнему статистически не значим, но тем не менее наводит на мысль, что лечение *A [treatment=0]* может оказаться лучше лечения *B [treatment=1]*, поскольку оценка параметра для лечения *B* больше оценки лечения *A* и следовательно связана с увеличенной вероятностью рецидивов в первые 12 месяцев. Линейный предиктор (свободный член + эффект лечения) - это оценка $\log(-\log(1-P(recur_{12,t}))$, где $P(recur_{12,t})$ - вероятность рецидива в течение 12 месяцев для лечения t(=*A* или *B*). Эти предсказанные вероятности генерируются для каждого наблюдения в наборе данных.

Прогноз вероятностей рецидива и выживания

Рисунок 290. Опции параметров узла извлечения

1. Для каждого пациента скоринг модели дает предсказанный результат и вероятность этого результата предсказания. Чтобы просмотреть предсказанные вероятности рецидивов, скопируйте сгенерированную модель на палитру и присоедините узел извлечения.
2. На вкладке Параметры в качестве узла извлечения введите `precur`.
3. Выберите для извлечения вариант **С условием**.
4. Нажмите кнопку с калькулятором, чтобы открыть Построитель выражений как условие **If**.

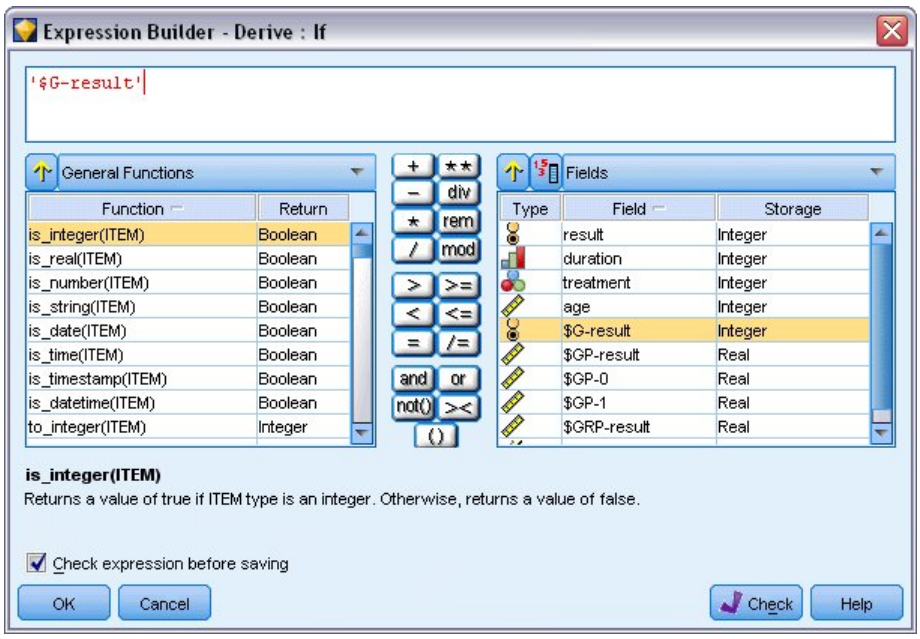


Рисунок 291. Узел извлечения: Построитель выражений для условия If

5. Вставьте в выражение поле *$G-result*.
6. Щелкните по **OK**.

   Поле извлечения *precur* будет принимать значение выражения **Then**, если значение поля *$G-result* равно 1, и значение выражения **Else**, если значение этого поля равно 0.

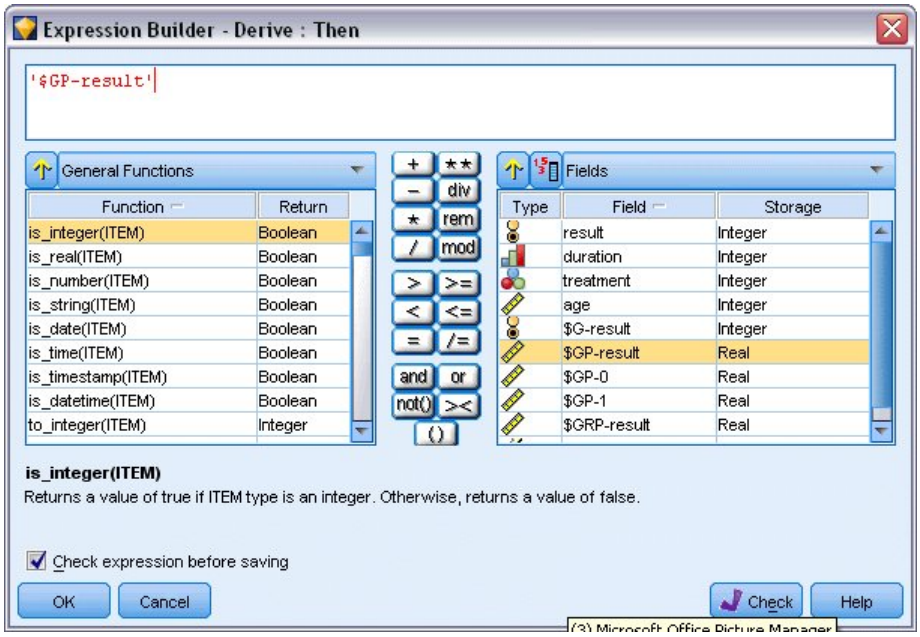


Рисунок 292. Узел извлечения: Построитель выражений для выражения Then

7. Нажмите кнопку с калькулятором, чтобы открыть Построитель выражений как выражение **Then**.
8. Вставьте поле *$GP-result* в выражение.
9. Щелкните по **OK**.

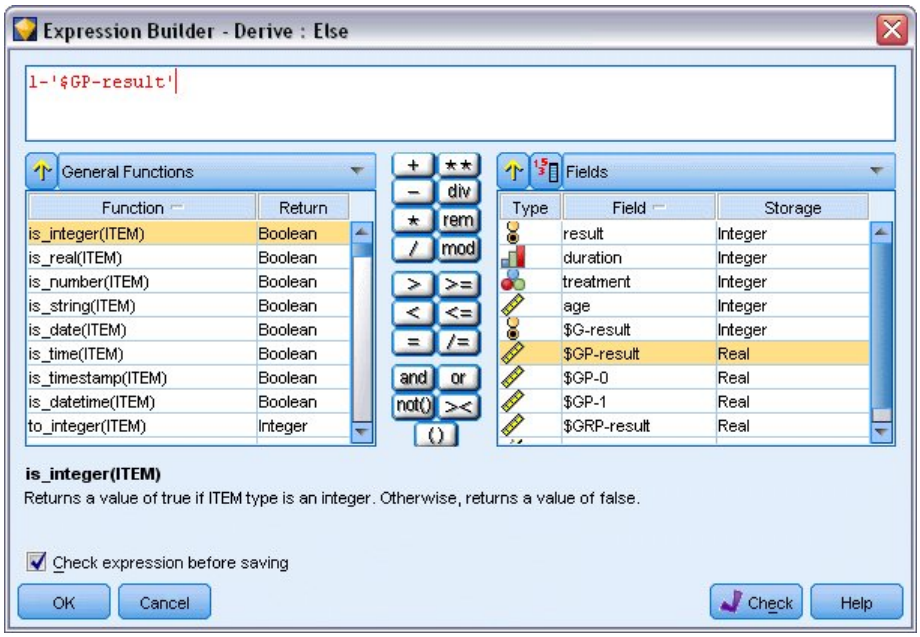
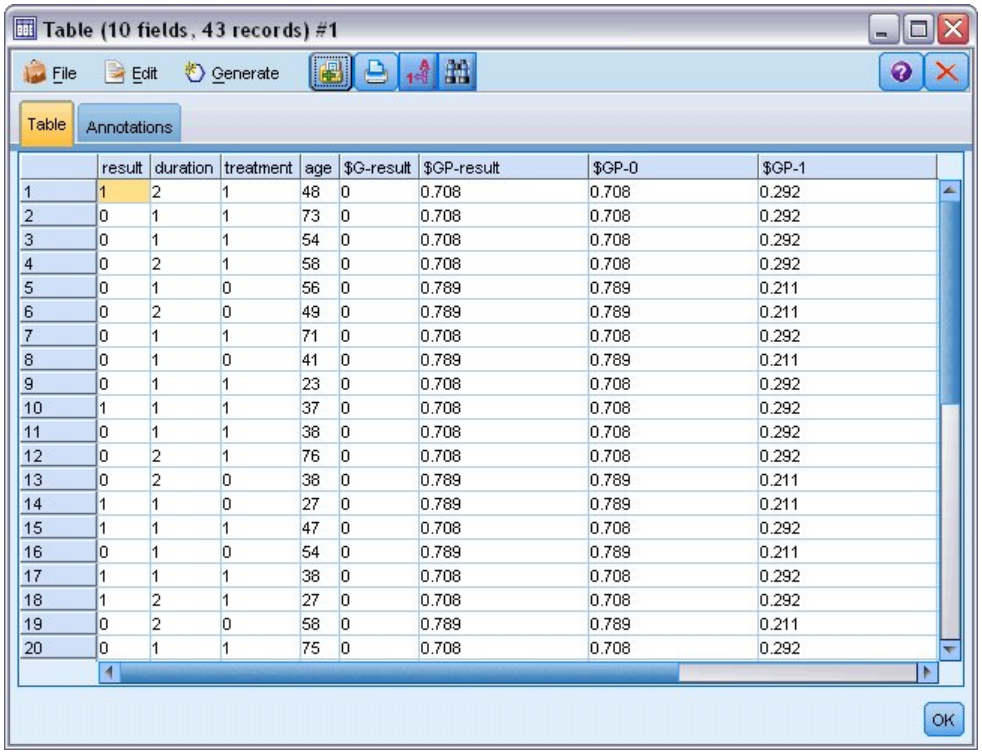


Рисунок 293. Узел извлечения: Построитель выражений для выражения Else

10. Нажмите кнопку с калькулятором, чтобы открыть Построитель выражений как выражение **Else**.
11. Введите в выражении `1-`, а затем вставьте в выражение поле *$GP-result*.
12. Щелкните по **OK**.

Рисунок 294. Опции параметров узла извлечения

13. Присоедините к узлу извлечения узел таблицы и выполните его.



	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

Рисунок 295. Предсказанная вероятность

Тому, что у пациентов, назначенных на лечение *A*, случится рецидив в первые 12 месяцев, соответствует оцененная вероятность 0,211; лечению *B* соответствует вероятность 0,292. Имейте в виду, что $1-P(recur_{12,\ t})$ - это вероятность выживания в течение 12 месяцев, которая аналитикам выживания может быть более интересна.

Моделирование вероятности рецидивов по периодам

Проблема с моделью при данных обстоятельствах состоит в игнорировании ею информации, собранной при первом исследовании; то есть в отсутствии рецидивов у многих пациентов в течение первых шести месяцев. "Лучшая" модель смоделировала бы бинарный отклик, записывающий, происходит ли событие в течение каждого интервала. Подгонка этой модели требует реорганизации исходного набора данных, который можно найти в файле *ulcer_recurrence_recoded.sav*. Этот файл содержит две дополнительные переменные:

- Переменная *Период* записывает, соответствует ли случай первому или второму периоду исследования.
- Переменная *Результаты по периодам* записывает, был ли рецидив у данного пациента в течение данного периода.

Каждый исходный случай (пациент) вносит только один случай на интервал, где тот остается в наборе рисков. Так, например, пациент 1 вносит два случая: один для первого периода исследования, в котором не было рецидива, а другой для второго периода исследования, в котором был рецидив. С другой стороны, пациент 10 вносит один случай, поскольку рецидив был записан в первом периоде. Пациенты 16, 28 и 34 были исключены из исследования по истечении шести месяцев, и поэтому они вносят в новый набор данных только один случай.

1. Добавьте узел источников файла статистики, указывающий на файл *ulcer_recurrence_recoded.sav* в папке *Demos*.

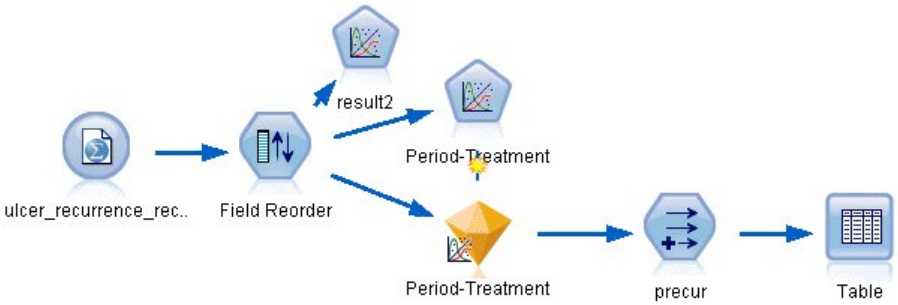


Рисунок 296. Поток примера для прогноза рецидива язвы

2. На вкладке Фильтр узла источника отфильтруйте поля *id*, *time* и *result*.

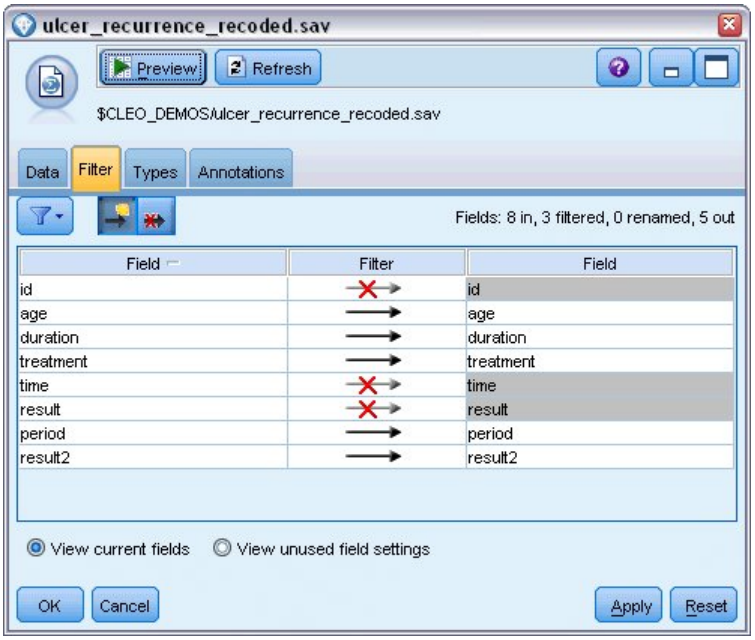


Рисунок 297. Отфильтровать нежелательные поля

3. На вкладке Типы узла источника задайте для поля *result2* роль **Назначение** и задайте его шкалу измерения как **Флаговую**. Для всех остальных полей нужно задать роль **Ввод**.

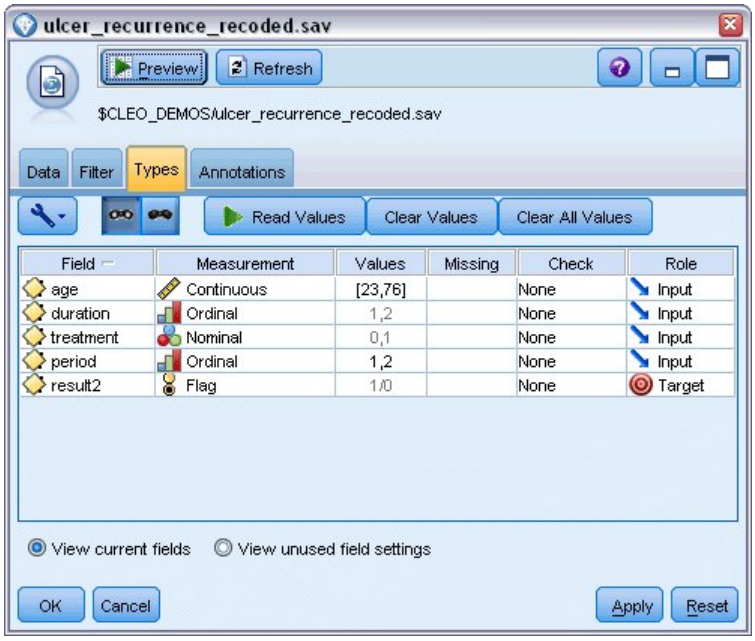


Рисунок 298. Задание роли поля

4. Добавьте узел переупорядочения полей и в качестве порядка входных данных задайте *period (период)*, *duration (продолжительность)*, *treatment (лечение)* и *age (возраст)*. Если сделать *period* первым вводимым полем (и не включать в модель свободный член), будет возможна подгонка полного набора фиктивных переменных для захвата эффектов периодов.

Рисунок 299. Переупорядочение полей для их ввода в модель нужным образом

5. На узле Обобщенная линейная регрессия щелкните по вкладке **Модель**.

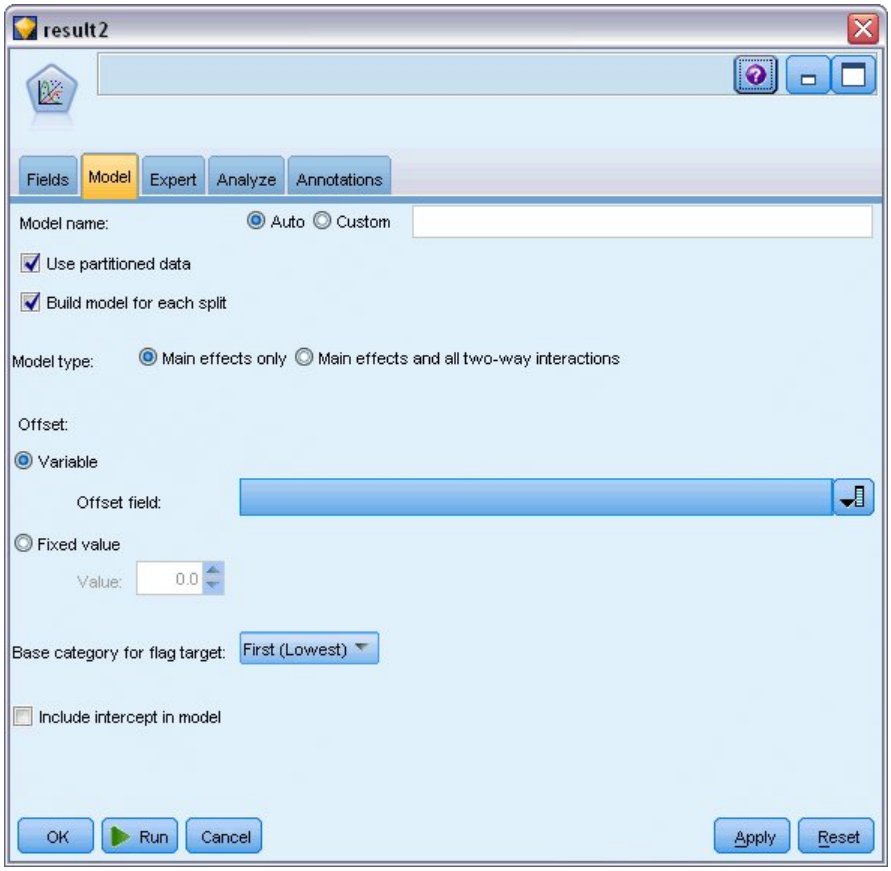


Рисунок 300. Выбор опций модели

6. В качестве эталонной категории для назначения выберите **Первая (с наименьшим значением)**. Это означает, что вторая категория будет представлять собой исследуемое событие, а ее влияние на модель - состоять в интерпретации оценок параметров.
7. Отмените выбор **Включить в модель свободный член**.
8. Щелкните по вкладке **Эксперт** и выберите **Эксперт**, чтобы активировать экспертные опции моделирования.

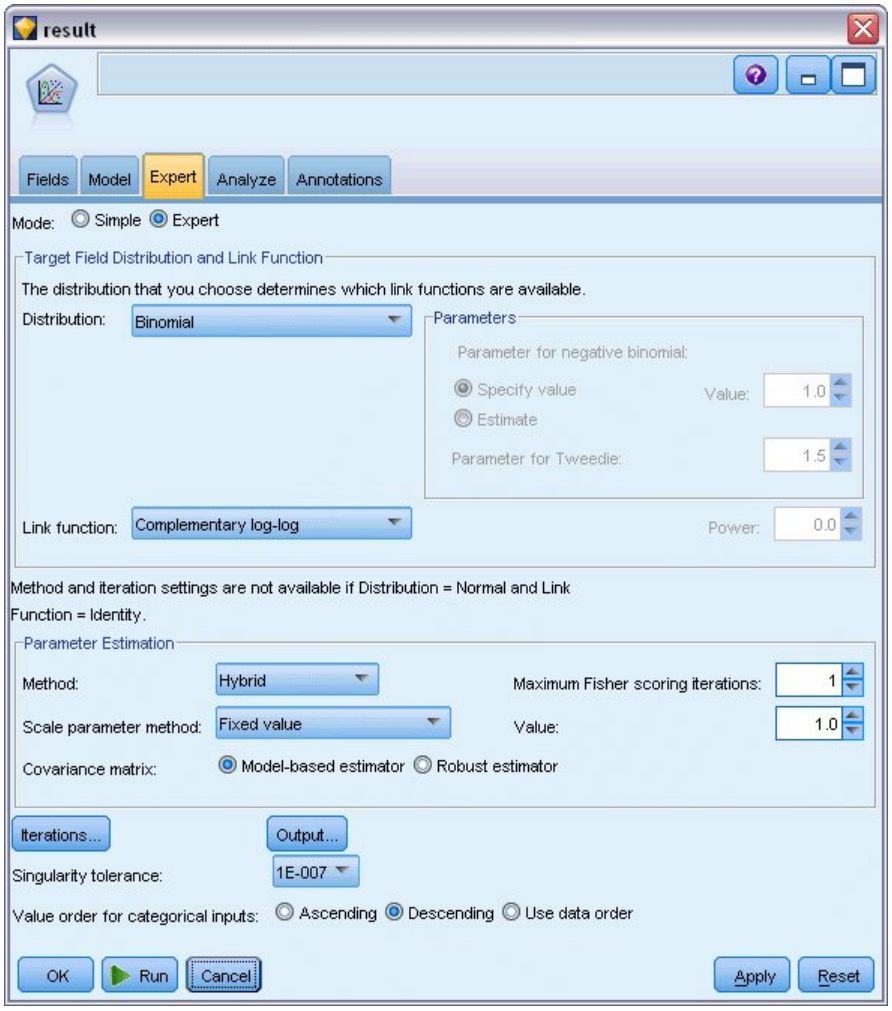


Рисунок 301. Выбор дополнительных опций

9. В качестве распределения выберите **Биномиальное**, а в качестве функции связи - **Дважды логарифмическая**.
10. В качестве метода для оценки параметра масштаба выберите **Фиксированное значение** и оставьте значение по умолчанию 1,0.
11. Выберите **По убыванию** в качестве порядка категорий для коэффициентов. Это значит, что первая категория каждого фактора будет его эталонной категорией; влияние этого выбора на модель заключается в интерпретации оценок параметров.
12. Запустите поток, чтобы создать слепок модели, который будет добавлен на холст потока, а также на палитру моделей в верхнем правом углу. Для просмотра подробностей модели щелкните правой кнопкой мыши по слепку и выберите **Изменить** или **Просмотреть**.

Проверки эффектов модели

Source	Type III		
	Wald Chi-Square	df	Sig.
period	.464	1	.496
duration	.000	1	.988
treatment	.117	1	.732
age	.314	1	.575

Dependent Variable: Result by period
Model: period, duration, treatment, age

Рисунок 302. Тестирование эффектов моделей для модели главных эффектов

Никакие из эффектов приведенной модели статистически не значимы; однако любые наблюдаемые различия в эффектах периодов и лечения представляют собой клинический интерес, поэтому мы выполним подгонку упрощенной модели, в которой используются только эти члены.

Подгонка упрощённой модели

1. На вкладке Поля узла обобщенной линейной модели щелкните по **Использовать пользовательские параметры**.
2. В качестве назначения выберите *result2*.
3. В качестве полей ввода выберите *period* и *treatment*.

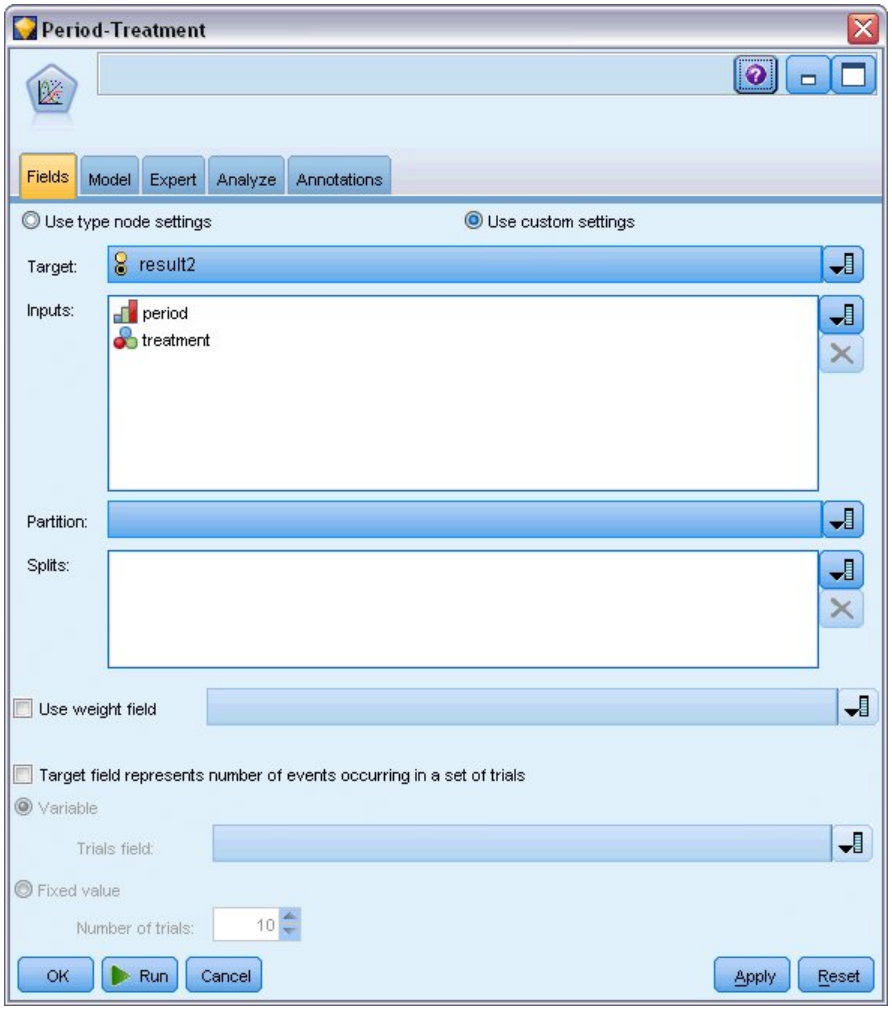


Рисунок 303. Выбор опции полей

4. Выполните приведённый узел и найдите сгенерированную модель, после чего скопируйте эту сгенерированную модель на палитру, присоедините узел таблицы и выполните ее.

Оценки параметров

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	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]	.	.	.	.	.	.
(Scale)	1[b]						

Dependent Variable: Result by period
Model: period, treatment

a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

Рисунок 304. Оценки параметров для модели только обработки

Эффект лечения по прежнему статистически не значим, но тем не менее наводит на мысль, что лечение *A* может оказаться лучше лечения *B*, поскольку оценка параметра для лечения *B* связана с увеличенной вероятностью рецидивов в первые 12 месяцев. Значения периодов статистически значимо отличаются от 0, но это обусловлено тем, что не подогнан свободный член. Эффект периода (различие между значениями линейного периода для *[period=1]* и *[period=2]*) статистически не значим, как показывает тестирование эффектов модели. Линейный предиктор (эффект периода + эффект лечения) - это оценка $\log(-\log(1-P(recur_{p,t}))$, где $P(recur_{p,t})$ - вероятность рецидива в период p(=1 или 2, представляющий шесть или 12 месяцев) для лечения t(=*A* или *B*). Эти предсказанные вероятности генерируются для каждого наблюдения в наборе данных.

Прогноз вероятностей рецидива и выживания

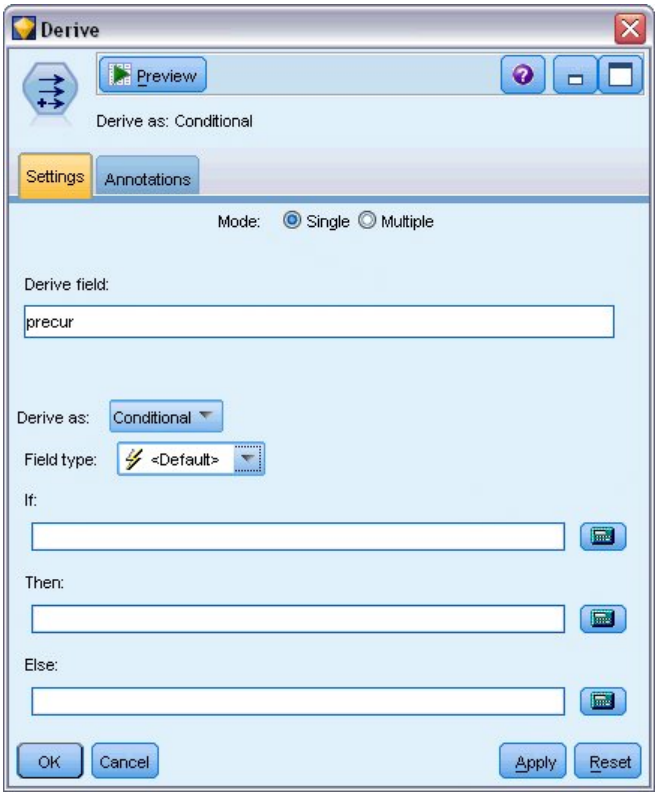


Рисунок 305. Опции параметров узла извлечения

1. Для каждого пациента скоринг модели дает предсказанный результат и вероятность этого результата предсказания. Чтобы просмотреть предсказанные вероятности рецидивов, скопируйте сгенерированную модель на палитру и присоедините узел извлечения.
2. На вкладке Параметры в качестве узла извлечения введите `precur`.
3. Выберите для извлечения вариант **С условием**.
4. Нажмите кнопку с калькулятором, чтобы открыть Построитель выражений как условие **If**.

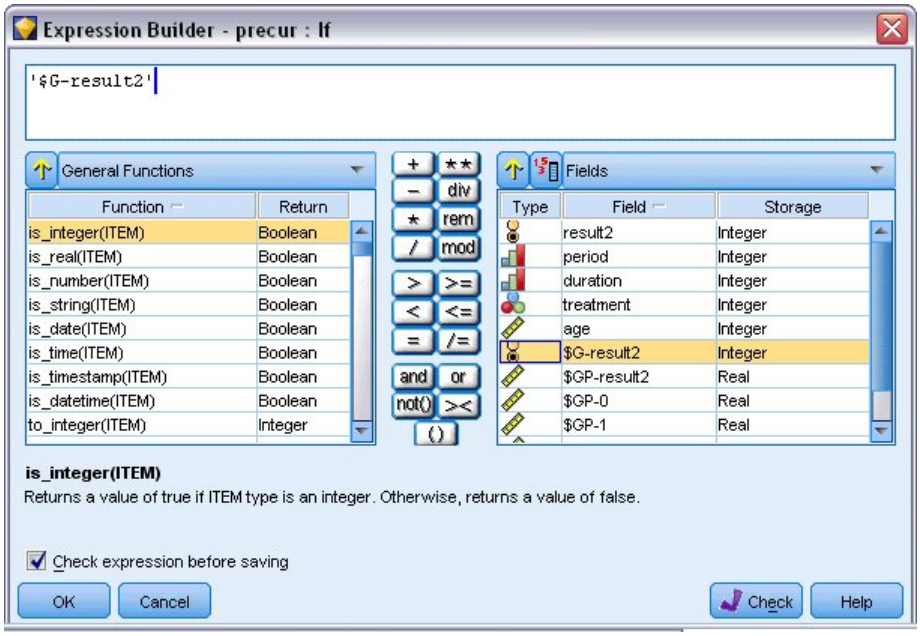


Рисунок 306. Узел извлечения: Построитель выражений для условия If

5. Вставьте в выражение поле *$G-result2*.
6. Щелкните по **OK**.

   Поле извлечения *precur* будет принимать значение выражения **Then**, если значение поля *$G-result2* равно 1, и значение выражения **Else**, если значение этого поля равно 0.

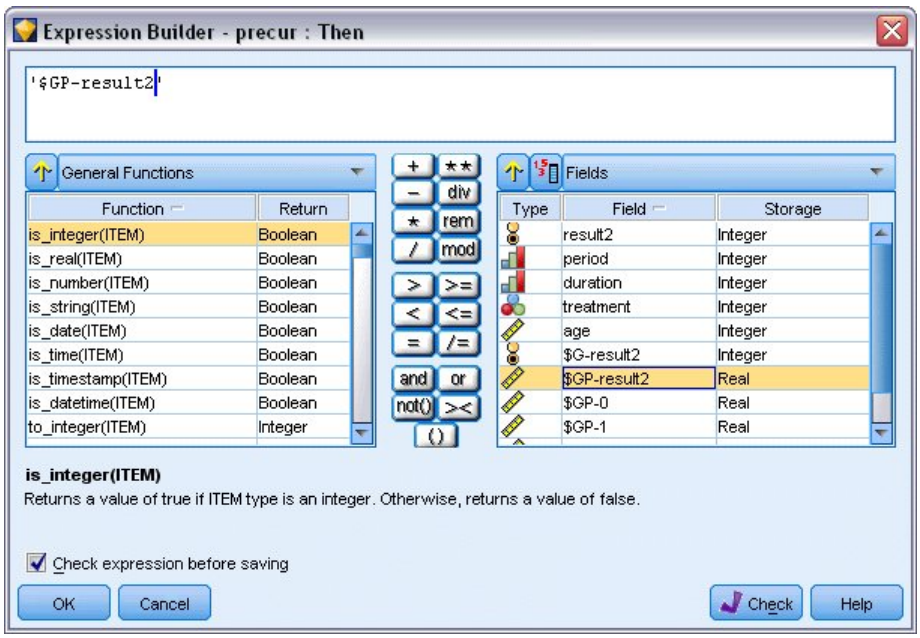


Рисунок 307. Узел извлечения: Построитель выражений для выражения Then

7. Нажмите кнопку с калькулятором, чтобы открыть Построитель выражений как выражение **Then**.
8. Вставьте поле *$GP-result2* в выражение.
9. Щелкните по **OK**.

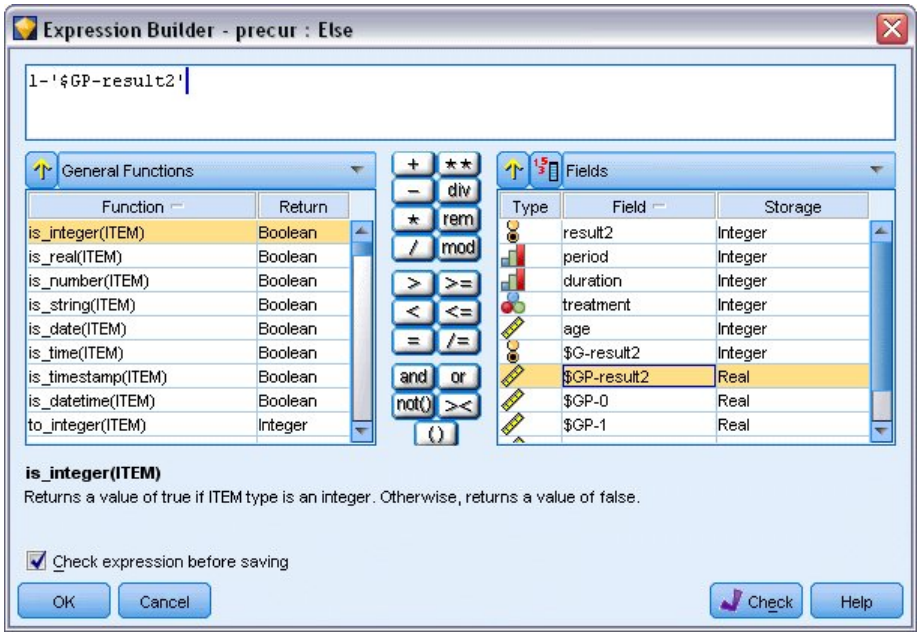


Рисунок 308. Узел извлечения: Построитель выражений для выражения Else

10. Нажмите кнопку с калькулятором, чтобы открыть Построитель выражений как выражение **Else**.
11. Введите в выражении `1-`, а затем вставьте в выражение поле *$GP-result2*.
12. Щелкните по **OK**.

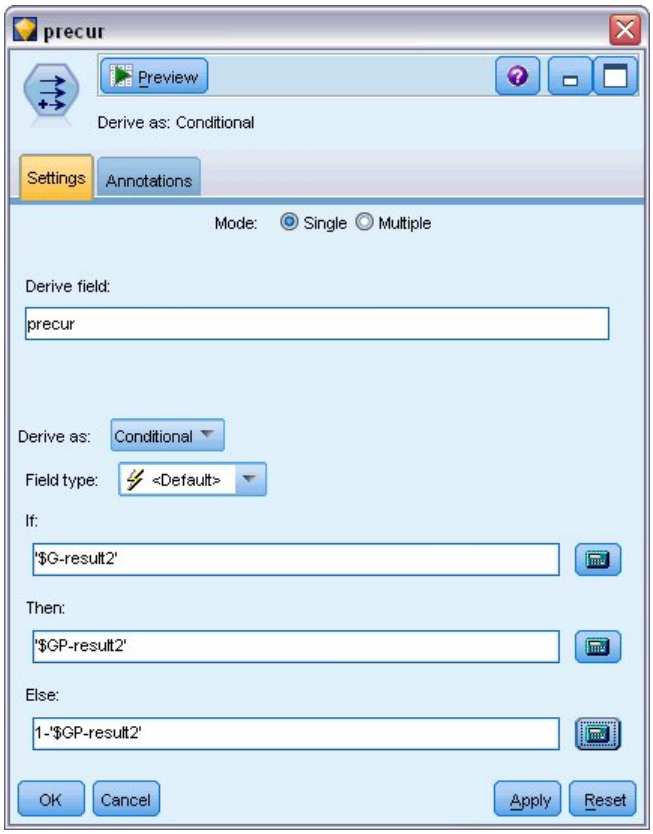
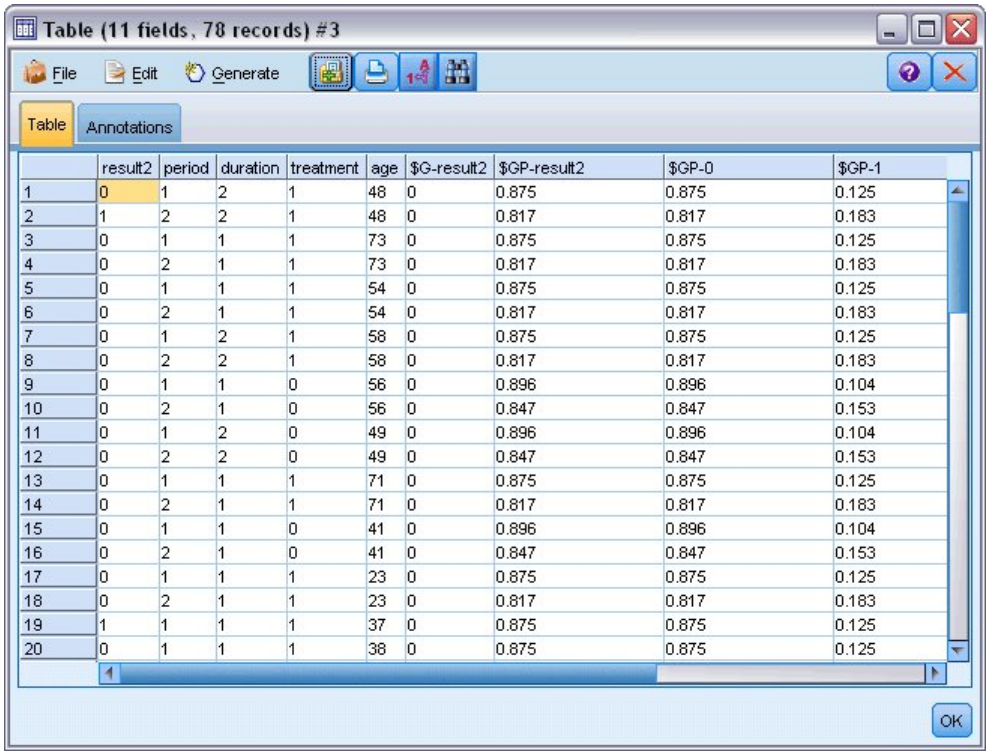


Рисунок 309. Опции параметров узла извлечения

13. Присоедините к узлу извлечения узел таблицы и выполните его.



	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

Рисунок 310. Предсказанная вероятность

Таблица 3. Оценка вероятностей рецидива

Обработка	**6 месяцев**	**12 месяцев**
A	0,104	0,153
D	0,125	0,183

Исходя из оцененных вероятностей рецидивов, вероятность выживания в течение 12 месяцев может быть оценена как $1-(P(recur_{1,t}) + P(recur_{2,t})\times(1-P(recur_{1,t})))$; соответственно для каждого лечения:

A: 1 – (0,104 + 0,153*0,896) = 0,759

B: 1 – (0,125 + 0,183*0,875) = 0,715

что снова демонстрирует статистически незначимую поддержку лечения *A* как лучшего лечения.

Итог

С помощью обобщенных линейных моделей вы выполнили подгонку ряда моделей дополняющей лог-лог регрессии для данных выживания, цензурированных по интервалам. Несмотря на наличие некоторых аргументов в пользу выбора лечения *A*, для достижения статистически значимого результата может потребоваться исследование большего объема. Однако есть несколько дальнейших путей исследования существующих данных.

- Может оказаться целесообразной повторная подгонка модели с эффектами взаимодействия, в частности, между факторами *Период* и *Группа испытуемых*.

Объяснение математических основ методов моделирования, используемых в IBM SPSS Modeler, смотрите в публикации *IBM SPSS Modeler: Руководство по алгоритмам*.

Связанные процедуры

Процедура Обобщенные линейные модели - мощный инструмент для подгонки самых разнообразных моделей.

- Процедура Обобщенные уравнения оценки расширяет обобщенную линейную модель, разрешая повторные измерения.
- Процедура Линейные смешанные модели позволяет выполнять подгонку моделей для количественных зависимых переменных со случайными компонентными и/или повторными измерениями.

Рекомендуемое чтение

Дополнительную информацию об обобщенных линейных моделях смотрите в следующих текстовых источниках:

Cameron, A. C., and P. K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press. Dobson, A. J. 2002. *An Introduction to Generalized Linear Models*, 2 ed. Boca Raton, FL: Chapman & Hall/CRC. Hardin, J. W., and J. M. Hilbe. 2003. *Generalized Linear Models and Extension*. Station, TX: Stata Press. McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

Глава 23. Использование регрессии Пуассона для анализа частоты повреждений судов (обобщенные линейные модели)

Обобщенную линейную модель можно использовать для подгонки регрессии Пуассона для анализа дискретных данных. Например, набор данных, представленный и проанализированный в другой работе [2], относится к повреждениям, причиненным грузовым судам волнами. Количество инцидентов можно смоделировать в виде распределения Пуассона при заданных значениях предикторов; полученная модель может помочь определить, суда каких типов повреждаются чаще всего.

Этот пример использует поток *ships_genlin.str*, в котором используется файл данных *ships.sav*. Файл данных находится в папке *Demos*, а файл потока - в подпапке *streams*.

Моделирование грубой оценки количеств в ячейках в этой ситуации может ввести в заблуждение, поскольку переменная *Суммарные месяцы эксплуатации* зависит от типа судна. Переменные, подобные этой, измеряющие величину "подверженности" риску, обрабатываются в обобщенной линейной модели как переменные смещения. Кроме того, регрессия Пуассона предполагает, что логарифм зависимой переменной линеен в предикторах. Таким образом, в подгонке регрессии Пуассона для количества происшествий с помощью линейных моделей нужно использовать *Логарифм суммарных месяцев эксплуатации*.

Подгонка регрессии Пуассона "со сверхрассеиванием"

1. Добавьте узел источников файла статистики, указывающий на файл *ships.sav* в папке *Demos*.

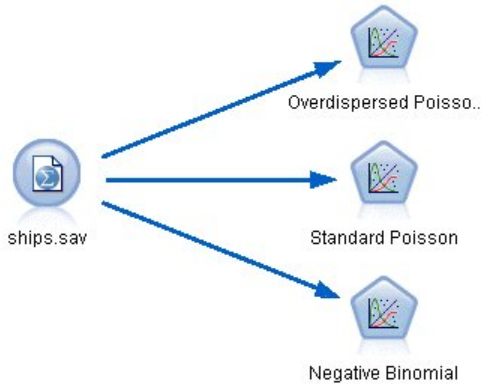


Рисунок 311. Поток примера для анализа частоты повреждений

2. На вкладке Фильтр узла источника исключите поле *months_service*. Подвергнутые логарифмическому преобразованию значения этой переменной содержатся в поле *log_months_service*, которое будет использоваться при анализе.

2. McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

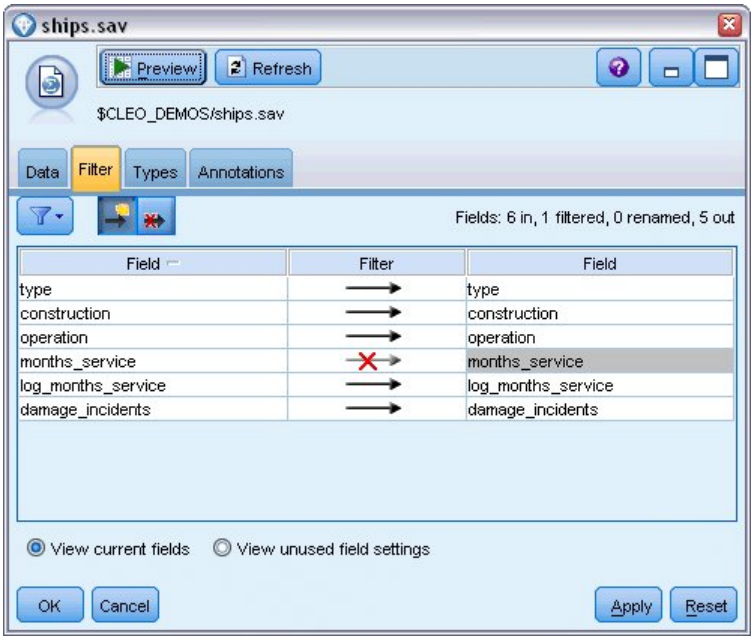


Рисунок 312. Фильтрация ненужного поля

(Другой вариант: для этого поля можно изменить роль можно на **Нет** на вкладке Типы вместо того, чтобы его исключать, или выбрать поля, которые вы хотите использовать на узле моделирования.)

3. На вкладке Типы узла источника задайте для поля *damage_incidents* роль **Назначение**. Для всех остальных полей нужно задать роль **Ввод**.
4. Нажмите кнопку **Чтение данных**, чтобы создать экземпляр данных.

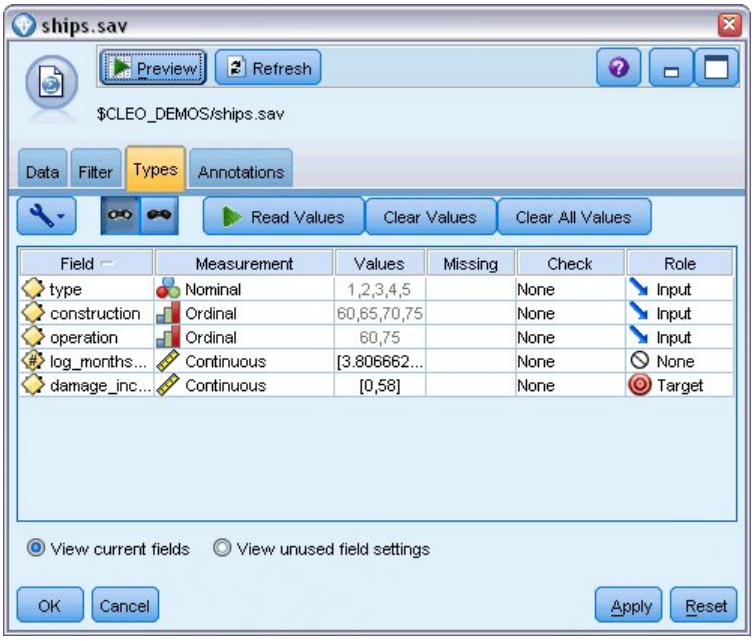


Рисунок 313. Задание роли поля

5. Присоедините к узлу источника узел Обобщенная линейная регрессия; на узле обобщенной линейной модели щелкните по вкладке **Модель**.

6. В качестве переменной смещения выберите *log_months_service*.

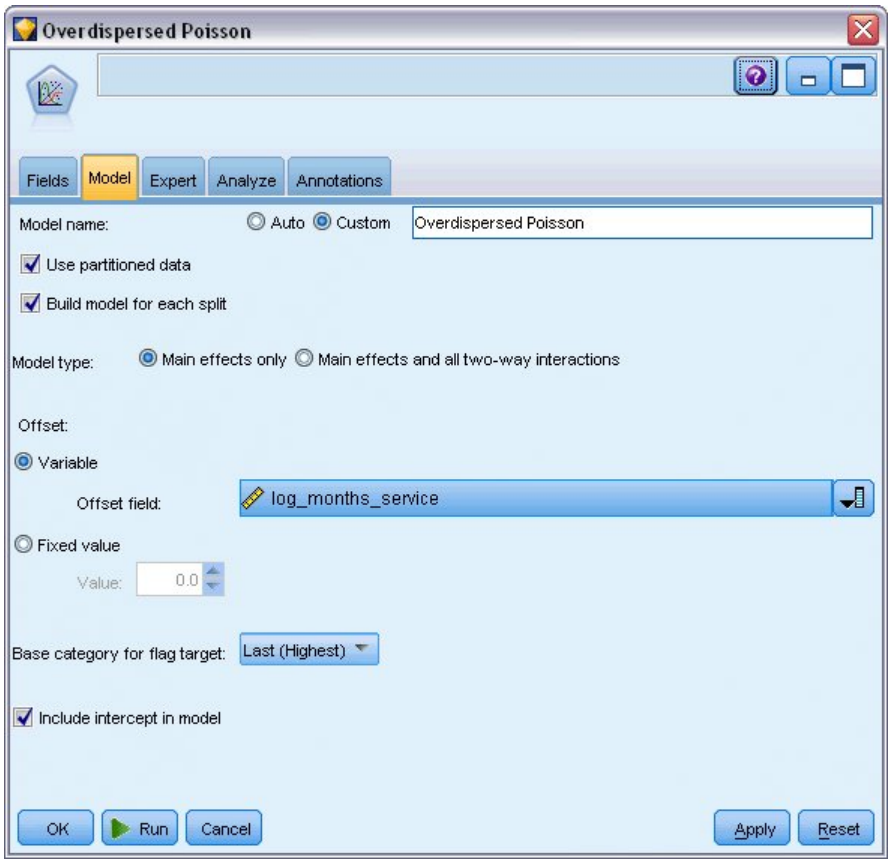


Рисунок 314. Выбор опций модели

7. Щелкните по вкладке **Эксперт** и выберите **Эксперт**, чтобы активировать экспертные опции моделирования.

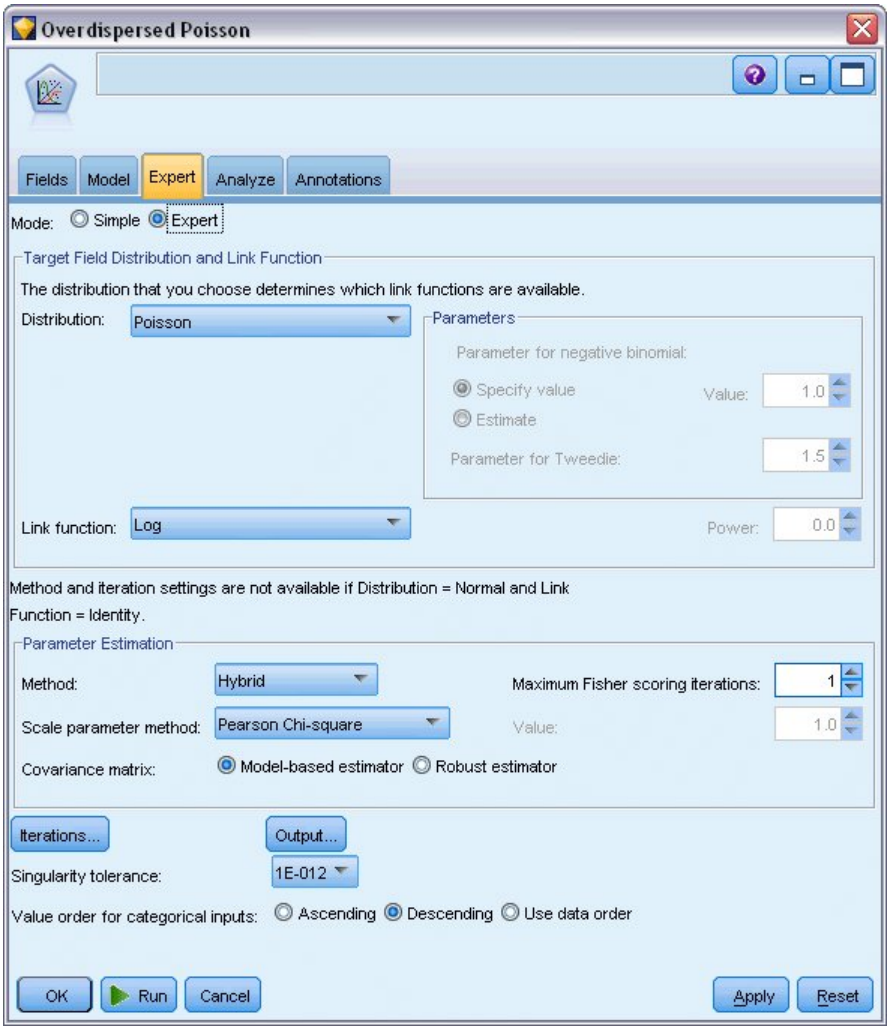


Рисунок 315. Выбор дополнительных опций

8. В качестве распределения для ответа выберите распределение **Пуассона**, а в качестве функции связи - **Логарифмическая**.
9. В качестве метода для оценки параметра масштаба выберите **Хи-квадрат Пирсона**. Обычно считается, что параметр масштаба в регрессии Пуассона равен 1, но Маккала и Нелдер применяют оценку хи-квадрат Пирсона для получения более консервативных оценок дисперсии и уровней значимости.
10. Выберите **По убыванию** в качестве порядка категорий для коэффициентов. Это значит, что первая категория каждого фактора будет его эталонной категорией; влияние этого выбора на модель заключается в интерпретации оценок параметров.
11. Нажмите кнопку **Выполнить**, чтобы создать слепок модели, который будет добавлен на холст потока, а также на палитру моделей в правом верхнем углу. Для просмотра подробностей модели щелкните правой кнопкой мыши по слепку и выберите **Изменить** или **Просмотреть**; затем щелкните по вкладке **Дополнительно**.

Статистики согласия

	Value	df	Value/df
Deviance	38.695	25	1.548
Scaled Deviance	22.883	25	
Pearson Chi-Square	42.275	25	1.691
Scaled Pearson Chi-Square	25.000	25	
Log Likelihood[a]	-68.281		
Akaike's Information Criterion (AIC)	154.562		
Finite Sample Corrected AIC (AICC)	162.062		
Bayesian Information Criterion (BIC)	168.299		
Consistent AIC (CAIC)	177.299		

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

a. The full log likelihood function is displayed and used in computing information criteria.

b. Information criteria are in small-is-better form.

Рисунок 316. Статистика критерия согласия

Таблица Статистика критерия согласия содержит измерения, полезные для сравнения конкурирующих моделей. Кроме того, *Значение/ст.св.* для статистики уклонения и статистики Хи-квадрат Пирсона дает соответствующие оценки для параметра масштаба. Эти значения для регрессии Пуассона должны быть близки к 1,0. Если они больше 1,0, это означает, что может оказаться правдоподобной подгонка модели со сверхрассеиванием.

Универсальный критерий

Likelihood Ratio Chi-Square	df	Sig.
107.633	8	.000

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

a. Compares the fitted model against the intercept-only model.

Рисунок 317. Универсальный критерий

Универсальный критерий - это критерий хи-квадрат отношения правдоподобия текущей модели относительно пустой модели (в данном случае модели со свободным членом). Уровень значимости меньше 0,05 означает, что текущая модель работает лучше пустой модели.

Проверки эффектов модели

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	2138.657	1	.000
type	15.415	4	.004
construction	17.242	3	.001
operation	6.249	1	.012

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

Рисунок 318. Тестирование эффектов моделей

Каждый член в модели проверяется на наличие у него какой-либо эффекта. У членов с уровнями значимости меньше 0,05 есть некоторый видимый эффект. Каждый из членов главных эффектов вносит вклад в модель.

Оценки параметров

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-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]	.	.	.	.	.	.
(Scale)	1.691[b]						

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

a. Set to zero because this parameter is redundant.

b. Computed based on the Pearson chi-square.

Рисунок 319. Оценки параметров

Таблица Оценки параметров содержит сводку эффектов каждого из предикторов. При том, что интерпретация коэффициентов в этой модели трудна из-за особенностей функции связи, знаки коэффициентов для ковариат и относительных значений коэффициентов для уровней факторов могут дать важное понимание эффектов предикторов в этой модели.

- Для ковариат положительные (отрицательные) коэффициенты означают положительные (обратные) взаимосвязи между предикторами и результатами выполнения. Увеличивающееся значение ковариаты с положительным коэффициентом соответствует увеличивающейся частоте случаев повреждений.
- Для факторов уровень факторов с более высоким коэффициентом означает более высокую частоту повреждения. Знак коэффициента для уровня факторов зависит от эффекта этого уровня факторов относительно эталонной категории.

На основе оценок параметров возможны следующие интерпретации:

- У типа судна *B [type=2]* частота повреждений статистически значимая (значение *p* 0,019) частота повреждений (оцененный коэффициент –0,543) ниже, чем у типа *A [type=1]* - эталонной категории. У типа

C [type=3] оцененный параметр фактически ниже, чем у типа *B*, но изменчивость оценки типа *C* затемняет эффект. Смотрите оценки маргинальных средних для всех связей между между уровнями факторов.

- У судов, построенных между 1965–69 *[construction=65]* и 1970–74 *[construction=70]* статистически значимая (значения $p < 0{,}001$) частота повреждений (оцененные коэффициенты 0,697 и 0,818 соответственно) выше, чем у судов, построенных между 1960 и 1964 *[construction=60]* - эталонной категории. Смотрите оценки маргинальных средних для всех связей между между уровнями факторов.
- У судов, введенных в эксплуатацию между 1975 и 1979 *[operation=75]*, статистически значимая (значение *p* 0,12) частота повреждений (оцененный коэффициент 0,384) выше, чем у судов, введенных в эксплуатацию между 1960 и 1974 *[operation=60]*.

Подгонка альтернативных моделей

Одна проблема с регрессией Пуассона "со сверхрассеиванием" состоит в том, что отсутствует формальный способ ее проверки относительно "стандартной" регрессии Пуассона. Однако один формальный тест, позволяющий определить, есть ли сверхрассеивание, подразумевает выполнение проверки отношения правдоподобия между "стандартной" регрессией Пуассона и отрицательной биномиальной регрессией со всеми остальными одинаково заданными параметрами. Если в регрессии Пуассона нет никакого сверхрассеивания, то у статистики −2×(логарифмическое правдоподобие для модели Пуассона − логарифмическое правдоподобие для модели отрицательной биномиальной регрессии) должно быть смешанное распределение с половиной его положительной вероятностной меры 0 и остатком в распределении хи-квадрат с одной степенью свободы.

1. В качестве метода для оценки параметра масштаба выберите **Фиксированное значение**. Значение по умолчанию - 1.

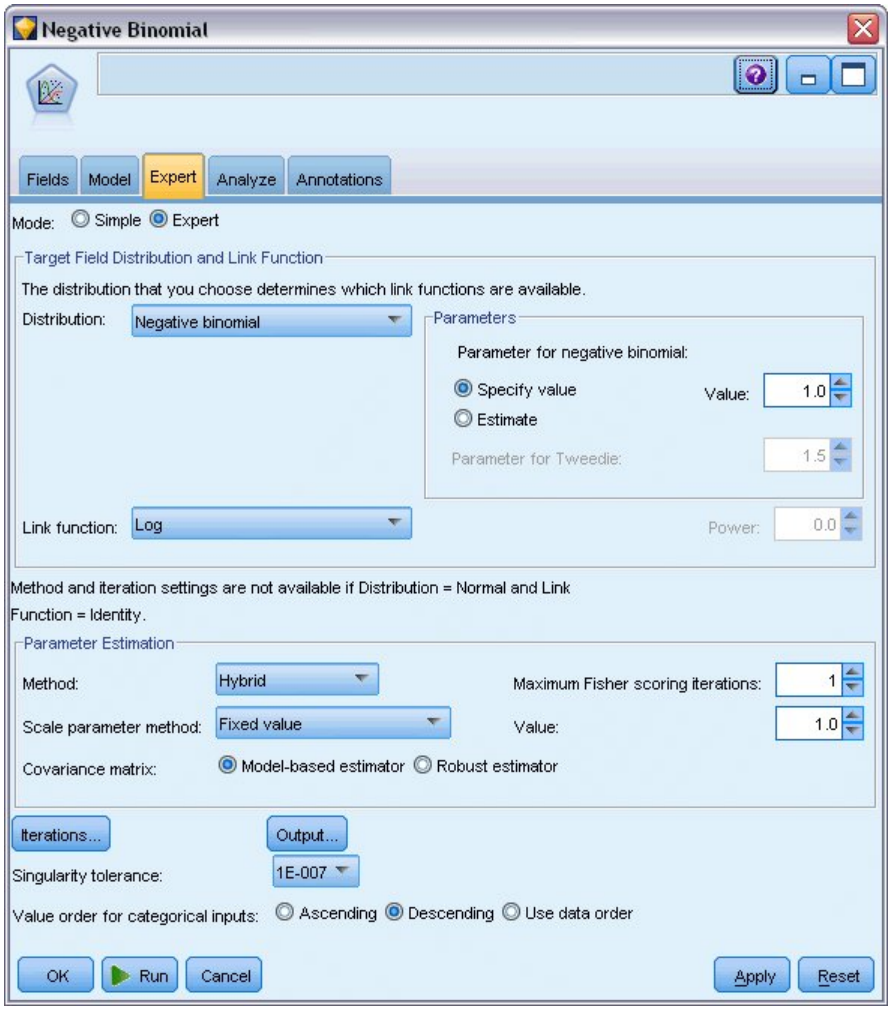


Рисунок 320. Вкладка Эксперт

2. Для подгонки отрицательной биномиальной регрессии скопируйте и вставьте узел обобщенной линейной модели, подсоедините его к узлу источника, откройте новый узел и щелкните по вкладке **Эксперт**.
3. В качестве распределения выберите **Отрицательное биномиальное**. Оставьте для вспомогательного параметра значение по умолчанию 1.
4. Запустите поток и просмотрите вкладку Дополнительно для вновь созданных слепков моделей.

Статистики согласия

	Value	df	Value/df
Deviance	38.695	25	1.548
Scaled Deviance	38.695	25	
Pearson Chi-Square	42.275	25	1.691
Scaled Pearson Chi-Square	42.275	25	
Log Likelihood[a]	-68.281		
Akaike's Information Criterion (AIC)	154.562		
Finite Sample Corrected AIC (AICC)	162.062		
Bayesian Information Criterion (BIC)	168.299		
Consistent AIC (CAIC)	177.299		

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

a. The full log likelihood function is displayed and used in computing information criteria.

b. Information criteria are in small-is-better form.

Рисунок 321. Статистика критерия согласия для стандартной регрессии Пуассона

Логарифмическое правдоподобие, сообщаемое для стандартной регрессии Пуассона, составляет –68,281. Сравните его с моделью отрицательной биномиальной регрессии.

	Value	df	Value/df
Deviance	11.145	25	.446
Scaled Deviance	11.145	25	
Pearson Chi-Square	8.815	25	.353
Scaled Pearson Chi-Square	8.815	25	
Log Likelihood[a]	-83.725		
Akaike's Information Criterion (AIC)	185.450		
Finite Sample Corrected AIC (AICC)	192.950		
Bayesian Information Criterion (BIC)	199.187		
Consistent AIC (CAIC)	208.187		

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

a. The full log likelihood function is displayed and used in computing information criteria.

b. Information criteria are in small-is-better form.

Рисунок 322. Статистика критерия согласия для отрицательной биномиальной регрессии

Логарифмическое правдоподобие, сообщаемое для отрицательной биномиальной регрессии, составляет –83,725. Фактически оно *меньше* логарифмического правдоподобия для регрессии Пуассона, что означает (в отсутствие необходимости проверки отношения правдоподобия), что эта отрицательная биномиальная регрессия не предлагает улучшения по сравнению с регрессией Пуассона.

Однако выбранное значение 1 для вспомогательного параметра отрицательного биномиального распределения может не оказаться оптимальным для этого набора данных. Другой способ, который можно опробовать для сверхрассеивания - выполнить подгонку модели отрицательной биномиальной регрессии со вспомогательным параметром, равным 0, и затребовать критерий множителей Лагранжав диалоговом окне Вывод вкладки Эксперт. Если этот критерий несуществен, сверхрассеивание не должно представлять собой проблему для этого набора данных.

Итог

С помощью обобщенных линейных моделей вы выполнили подгонку трех отличающихся моделей для дискретных данных. Доказано, что отрицательная биномиальная регрессия не дает никакого улучшения по сравнению с регрессией Пуассона. Регрессия Пуассона со сверхрассеиванием выглядит разумной альтернативой стандартной модели Пуассона, но формального теста для выбора между ними не существует.

Объяснение математических основ методов моделирования, используемых в IBM SPSS Modeler, смотрите в публикации *IBM SPSS Modeler: Руководство по алгоритмам*.

Связанные процедуры

Процедура Обобщенные линейные модели - мощный инструмент для подгонки самых разнообразных моделей.

- Процедура Обобщенные уравнения оценки расширяет обобщенную линейную модель, разрешая повторные измерения.
- Процедура Линейные смешанные модели позволяет выполнять подгонку моделей для количественных зависимых переменных со случайными компонентными и/или повторными измерениями.

Рекомендуемое чтение

Дополнительную информацию об обобщенных линейных моделях смотрите в следующих текстовых источниках:

Cameron, A. C., and P. K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press. Dobson, A. J. 2002. *An Introduction to Generalized Linear Models*, 2 ed. Boca Raton, FL: Chapman & Hall/CRC. Hardin, J. W., and J. M. Hilbe. 2003. *Generalized Linear Models and Extension*. Station, TX: Stata Press. McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

Глава 24. Подгонки гамма-регрессии для страховых исков по автомобилям (Обобщенные линейные модели)

Обобщенную линейную модель можно использовать с целью подгонки гамма-регрессии для анализа данных положительного диапазона. Например, набор данных, представленный и проанализированный в другой работе [3], касается исков за повреждение автомобилей.' Среднюю сумму иска можно смоделировать как имеющую гамма-распределение, использующее функцию инверсной связи для соотношения среднего зависимой переменной с линейной комбинацией предикторов. Для учёта переменного числа исков, используемого для вычисления средних сумм исков, надо задать *Число исков* в качестве веса масштабирования.

Этот пример использует поток *car-insurance_genlin.str*, в котором используется файл данных *car_insurance_claims.sav*. Файл данных находится в папке *Demos*, а файл потока - в подпапке *streams*.

Создание потока

1. Добавьте узел источников файла статистики, указывающий на файл *car_insurance_claims.sav* в папке *Demos*.

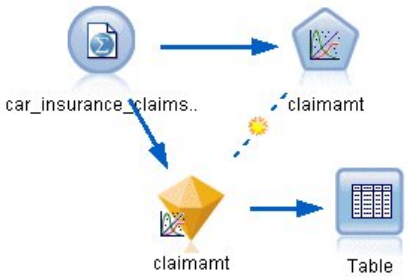


Рисунок 323. Поток примера для предсказания страховых исков по автомобилям

2. На вкладке Типы узла источника задайте для поля *claimamt* роль **Назначение**. Для всех остальных полей нужно задать роль **Ввод**.
3. Нажмите кнопку **Чтение данных**, чтобы создать экземпляр данных.

3. McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

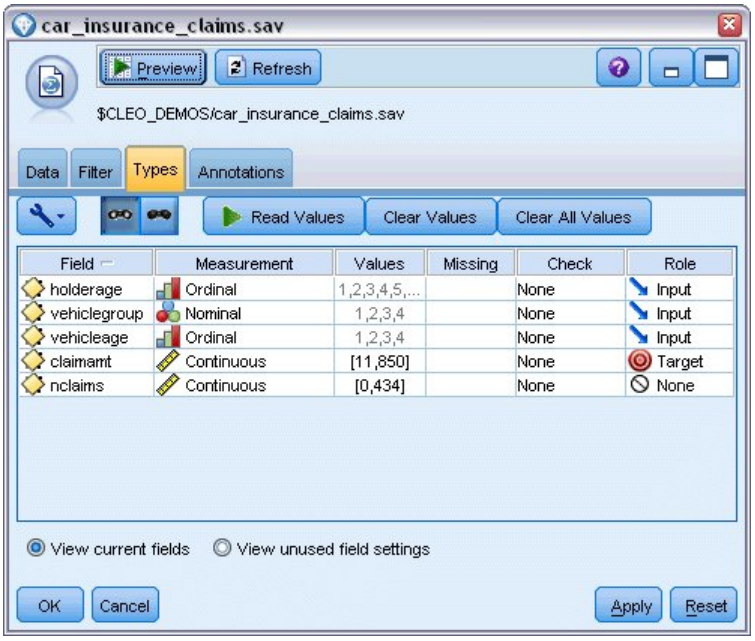


Рисунок 324. Задание роли поля

4. Присоедините к узлу источника узел Обобщенная линейная регрессия; на узле обобщенной линейной модели щелкните по вкладке Поля.
5. В качестве поля масштабного веса выберите *nclaims*.

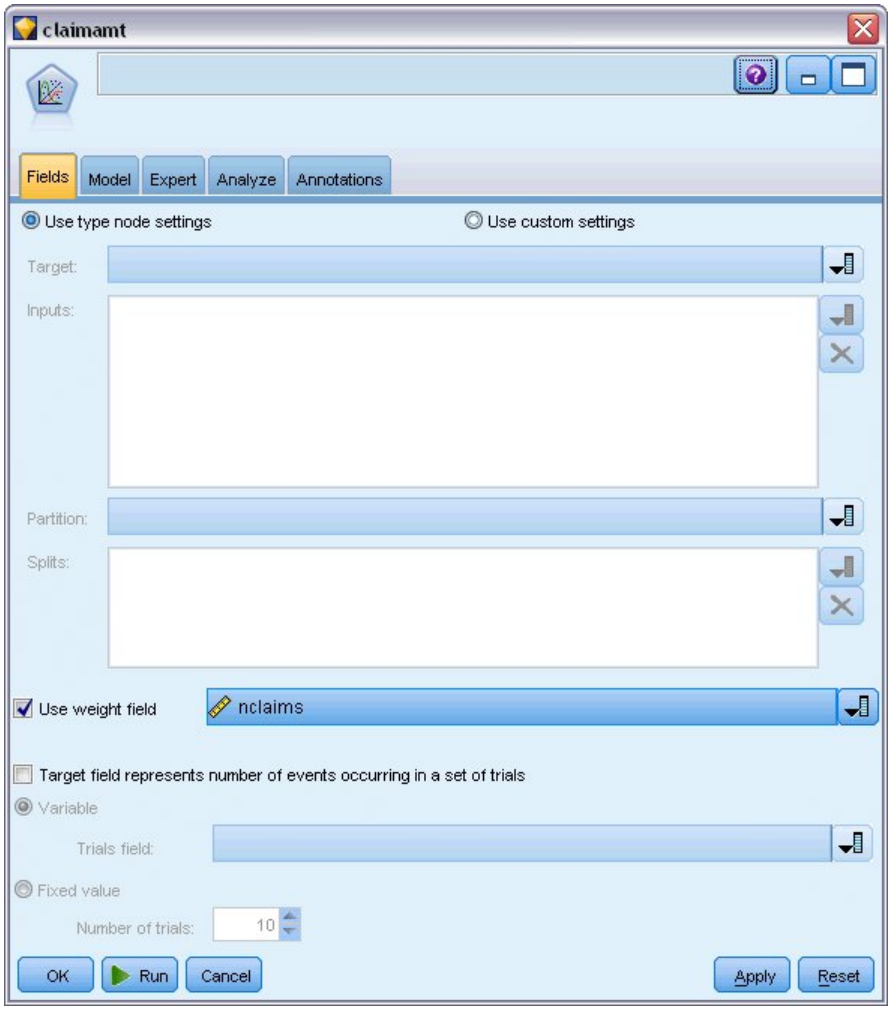


Рисунок 325. Выбор опции полей

6. Щелкните по вкладке **Эксперт**, чтобы активировать дополнительные опции моделирования.

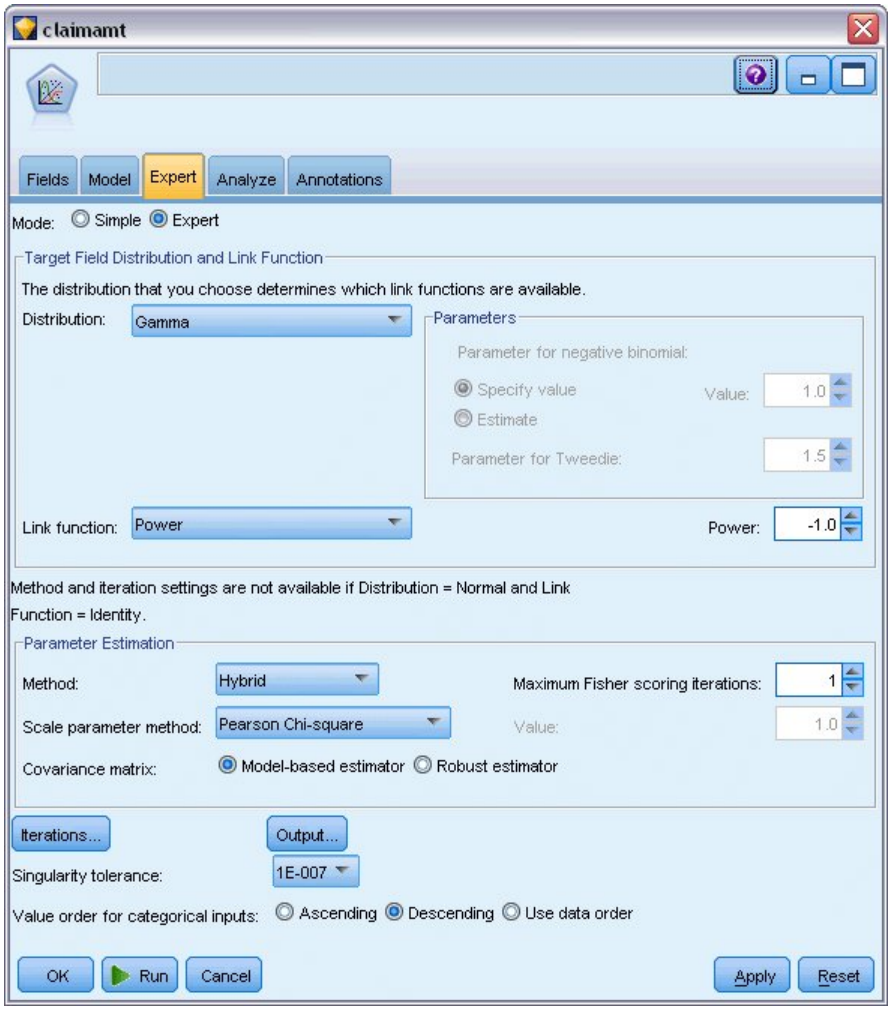


Рисунок 326. Выбор дополнительных опций

7. В качестве распределения откликов выберите **Гамма**.
8. В качестве функции связи выберите **Степенная**, а в качестве показателя ее степени введите `-1.0`. Это обратная связь.
9. В качестве метода для оценки параметра масштаба выберите **Хи-квадрат Пирсона**. Это метод, используемый Маккала и Нелдером, поэтому мы следуем ему здесь, чтобы повторить их результаты.
10. Выберите **По убыванию** в качестве порядка категорий для коэффициентов. Это значит, что первая категория каждого фактора будет его эталонной категорией; влияние этого выбора на модель заключается в интерпретации оценок параметров.
11. Нажмите кнопку **Выполнить**, чтобы создать слепок модели, который будет добавлен на холст потока, а также на палитру моделей в верхнем правом углу. Для просмотра подробностей модели щелкните правой кнопкой мыши по слепку модели и выберите **Изменить** или **Просмотреть**; затем выберите вкладку Дополнительно.

Оценки параметров

Parameter	B	Std. Error	95% Wald Confidence Interval Lower	95% Wald Confidence Interval Upper	Hypothesis Test Wald Chi-Square	Hypothesis Test df	Hypothesis Test Sig.
(Intercept)	.003411	.000418	.002591	.004230	66.593	1	.000
[holderage=8]	.000920	.000416	.000105	.001735	4.898	1	.027
[holderage=7]	.000916	.000408	.000117	.001716	5.046	1	.025
[holderage=6]	.000969	.000405	.000176	.001763	5.740	1	.017
[holderage=5]	.001370	.000419	.000548	.002192	10.682	1	.001
[holderage=4]	.000462	.000411	-.000342	.001267	1.268	1	.260
[holderage=3]	.000350	.000412	-.000458	.001158	.720	1	.396
[holderage=2]	.000101	.000436	-.000754	.000956	.054	1	.816
[holderage=1]	.000000[a]	.	.	.	.	.	.
[vehiclegroup=4]	-.001421	.000181	-.001775	-.001067	61.883	1	.000
[vehiclegroup=3]	-.000614	.000170	-.000947	-.000281	13.039	1	.000
[vehiclegroup=2]	.000038	.000169	-.000293	.000368	.050	1	.823
[vehiclegroup=1]	.000000[a]	.	.	.	.	.	.
[vehicleage=4]	.004154	.000442	.003287	.005021	88.175	1	.000
[vehicleage=3]	.001651	.000227	.001207	.002096	53.013	1	.000
[vehicleage=2]	.000366	.000101	.000169	.000564	13.191	1	.000
[vehicleage=1]	.000000[a]	.	.	.	.	.	.
(Scale)	1.209[b]			.001	.0004	.000	.002

Dependent Variable: Average cost of claims
Model: (Intercept), holderage, vehiclegroup, vehicleage

a. Set to zero because this parameter is redundant.

b. Computed based on the Pearson chi-square.

Рисунок 327. Оценки параметров

Универсальный критерий и проверки эффектов модели (не показаны) указывают, что полученная модель работает лучше пустой модели и что каждый из главных эффектов вносит вклад в модель. Таблица оценок параметров содержит значения, одинаковые с полученными Маккала и Нелдером (McCullagh and Nelder) для уровней факторов и параметра масштаба.

Итог

С помощью обобщенных линейных моделей вы выполнили подгонку гамма-регрессии для данных исков. Обратите внимание на то, что для этой модели использовалась каноническая функция связи для гамма-распределения, но логарифмическая функция связи также может дать правдоподобные результаты. В целом трудно, если возможно вообще, непосредственно сравнить модели с отличающимися функциями связи; однако логарифмическая функция связи - это частный случай степенной функции связи, где показатель степени - 0, поэтому можно сравнить отклонения модели с логарифмический функцией связи и модели со степенной функцией связи, чтобы определить, какая из них дает лучшую подгонку (смотрите, например, раздел 11.3 публикации Маккала и Нелдера - McCullagh and Nelder).

Объяснение математических основ методов моделирования, используемых в IBM SPSS Modeler, смотрите в публикации *IBM SPSS Modeler: Руководство по алгоритмам*.

Связанные процедуры

Процедура Обобщенные линейные модели - мощный инструмент для подгонки самых разнообразных моделей.

- Процедура Обобщенные уравнения оценки расширяет обобщенную линейную модель, разрешая повторные измерения.
- Процедура Линейные смешанные модели позволяет выполнять подгонку моделей для количественных зависимых переменных со случайными компонентными и/или повторными измерениями.

Рекомендуемое чтение

Дополнительную информацию об обобщенных линейных моделях смотрите в следующих текстовых источниках:

Cameron, A. C., and P. K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press. Dobson, A. J. 2002. *An Introduction to Generalized Linear Models*, 2 ed. Boca Raton, FL: Chapman & Hall/CRC. Hardin, J. W., and J. M. Hilbe. 2003. *Generalized Linear Models and Extension*. Station, TX: Stata Press. McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

Глава 25. Классификация образцов клеток (SVM)

Механизм опорных векторов (Support Vector Machine, SVM) - это способ классификации и построения регрессии, особенно подходящий для больших наборов данных. Большой набор данных содержит много предикторов, что может встретиться в области биоинформатики (применение информационных технологий к биохимическим и биологическим данным).

В медицинских исследованиях получен набор данных, содержащих характеристики многих образцов человеческих клеток от пациентов, для которых предполагается риск развития рака. Анализ исходных данных показал, что для здоровых и злокачественных клеток многие характеристики существенно отличаются. Медики хотят разработать модель SVM, которая сможет использовать значения характеристик клеток в образцах от других пациентов, чтобы получить раннюю диагностику нормальности или злокачественности новых образцов.

В этом примере используется поток с именем *svm_cancer.str*, доступный в папке *Demos* в подпапке *streams*. Файл данных - это *cell_samples.data*. Дополнительную информацию смотрите в разделе “Папка demos” на стр. 4.

Этот пример основан на наборе данных, общедоступном в репозитории UCI Machine Learning. Этот набор данных состоит из нескольких сотен записей образцов человеческих клеток, каждая из которых содержит значения набора клеточных характеристик. В каждой записи есть следующие поля:

Имя поля	**Описание**
ID	Идентификатор пациента
Clump	Консистенция колонии
UnifSize	Однородность размеров клеток
UnifShape	Однородность формы клеток
MargAdh	Граничная адгезия
SingEpiSize	Размер одной эпителиальной клетки
BareNuc	Голые ядра
BlandChrom	Пассивный хроматин
NormNucl	Обычные ядрышки
Mit	Митозы
Class	Здоровая или злокачественная

Для целей этого примера мы используем набор данных с относительно малым числом предикторов в каждой записи.

Создание потока

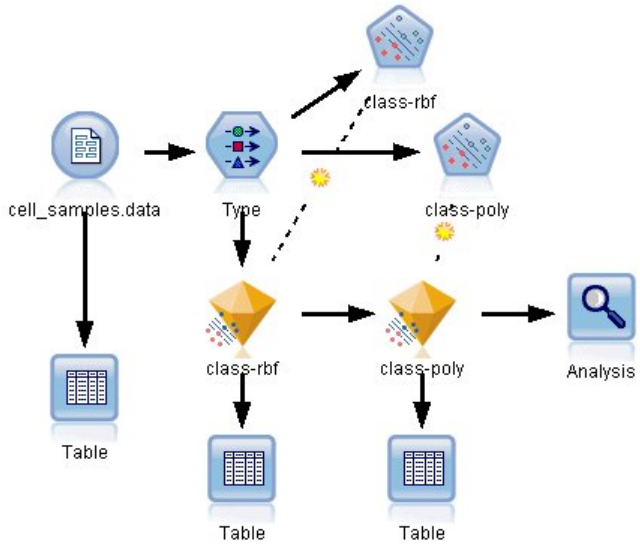
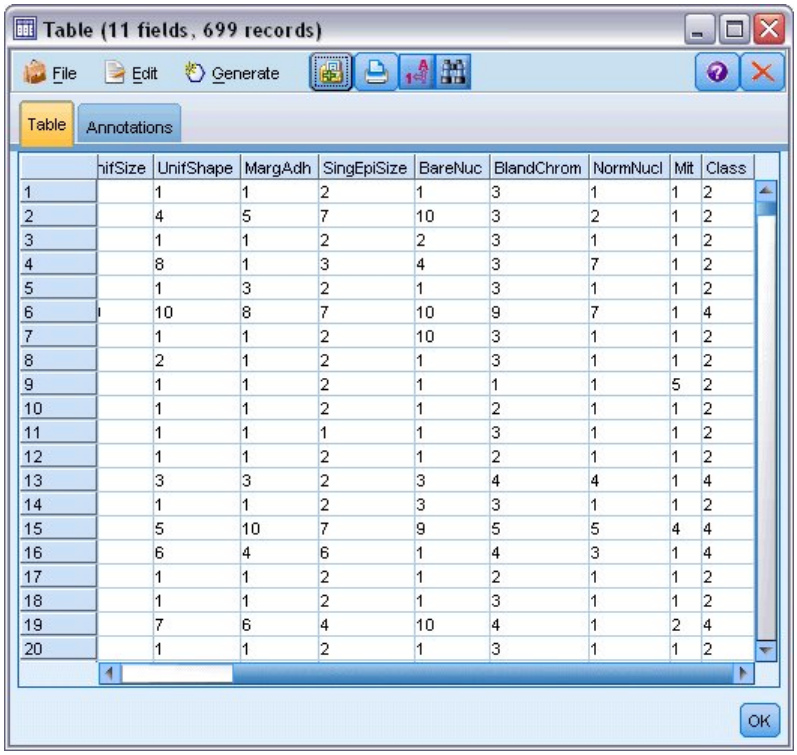


Рисунок 328. Пример потока для иллюстрации моделирования SVM

1. Создайте новый поток и добавьте исходный узел Файл переменных, указывающий на файл *cell_samples.data* в папке *Demos* вашей установки IBM SPSS Modeler.

   Давайте посмотрим на данные в исходном файле.
2. Добавьте узел Таблица к потоку.
3. Присоедините узел Таблица к узлу Файл переменных и запустите поток.



	ifSize	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

Рисунок 329. Исходные данные для SVM

Поле *ID* содержит идентификаторы пациентов. Характеристики образцов клеток от каждого пациента находятся в полях от *Clump* до *Mit*. Значения ранжируются от 1 до 10, где 1 - ближайшее значение для здоровых клеток.

Поле *Class* содержит диагноз, подтвержденный отдельными медицинскими процедурами, то есть здоровые ли это клетки (значение = 2), или злокачественные (значение = 4).

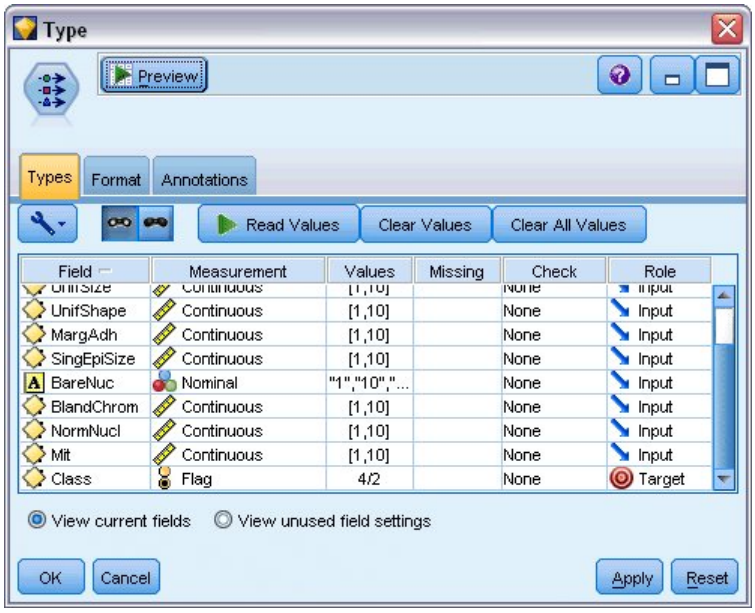


Рисунок 330. Параметры узла Тип

4. Добавьте узел Тип и присоедините его к узлу Файл переменных.

5. Откройте узел Тип.

   Мы хотим, чтобы эта модель предсказывала значение в поле *Class* (то есть здоровая ли клетка (=2), или злокачественная (=4)). Так как в этом поле может быть только одно из двух возможных значений, для отображения этого следует изменить ее уровень измерения.
6. В столбце **Измерение** для поля *Class* (последний в списке), щелкните по значению **Непрерывное** и замените его на **Флаг**.
7. Нажмите кнопку **Прочесть значения**.
8. В столбце **Роль** задайте роли для поля *ID* (идентификатора пациента) значение **Нет**, так как это поле не будет использоваться ни как предиктор, ни как поле назначения для модели.
9. Для поля назначения *Class* задайте роль **Назначение** и оставьте значением роли всех остальных полей (предикторов) **Вход**.
10. Щелкните по **OK**.

    Узел SVM предлагает выбор функций ядра для выполнения своей обработки. Так как нет простого способа определения, какая из функций лучше всего сработает с любым данным набором данных, будем выбирать различные функции по очереди и сравнивать их результаты. Начнем с функции по умолчанию RBF (Radial Basis Function, радиальная базовая функция).

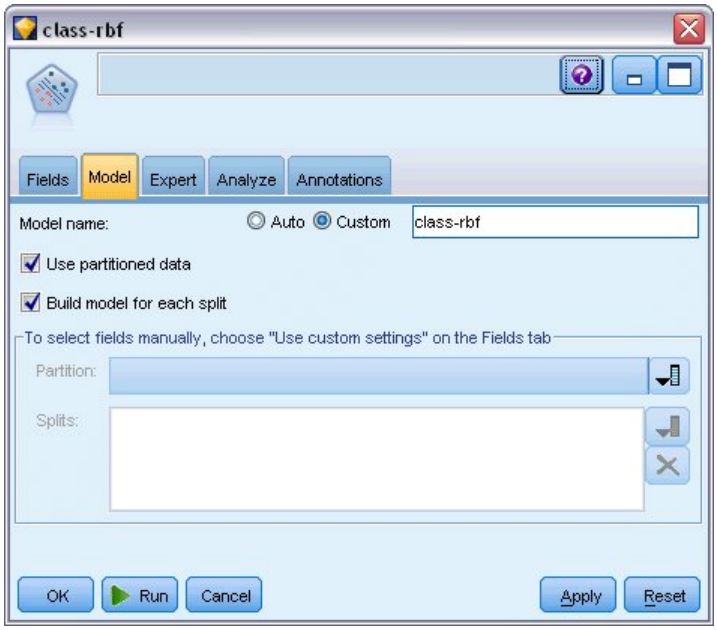


Рисунок 331. Параметры вкладки модели

11. На палитре Моделирование присоедините узел SVM к узлу Тип.
12. Откройте узел SVM. На вкладке **Модель** выберите опцию **Пользовательское** для **Имени модели** и введите *class-rbf* в соответствующем текстовом поле.

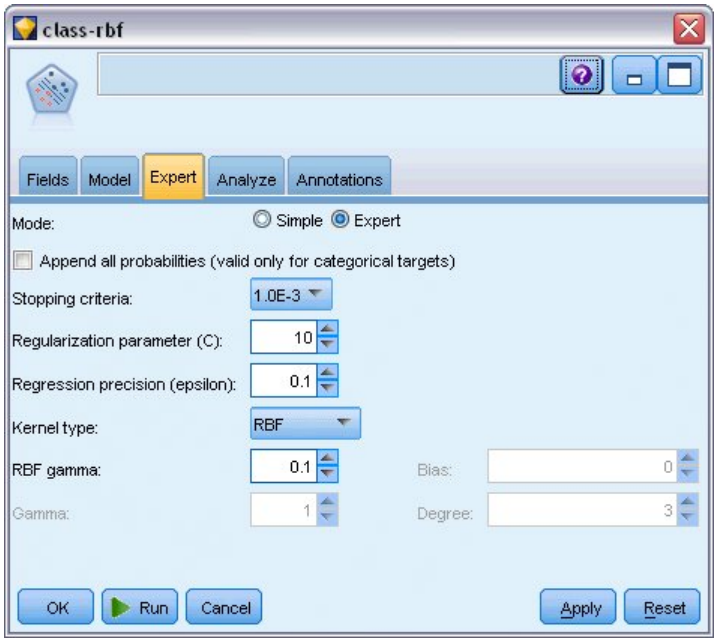


Рисунок 332. Параметры вкладки Эксперт по умолчанию

13. На вкладке **Эксперт** задайте для **Режима** значение **Эксперт** для понятности, но для всех других опций по умолчанию оставьте существующие значения. Обратите внимание на то, что по умолчанию для **Типа ядра** задано значение **RBF**. В простом режиме все эти опции отключены.

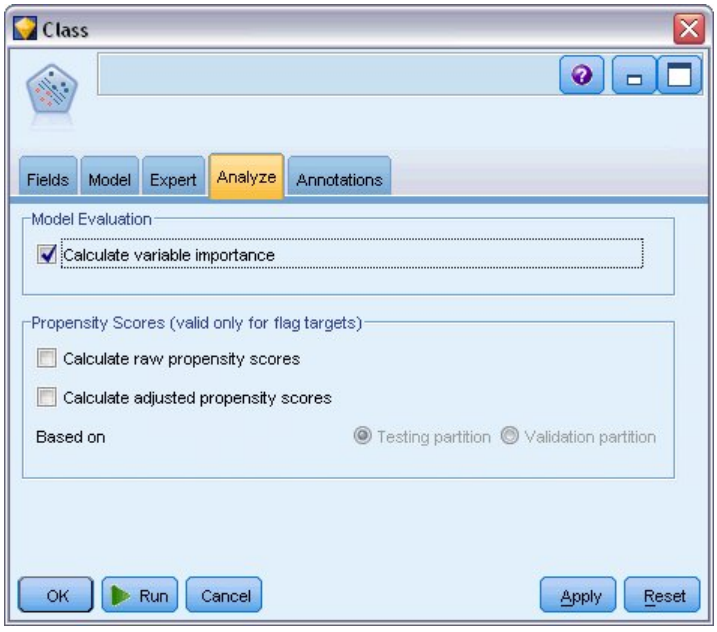


Рисунок 333. Параметры вкладки Анализ

14. На вкладке **Анализ** включите переключатель **Вычислить важность переменных**.
15. Нажмите кнопку **Выполнить**. Слепок модели будет размещен в потоке и на палитре Модели в правом верхнем углу экрана.
16. Дважды щелкните по слепку модели в потоке.

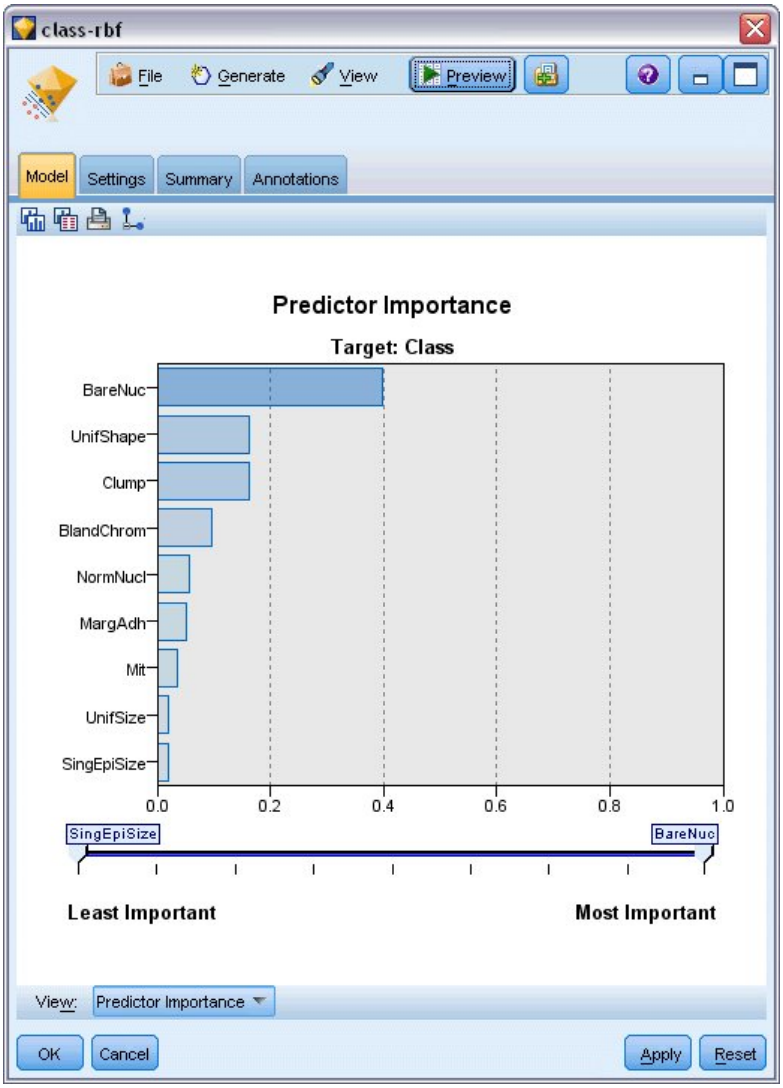
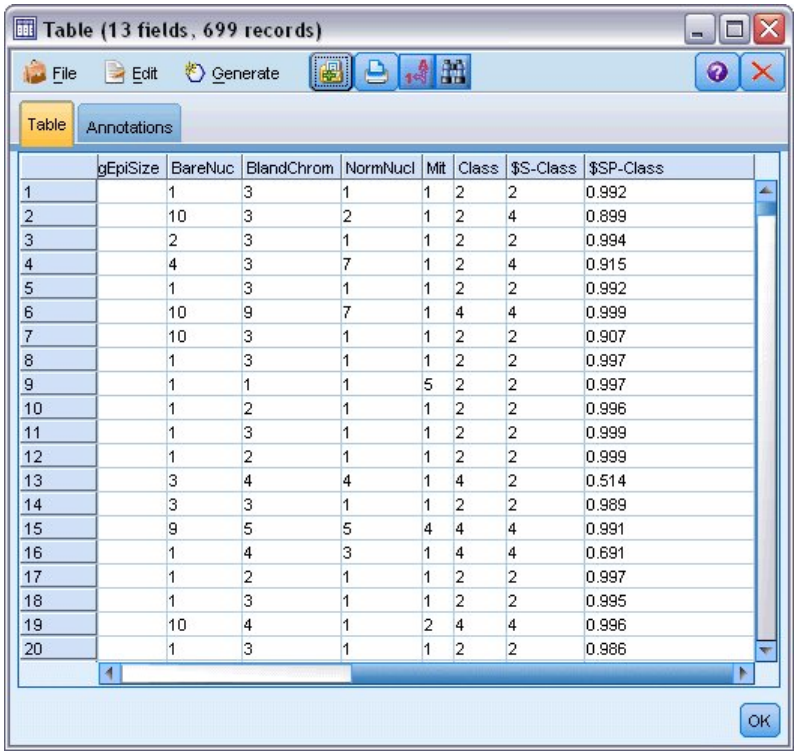


Рисунок 334. Диаграмма Важность предиктора

На вкладке Модель диаграмма Важность предикторов показывает относительное влияние разных полей на предсказание. Здесь легко увидеть, что влияние *BareNuc* наибольшее, хотя *UnifShape* и *Clump* также весьма существенны.

1. Щелкните по **OK**.
2. Присоедините к слепку модели *class-rbf* узел таблицы.
3. Откройте узел Таблица и щелкните по **Выполнить**.



	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

Рисунок 335. Добавленные поля для предсказания и значения достоверности

4. Эта модель создала два дополнительных поля. Прокрутите выходную таблицу направо, чтобы увидеть их:

Имя нового поля	**Описание**
$S-Class	Значение для поля *Class*, предсказанное моделью.
$SP-Class	Оценка склонности для этого предсказания (вероятность правильности этого предсказания в диапазоне от 0,0 до 1,0).

Просто взглянув на эту таблицу, можно заметить, что оценки склонности (в столбце *$SP-Class*) для большинства записей довольно высокие.

Однако есть несколько существенных исключений; например, запись для пациента 1041801 в строке 13, где значение 0,514 неприемлемо низкое. Кроме этого, при сравнении полей *Class* и *$S-Class* видно, что модель сделала несколько неправильных предсказаний, даже при относительно высоких в этих случаях оценках склонности (например, строки 2 и 4).

Посмотрим, сможем ли мы улучшить результаты, выбрав функцию другого типа.

Проверка другой функции

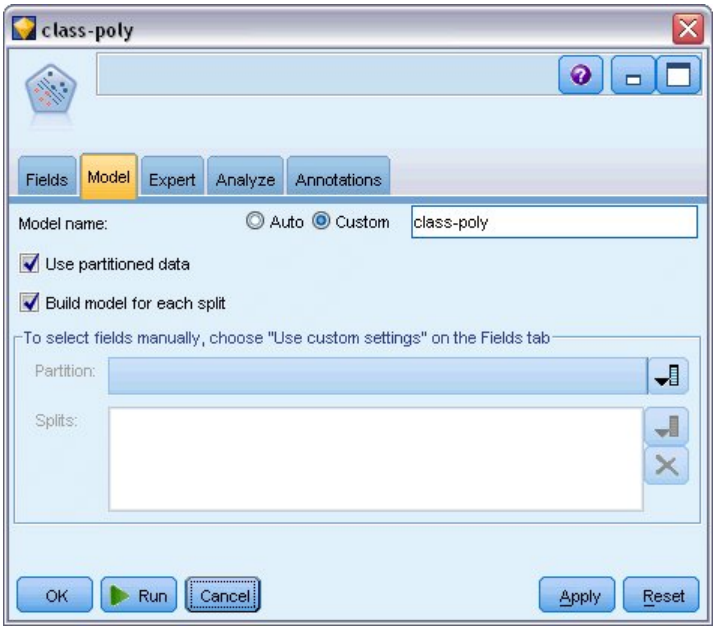


Рисунок 336. Задание нового имени для модели

1. Закройте окно вывода Таблица.
2. Присоедините второй узел Моделирование SVM к узлу Тип.
3. Откройте новый узел SVM.
4. На вкладке **Модель** выберите опцию Пользовательское и ведите имя модели *class-poly*.

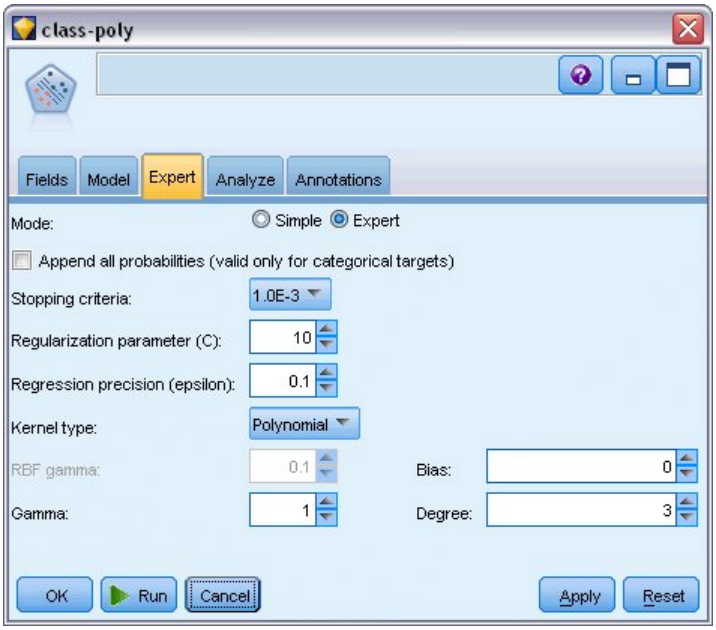
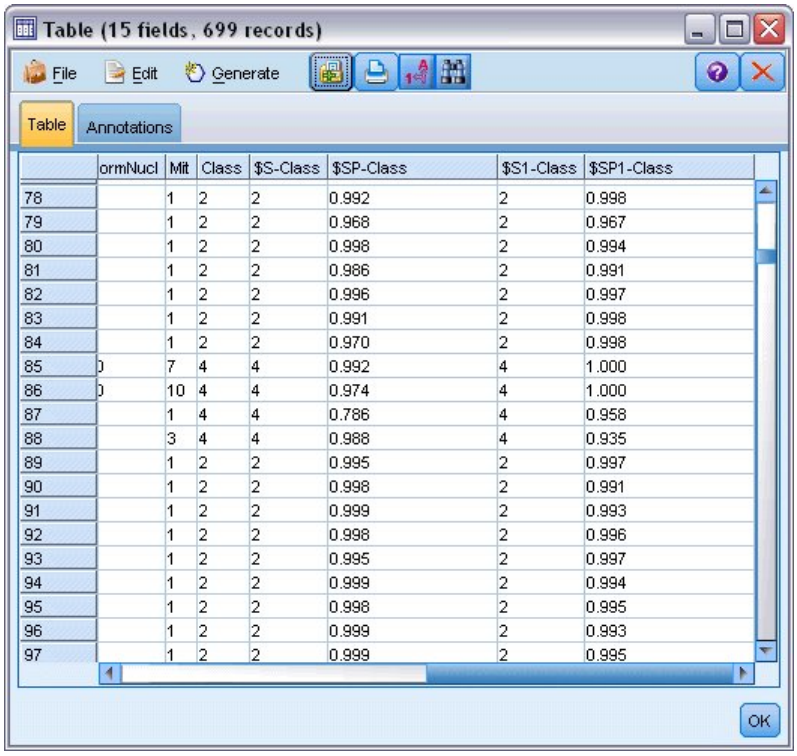


Рисунок 337. Параметры вкладки Эксперт для полиномиальной модели

5. На вкладке **Эксперт** задайте для **Режима** значение **Эксперт**.

6. Задайте для **Типа ядра** значение **Полиномиальное** и щелкните по **Выполнить**. Слепок модели *class-poly* будет добавлен в поток и также на палитре Модели в правом верхнем углу экрана.
7. Присоедините слепок модели *class-rbf* к слепку модели *class-poly* (выберите опцию **Заменить** в диалоговом окне предупреждения).
8. Присоедините к слепку *class-poly* узел таблицы.
9. Откройте узел Таблица и щелкните по **Выполнить**.

Сравнение результатов



	ormNucl	Mit	Class	$S-Class	$SP-Class	$S1-Class	$SP1-Class
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

Рисунок 338. Добавленные поля для полиномиальной функции

1. Прокрутите выходную таблицу направо, чтобы увидеть вновь добавленные поля.

   Сгенерированные поля для функций полиномиального типа называются *$S1-Class* и *$SP1-Class*.

   Результаты для полиномиальной модели выглядят гораздо лучше. Многие оценки склонности составляют значение 0,995 или лучше, что воодушевляет.
2. Для подтверждения усовершенствования модели присоедините узел Анализ к слепку модели *class-poly*.

Откройте узел Анализ и нажмите кнопку **Выполнить**.

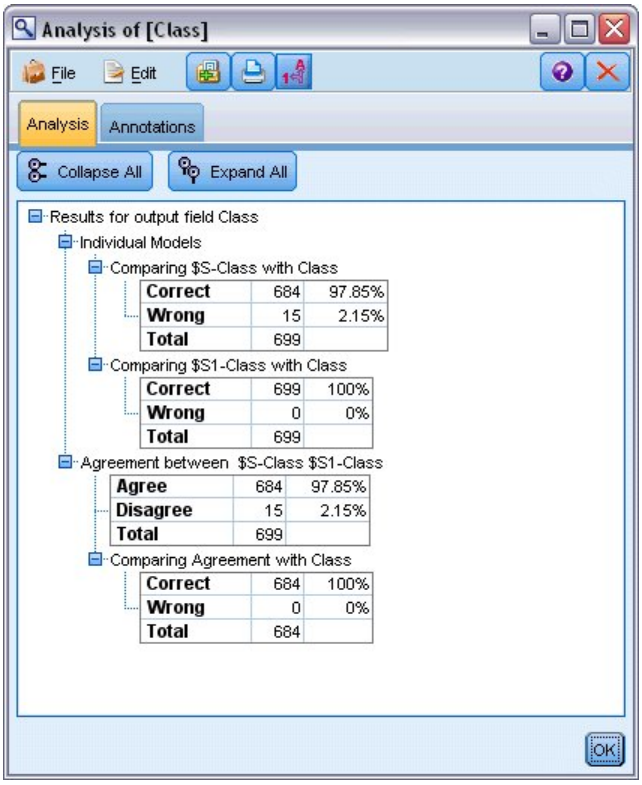


Рисунок 339. Узел Анализ

Способ с использованием узла Анализ позволяет вам сравнить два или более слепков моделей одного типа. Выходные данные узла Анализ показывают, что функция RBF правильно предсказывает 97,85% наблюдений, что весьма хорошо. Однако выходные данные показывают, что полиномиальная функция правильно предсказала диагноз в каждом отдельном наблюдении. На практике вы вряд ли встретите точность в 100%, но узел Анализ можно использовать для определения, приемлема ли точность модели для фактического применения.

На самом деле, ни одна из функций другого типа (сигмоидная или линейная) не работает так хорошо, как полиномиальная, на этом конкретном наборе данных. Однако с другим набором данных результаты вполне могут отличаться, поэтому всегда следует испробовать полный диапазон опций.

Итог

Вы использовали различные типы функций ядра SVM для предсказания классификации по некоторому числу атрибутов. Вы увидели, как различные ядра приводят к разным результатам для одного набора данных и как можно оценивать улучшение одной модели по сравнению с другой.

Глава 26. Использование регрессии Кокса для моделирования времени до оттока клиента

Прикладывая усилия к сокращение оттока клиентов, телекоммуникационная компания хочет, в частности, смоделировать "время текучести", чтобы определить факторы, связанные с клиентами, быстро переключающимися на услуги других компаний. Для этого определяется случайная выборка клиентов, и из базы данных берутся данные о времени использования услуг, продолжают ли эти клиенты активно использовать услуги, и данные различных демографических полей.

Этот пример использует поток *telco_coxreg.str*, в котором используется файл данных *telco.sav*. Файл данных находится в папке *Demos*, а файл потока - в подпапке *streams*. Дополнительную информацию смотрите в разделе “Папка demos” на стр. 4.

Построение подходящей модели

1. Добавьте узел источников файла статистики, указывающий на файл *telco.sav* в папке *Demos*.

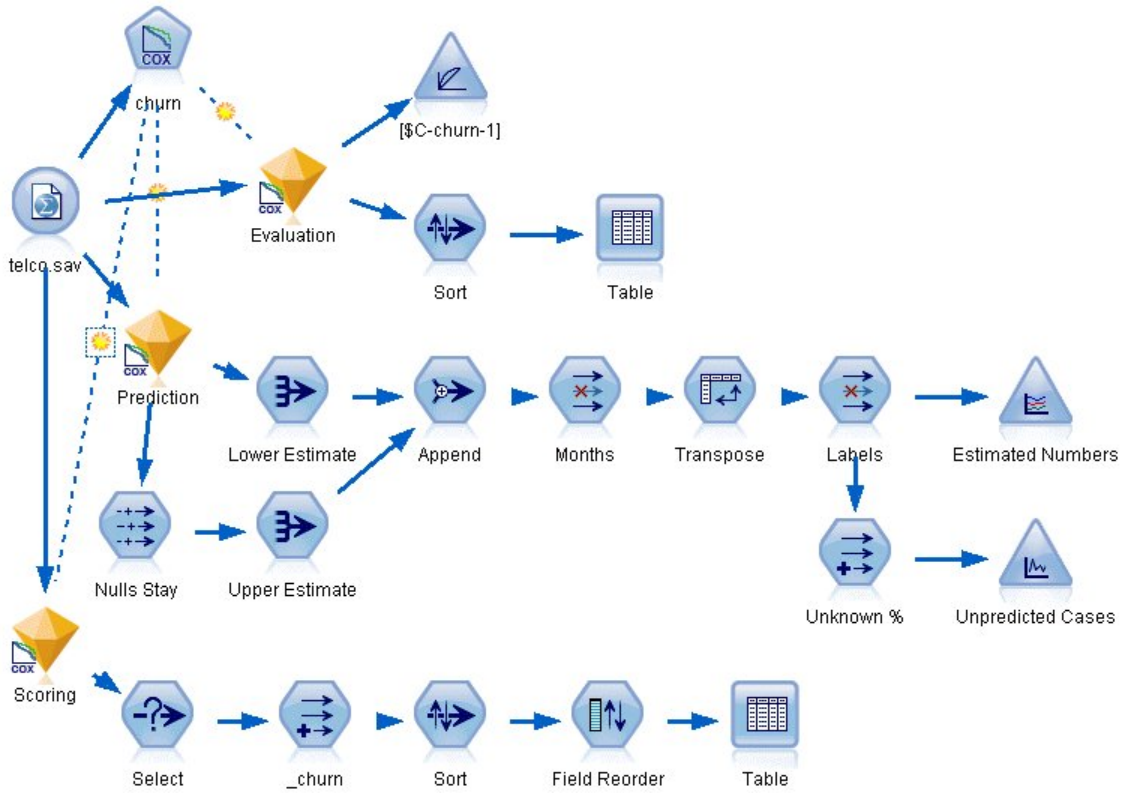


Рисунок 340. Пример потока для анализа времени до оттока

2. На вкладке Фильтр узла Источник исключите поля *region*, *income*, с *longten* по *wireten* и с *loglong* по *logwire*.

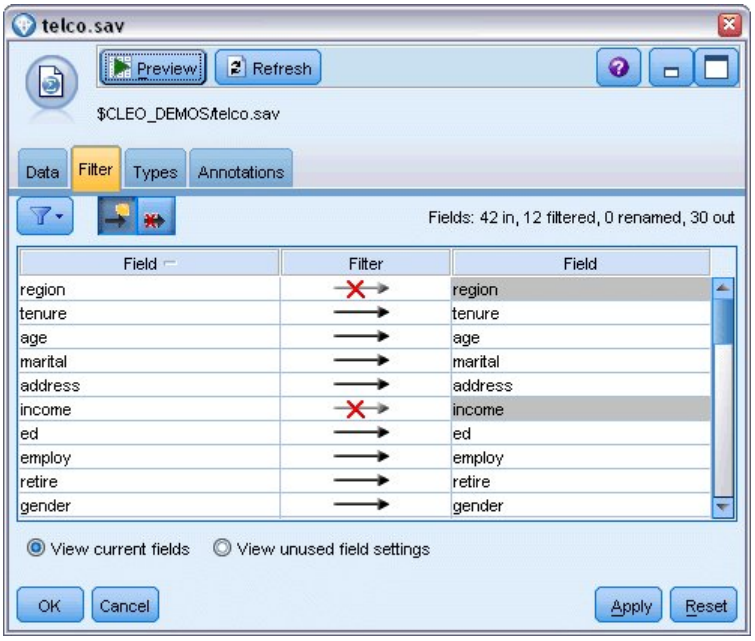


Рисунок 341. Фильтрация ненужных полей

(Другой вариант - не исключать эти поля, а задать для них роль **Нет** на вкладке Тип, или выбрать нужные поля в узле моделирования.)

3. На вкладке Типы узла Источник задайте для поля *churn* (отток) роль **Назначение** и тип измерений **Флаг**. Для всех остальных полей нужно задать роль **Ввод**.
4. Нажмите кнопку **Чтение данных**, чтобы создать экземпляр данных.

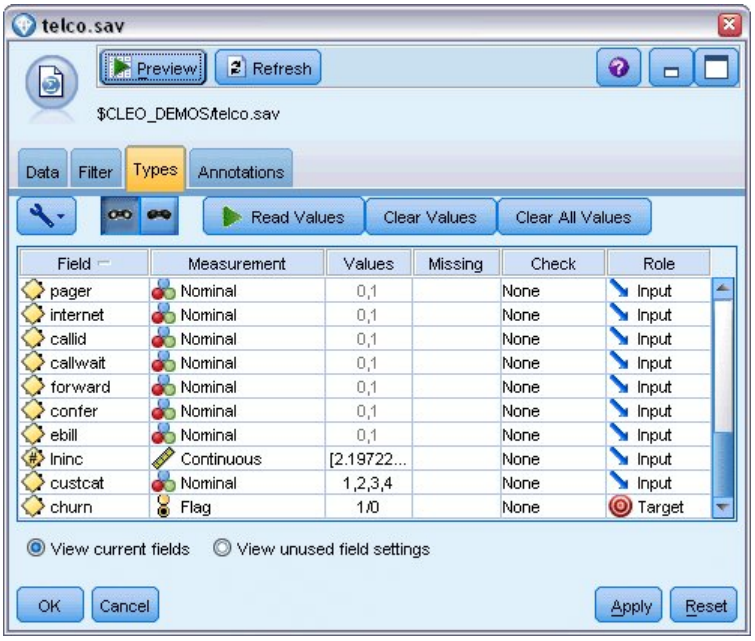


Рисунок 342. Задание роли поля

5. Подсоедините узел Кокс к узлу источника; на вкладке **Поля** выберите как переменную времени *tenure* (длительность предоставления услуг).

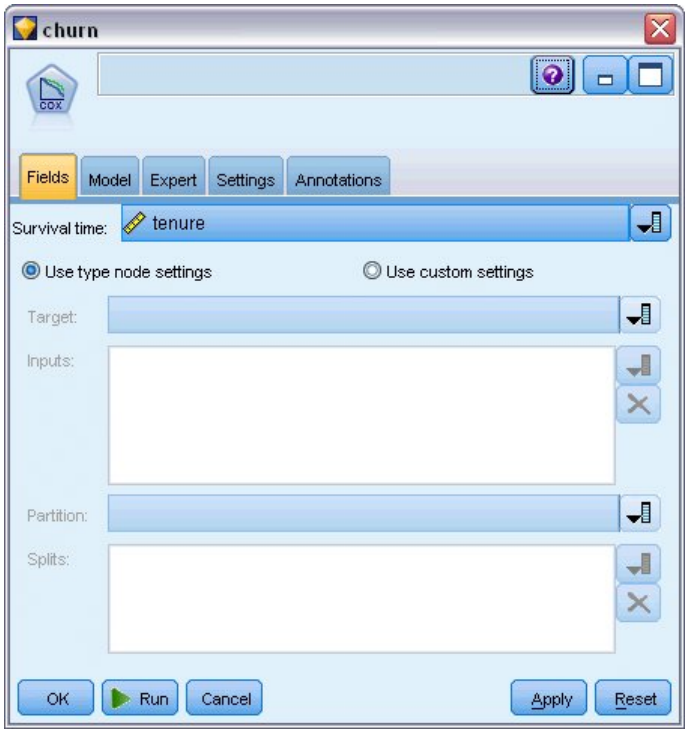


Рисунок 343. Выбор опции полей

6. Щелкните по вкладке **Модель**.
7. Выберите метод выбора переменной **Пошаговый**.

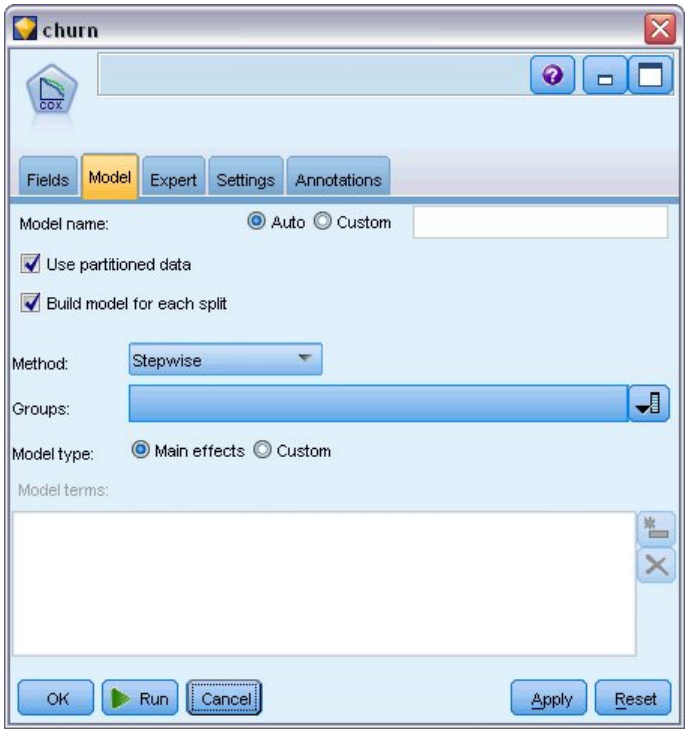


Рисунок 344. Выбор опций модели

8. Щелкните по вкладке **Эксперт** и выберите **Эксперт**, чтобы активировать экспертные опции моделирования.
9. Щелкните по **Вывод**.

Рисунок 345. Выбор дополнительных параметров вывода

10. Выберите генерируемые графики **Выживание** и **Риск**, затем нажмите **OK**.
11. Щелкните по **Запуск**, чтобы создать слепок модели, который добавляется в поток, а также на палитру Модели в верхнем правом углу. Чтобы посмотреть сведения о нем, дважды щелкните по слепку в потоке. Сначала посмотрим на вкладку Расширенный вывод.

Цензурированные наблюдения

		N	Percent
Cases available in analysis	Event[a]	274	27.4%
	Censored	726	72.6%
	Total	1000	100.0%
Cases dropped	Cases with missing values	0	.0%
	Cases with negative time	0	.0%
	Censored cases before the earliest event in a stratum	0	.0%
	Total	0	.0%
Total		1000	100.0%

a. Dependent Variable: Months with service

Рисунок 346. Сводный отчет обработки наблюдений

Переменная состояния задает, имело ли событие место для данного наблюдения. Если событие не имело место, про такое наблюдение говорят, что оно цензурируется. Цензурированные наблюдения не используются при вычислении коэффициентов регрессии, но используются при вычислении базового риска. В сводке обработки наблюдений показано 726 цензурированных наблюдений. Это те клиенты, которые не перешли к другим поставщикам.

Категориальное кодирование переменных

		Frequency	(1)[b]	(2)	(3)	(4)
marital[a]	0=Unmarried	505	1			
	1=Married	495	0			
ed[a]	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[a]	.00=No	953	1			
	1.00=Yes	47	0			
gender[a]	0=Male	483	1			
	1=Female	517	0			
tollfree[a]	0=No	526	1			
	1=Yes	474	0			
equip[a]	0=No	614	1			
	1=Yes	386	0			
callcard[a]	0=No	322	1			
	1=Yes	678	0			
wireless[a]	0=No	704	1			
	1=Yes	296	0			
multline[a]	0=No	525	1			
	1=Yes	475	0			
voice[a]	0=No	696	1			
	1=Yes	304	0			
pager[a]	0=No	739	1			
	1=Yes	261	0			
internet[a]	0=No	632	1			
	1=Yes	368	0			
callid[a]	0=No	519	1			
	1=Yes	481	0			
callwait[a]	0=No	515	1			
	1=Yes	485	0			
forward[a]	0=No	507	1			
	1=Yes	493	0			
confer[a]	0=No	498	1			
	1=Yes	502	0			
ebill[a]	0=No	629	1			
	1=Yes	371	0			
custcat[a]	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	

Рисунок 347. Кодировка категориальных переменных

На кодирование категориальных переменных, особенно дихотомических, удобно опираться при интерпретации коэффициентов регрессии категориальных ковариат. По умолчанию опорная категория - это "последняя" категория. Поэтому, например, клиент, который *Состоит в браке*, в файле данных задается значением переменной 1, но для целей регрессии кодируется как 0.

Выбор переменных

Step	-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
		Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
1[a]	3392.536	162.303	1	.000	133.828	1	.000	133.828	1	.000
2[b]	3087.314	249.392	2	.000	305.222	1	.000	439.050	2	.000
3[c]	3027.085	328.426	3	.000	60.229	1	.000	499.279	3	.000
4[d]	2990.790	347.197	4	.000	36.294	1	.000	535.574	4	.000
5[e]	2973.790	362.673	5	.000	17.000	1	.000	552.574	5	.000
6[f]	2958.796	376.140	6	.000	14.994	1	.000	567.568	6	.000
7[g]	2945.503	384.717	7	.000	13.293	1	.000	580.861	7	.000
8[h]	2936.993	417.341	8	.000	8.510	1	.004	589.371	8	.000
9[i]	2926.000	423.911	9	.000	10.994	1	.001	600.364	9	.000
10[j]	2917.551	428.078	10	.000	8.449	1	.004	608.813	10	.000
11[k]	2913.308	436.837	11	.000	4.243	1	.039	613.056	11	.000
12[l]	2908.078	440.158	12	.000	5.230	1	.022	618.286	12	.000

a. Variable(s) Entered at Step Number 1: callcard
b. Variable(s) Entered at Step Number 2: longmon
c. Variable(s) Entered at Step Number 3: equip
d. Variable(s) Entered at Step Number 4: employ
e. Variable(s) Entered at Step Number 5: multline
f. Variable(s) Entered at Step Number 6: voice
g. Variable(s) Entered at Step Number 7: address
h. Variable(s) Entered at Step Number 8: equipmon
i. Variable(s) Entered at Step Number 9: ebill
j. Variable(s) Entered at Step Number 10: callid
k. Variable(s) Entered at Step Number 11: internet
l. Variable(s) Entered at Step Number 12: reside
m. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 3526.364
n. Beginning Block Number 1. Method = Forward Stepwise (Likelihood Ratio)

Рисунок 348. Универсальные критерии

В процессе построения модели используется прямой шаговый алгоритм. Универсальные критерии - это показатели того, насколько хорошо работает модель. Изменение хи-квадрата по сравнению с предыдущим шагом - это разность между двоичным логарифмом правдоподобия модели на предыдущем шаге и на текущем шаге. Если шаг добавлял переменную, добавление оправдано при значимости изменения менее 0,05. Если шаг удалял переменную, удаление оправдано при значимости изменения более 0,10. За двенадцать шагов в модель добавлены двенадцать переменных.

		B	SE	Wald	df	Sig.	Exp(B)
Step 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

Рисунок 349. Переменные в уравнении (только шаг 12)

Окончательная модель содержит переменные *address*, *employ*, *reside*, *equip*, *callcard*, *longmon*, *equipmon*, *multline*, *voice*, *internet*, *callid* и *ebill* (адрес, занятость, место жительства, оборудование, телефонная карта, longmon, equipmon, несколько линий, голосовые услуги, интернет, id вызова и электронный счет). Чтобы понять влияние отдельных предикторов, посмотрим на выражение Exp(B), которое можно трактовать как предсказанное изменение риска при изменении предиктора на единицу.

- Значение Exp(B) для переменной *address* означает, что риск оттока уменьшается на 100%–(100 %×0,966)=3,4% с каждым годом, прожитым клиентом по неизменному адресу. Риск оттока для клиента, прожившего пять лет по неизменному адресу, уменьшается на $100\%-(100\%\times0{,}966^5)=15{,}88\%$.
- Значение Exp(B) для переменной *callcard* означает, что риск оттока уменьшается для клиента, не подписавшегося на услугу телефонной карты, в 2,175 раза по сравнению с клиентом, подписанным на такую услугу. Вспомним, что кодированием категориальных переменных для регрессии задано *Нет* = 1.
- Значение Exp(B) для переменной *internet* означает, что риск оттока уменьшается для клиента, не подписавшегося на услугу интернет-доступа, в 0,697 раза по сравнению с клиентом, подписанным на такую услугу. Это тревожно, поскольку показывает, что клиенты, пользующиеся услугой, покидают компанию быстрее, чем клиенты без этой услуги.

		Score	df	Sig.
Step 12	age	.122	1	.726
	marital	.648	1	.421
	income	1.476	1	.224
	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
	custcat	.864	3	.834
	custcat(1)	.466	1	.495
	custcat(2)	.450	1	.502
	custcat(3)	.019	1	.889

Рисунок 350. Переменные, не вошедшие в модель (только шаг 12)

Все переменные, оставленные за пределами модели, получили при оценке значимость больше 0,05. Однако значимость переменных *tollfree* и *cardmon*, хотя и превышала 0,05, но не намного. Эти переменные могут представлять интерес при дальнейших исследованиях.

Средние ковариат

	Mean
age	41.684
marital	.505
address	11.551
income	77.535
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
custcat(1)	.266
custcat(2)	.217
custcat(3)	.281

Рисунок 351. Средние ковариат

Эта таблица содержит среднее значение каждой переменной-предиктора. Эта таблица полезна как справочник при рассмотрении графиков выживания, построенных для средних значений. Имейте в виду, однако, рассматривая средние значения индикаторных переменных для категориальных предикторов, что "средний" клиент на самом деле не существует. Даже когда все предикторы - количественные, вы вряд ли найдете клиента, у которого значения всех ковариат близки к средним. Если нужно посмотреть кривую выживания для конкретного наблюдения, можете изменить значения ковариат, при которых строится кривая выживания, в диалоговом окне Графики. Если нужно посмотреть кривую выживания для конкретного наблюдения, можете изменить значения ковариат, при которых строится кривая выживания, в группе Графики диалогового окна Расширенный вывод.

Кривая выживания

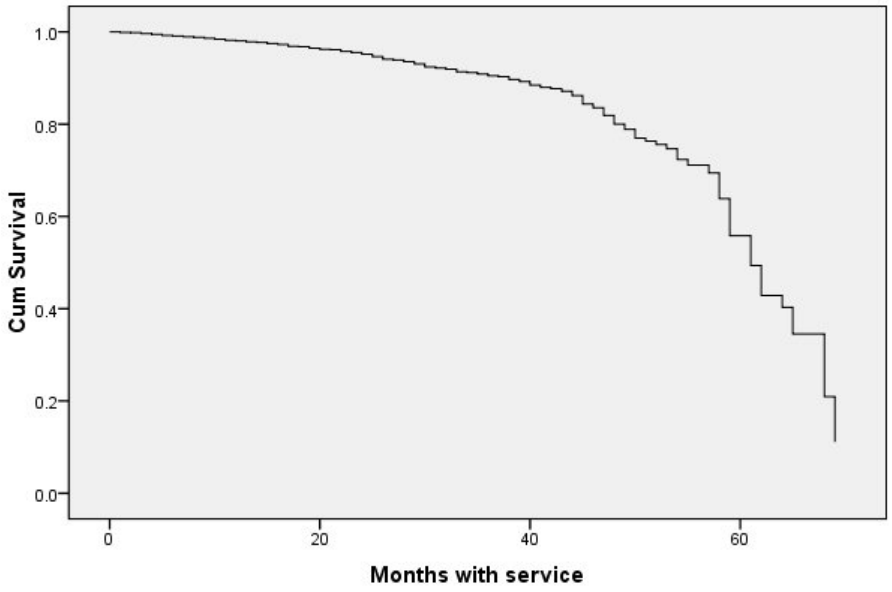


Рисунок 352. Кривая выживания для "среднего" клиента

Основная кривая выживания - это наглядное представление предсказанного моделью времени до оттока для "среднего" клиента. По горизонтальной оси отложено время до события. По вертикальной оси отложена вероятность выживания. Таким образом, всякая точка на кривой выживания показывает вероятность, что "средний" клиент останется клиентом после данного времени. После 55 месяцев кривая выживания теряет гладкость. Остается немного клиентов, пробывших с компанией столь долгое время, и в связи со скудостью доступной информации кривая идет уступами.

Кривая риска

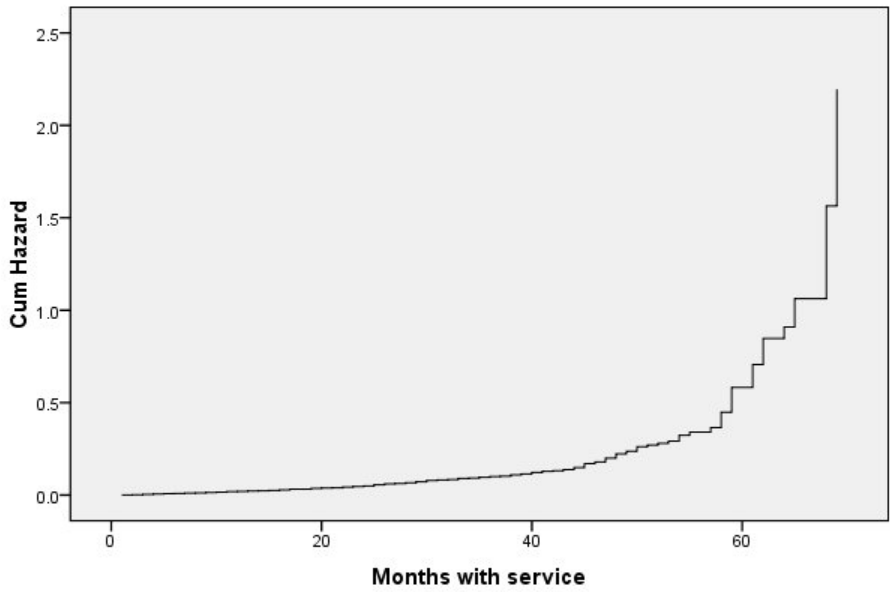


Рисунок 353. Кривая риска для "среднего" клиента

Основная кривая риска - это наглядное представление кумулятивного предсказанного моделью потенциала к оттоку для "среднего" клиента. По горизонтальной оси отложено время до события. По вертикальной оси отложен кумулятивный риск, равный логарифму вероятности выживания со знаком минус. После 55 месяцев кривая риска, как и кривая выживания, теряет гладкость, по той же причине.

Оценка

Пошаговые методы выбора гарантируют, что в модели будут только "статистически значимые" предикторы, но не гарантируют, что модель будет хорошо прогнозировать поле назначения. Чтобы гарантировать такое, необходимо проанализировать оцененные записи.

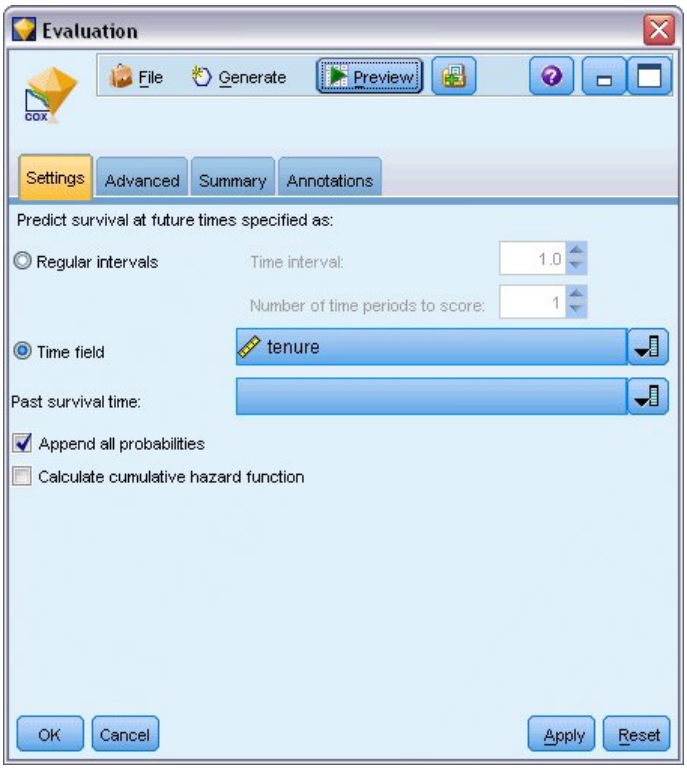


Рисунок 354. Слепок Cox: Вкладка Параметры

1. Поместите слепок модели на холст и подсоедините его к узлу источника, откройте слепок и щелкните по вкладке Параметры.
2. Выберите **Поле времени** и задайте *tenure* (длительность предоставления услуг). Каждая запись будет оценена по длительности предоставления услуг.
3. Выберите **Добавить все вероятности**.

   Будут созданы оценки ухода клиента к другим поставщикам с порогом отсечения 0,5; если склонность клиента к оттоку выше 0,5, клиент будет оцениваться как уходящий. Такое пороговое значение не следует считать чудодейственным средством на все случаи жизни, и иногда лучшие результаты дают другие пороговые значения. Один из способов выбрать порог отсечения - воспользоваться узлом Оценка.

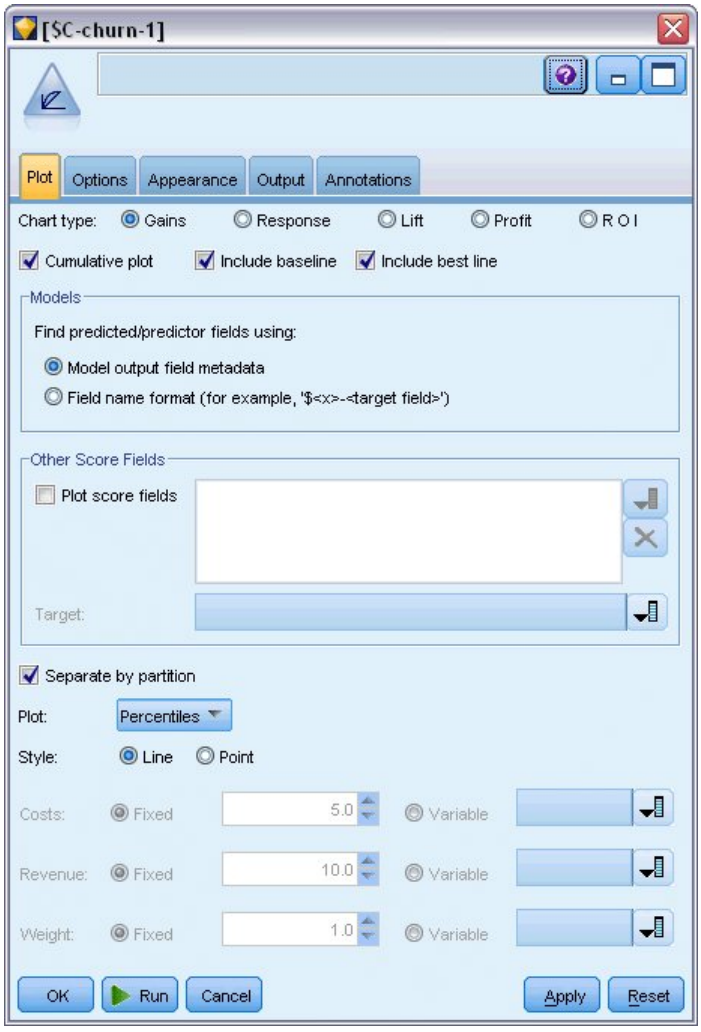


Рисунок 355. Узел Оценка: Вкладка График

4. Подсоедините узел Оценка к слепку модели; на вкладке График выберите **Включить лучший уровень**.
5. Щелкните по вкладке **Параметры**.

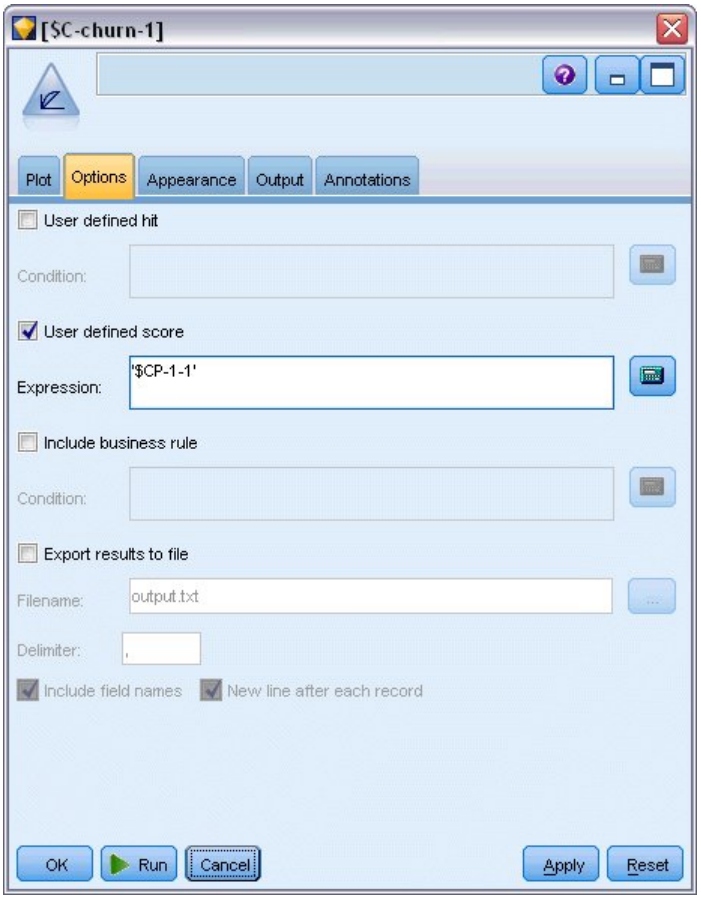


Рисунок 356. Узел Оценка: Вкладка Параметры

6. Выберите **Пользовательская оценка** и введите выражение `'$CP-1-1'`. Это генерируемое моделью поле, соответствующее склонности к оттоку.
7. Нажмите кнопку **Выполнить**.

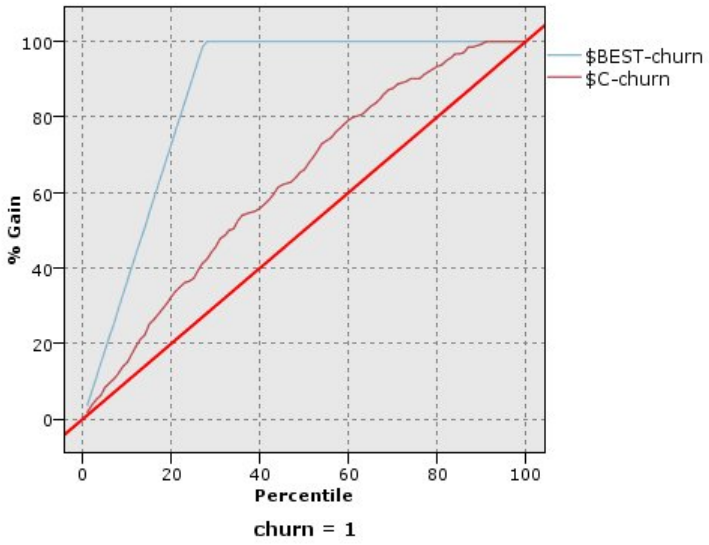


Рисунок 357. Диаграмма выигрыша

Кумулятивная диаграмма выигрыша содержит процент общего числа наблюдений в данной категории, "выигранное" нацеливанием на процента от общего числа наблюдений. Например, одна точка на кривой имеет координаты (10%, 15%), и это значит, что, если оценить набор данных при помощи модели и отсортировать все наблюдения по прогнозируемой склонности к оттоку, то ожидается, что верхние 10% содержат примерно 15% всех наблюдений, которые фактически попадут в категорию *1* (уходящие). Аналогичным образом верхние 60% содержат примерно 79,2% уходящих. Если выбрать 100% оцененного набора данных, получим всех уходящих в наборе данных.

Диагональ представляет "базовую" кривую; если случайно выбрать 20% записей из оцененного набора данных, ожидаемый "выигрыш" составит 20% всех записей, которые фактически попадут в категорию *1*. Чем выше проходит кривая над базовой, тем больше выигрыш. "Оптимальная" линия задает кривую "совершенной" модели, которая всем уходящим отток назначает более высокую оценку склонности к оттоку, чем остальным. При помощи кумулятивной диаграммы выигрыша удобно выбрать порог отсечения классификации как процент, который соответствует желательному выигрышу, и затем отобразить этот процент на соответствующее значение порога отсечения.

Величина "желательного" выигрыша зависит от стоимости ошибок типа I и типа II. Какова стоимость ошибочного классифицирования уходящего как не уходящего (тип I)? Какова стоимость ошибочного классифицирования не уходящего клиента как уходящего (тип II)? Если важнее прежде всего сохранить клиента, следует сократить ошибки типа I; например, можно на кумулятивной диаграмме выигрыша уделить повышенное внимание клиентам в верхних 60% с предсказанной склонностью попасть в категорию *1*, в которые попадают 79,2% потенциального оттока, однако вы потратите время и ресурсы, которые можно было израсходовать на привлечение новых клиентов. Если приоритет - снизить расходы на сохранение текущей клиентской базы, следует сократить ошибки типа II. На этой диаграмме можно, соответственно, уделить повышенное внимание клиентам в верхних 20%, в которые попадают 32,5% всех уходящих. Обычно важны оба фактора, и приходится подбирать правило решения при классификации клиентов, дающее оптимальное сочетание чувствительности и избирательности.

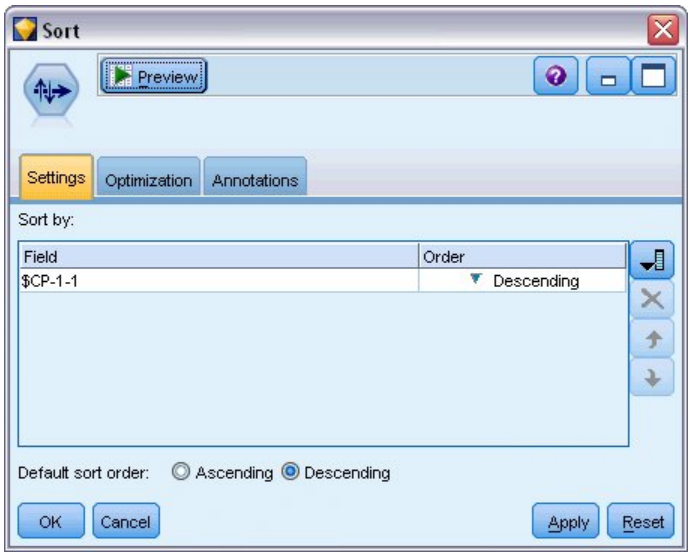


Рисунок 358. Узел Сортировка: Вкладка Параметры

8. Пусть вы решили, желательный выигрыш составляет 45,6%, которые соответствуют верхним 30% записей. Чтобы найти нужный порог отсечения классификации, подсоедините к слепку модели узел сортировки.
9. На вкладке Параметры выберите сортировку по *$CP-1-1* в убывающем порядке и нажмите **OK**.

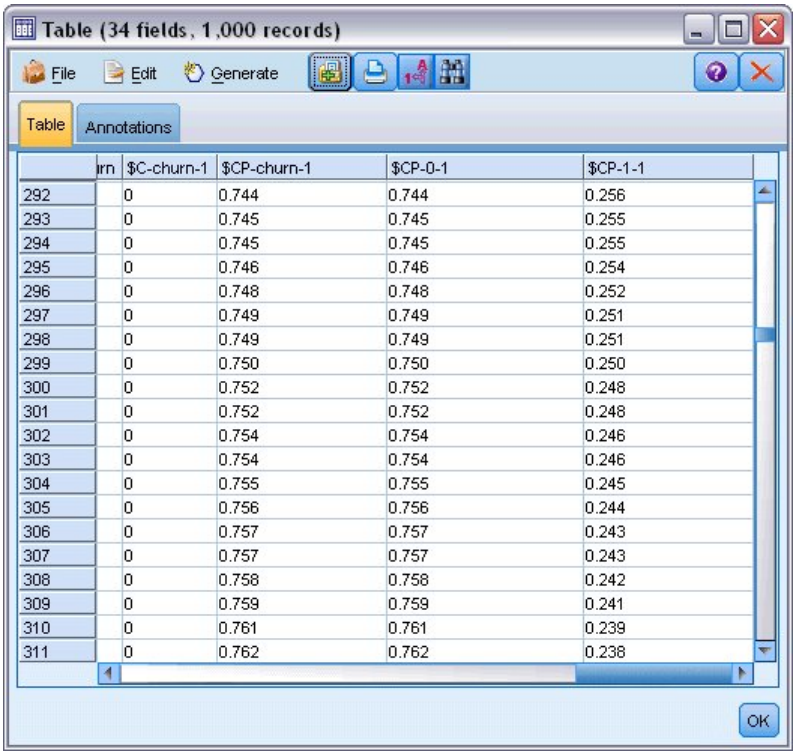


Рисунок 359. Таблица

10. Присоедините узел Таблица к узлу Сортировка.
11. Откройте узел Таблица и щелкните по **Выполнить**.

Прокрутив результаты, вы увидите, что значение *$CP-1-1* составляет 0,248 для 300-й записи. Если задать 0,248 как порог отсечения классификации, примерно 30% клиентов, оцененных как уходящие, охватят примерно 45% общего числа фактически уходящих.

Отслеживание ожидаемого числа сохраненных клиентов

Получив удовлетворительную модель, вы хотите проследить ожидаемое число клиентов в наборе данных, которые сохранится на протяжении последующих двух лет. Нетривиальный вызов представляют собой пустые значения, при которых у клиента общая длительность предоставления услуг (будущее время + *длительность предоставления услуг*) выпадает за пределы диапазона времени выживания в данных, использованных для обучения модели. Один из способов обработать эти данные - создать два набора прогнозов, в одном из которых клиенты с пустыми значениями отнесены к оттоку, а в другом - к сохранившемся клиентам. Таким образом можно установить верхние и нижние границы для ожидаемого числа сохраненных клиентов.

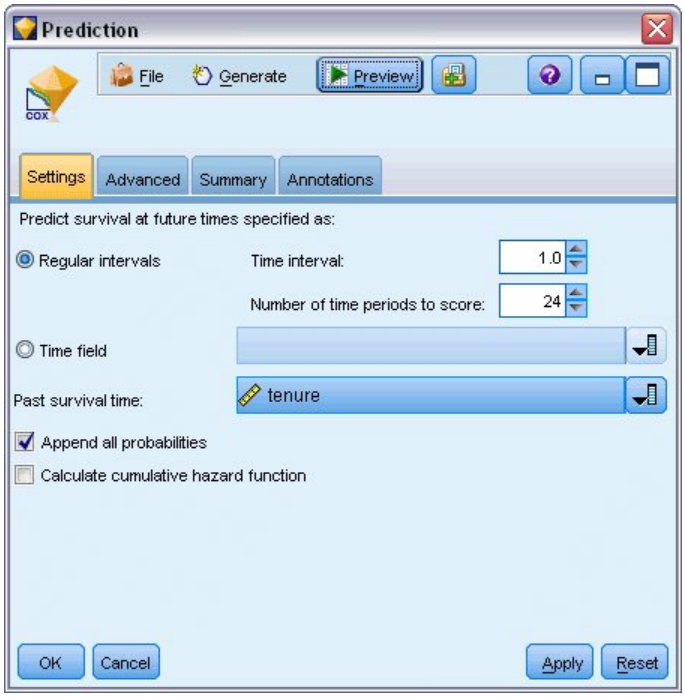


Рисунок 360. Слепок Cox: Вкладка Параметры

1. Дважды щелкните по слепку модели на палитре моделей (или скопируйте и вставьте слепок на холст потока), и подсоедините новый слепок к узлу Источник.
2. Откройте слепок и перейдите на вкладку Параметры.
3. Убедитесь, что включен переключатель **Определенные интервалы**, и укажите временной интервал `1,0` и число оцениваемых периодов `24`. Тем самым вы зададите, что каждая запись будет оценена для каждого из последующих 24 месяцев.
4. Выберите *tenure* (длительность предоставления услуг) как поле, задающее прошедшее время выживания. Алгоритм для оценки учитывает продолжительность времени каждого заказчика как заказчика компании.
5. Выберите **Добавить все вероятности**.

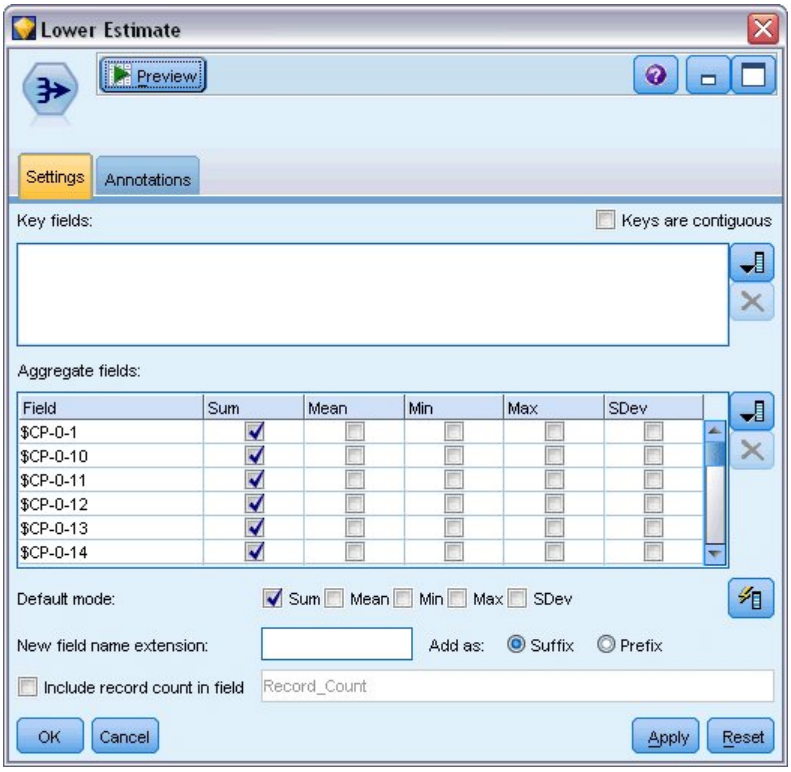


Рисунок 361. Узел Агрегирование: Вкладка Параметры

6. Подсоедините узел агрегирования к слепку модели; на вкладке Параметры выключите переключатель **Среднее** как режим по умолчанию.
7. Выберите от *$CP-0-1* по *$CP-0-24* поля формы *$CP-0-n*, как поля для агрегирования. Это легче всего сделать, если в диалоговом окне Выбрать поля отсортировать поля по имени (то есть в алфавитном порядке).
8. Отмените выбор **Включить количество записей в поле**.
9. Щелкните по **OK**. Этот узел создает прогнозы "нижней границы".

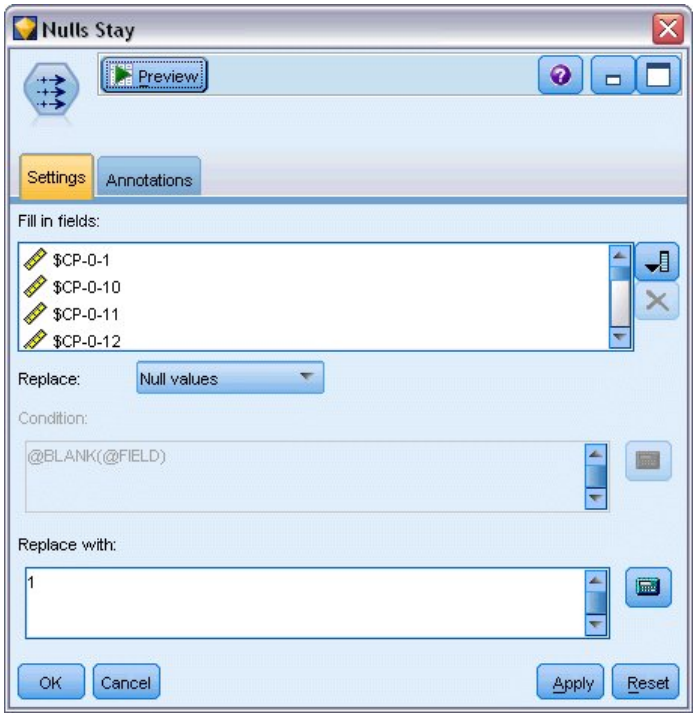


Рисунок 362. Узел Заполнение: Вкладка Параметры

10. Подсоедините узел заполнения к слепку Coxreg, к которому мы сейчас подсоединили узел агрегирования; на вкладке Параметры выберите с *$CP-0-1* по *$CP-0-24* из полей вида *$CP-0-n* как поля для заполнения. Это легче всего сделать, если в диалоговом окне Выбрать поля отсортировать поля по имени (то есть в алфавитном порядке).
11. Выберите, что **Пустые значения** заменяются на значение 1.
12. Щелкните по **OK**.

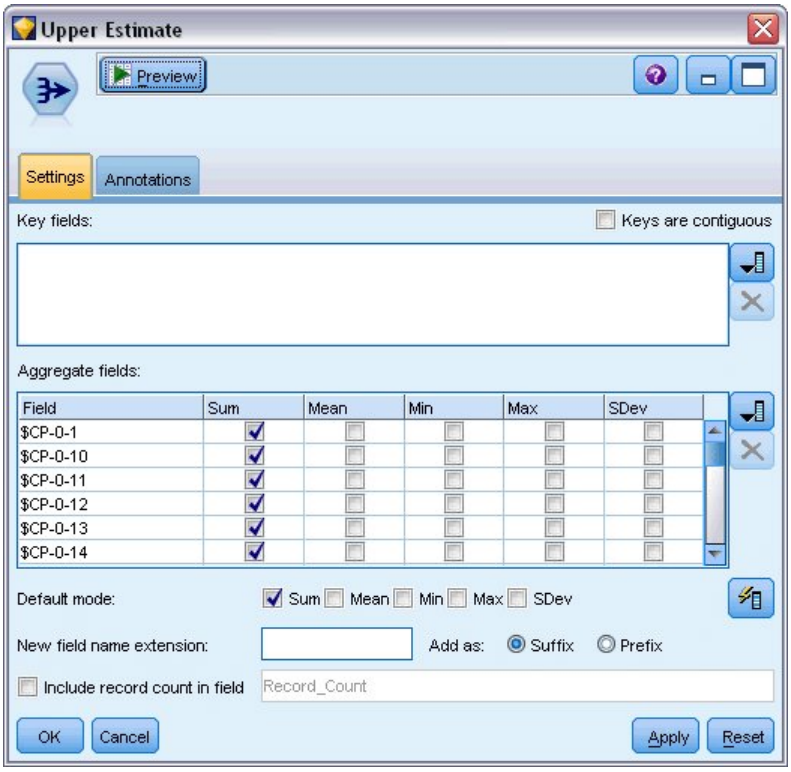


Рисунок 363. Узел Агрегирование: Вкладка Параметры

13. Подсоедините узел агрегирования к узлу заполнения; на вкладке Параметры выключите переключатель **Среднее** как режим по умолчанию.
14. Выберите от *$CP-0-1* по *$CP-0-24* поля формы *$CP-0-n*, как поля для агрегирования. Это легче всего сделать, если в диалоговом окне Выбрать поля отсортировать поля по имени (то есть в алфавитном порядке).
15. Отмените выбор **Включить количество записей в поле**.
16. Щелкните по **OK**. Этот узел создает прогнозы "верхней границы".

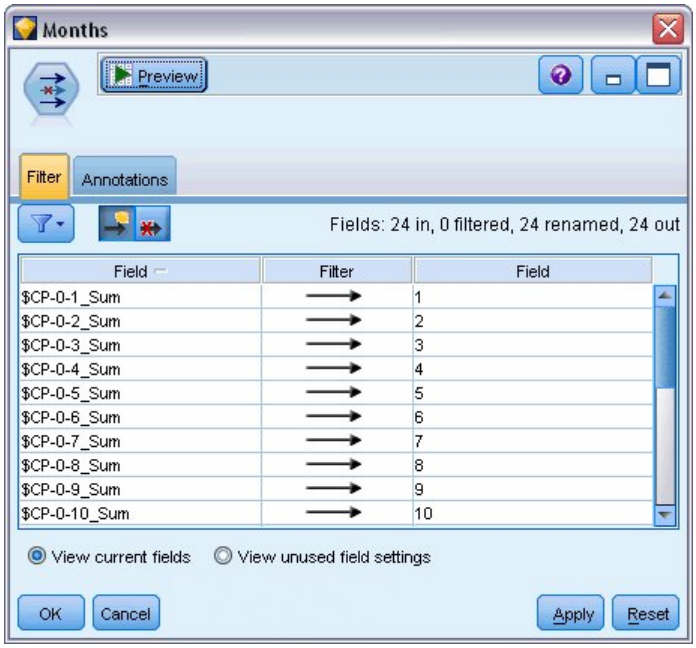


Рисунок 364. Узел Фильтр: Вкладка Параметры

17. Подсоедините узел добавления к двум узлам агрегирования, затем подсоедините узел Фильтр к узлу добавления.
18. На вкладке Параметры узла Фильтр переименуйте поля с *1* по *24*. При помощи узла транспонирования эти имена полей станут значениями оси *x* на диаграммах последующих узлов.

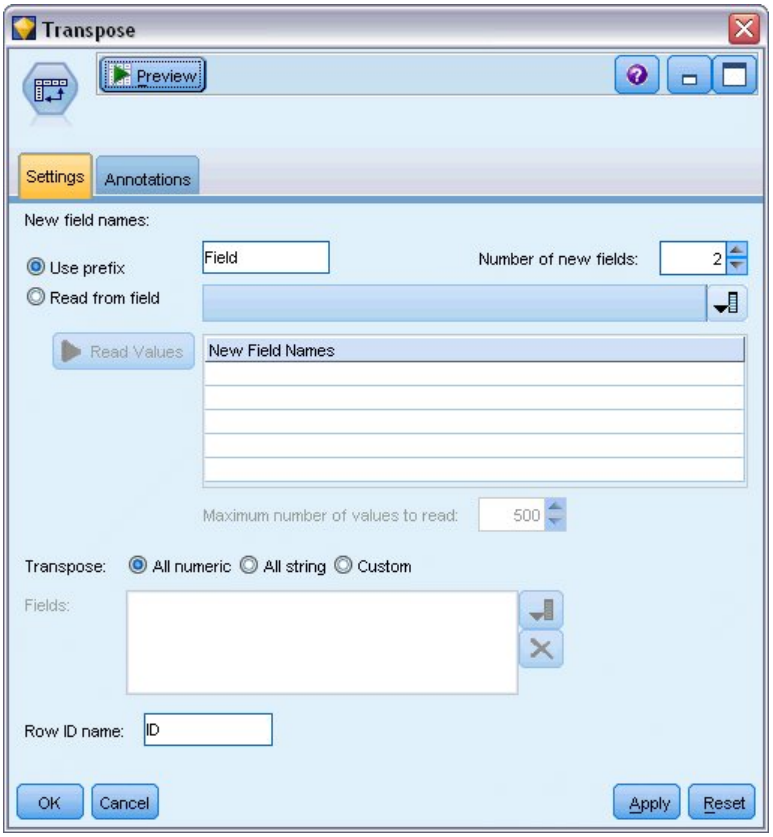


Рисунок 365. Узел Транспонирование: Вкладка Параметры

19. Присоедините узел Транспонирование к узлу Фильтр.
20. Введите 2 как число новых полей.

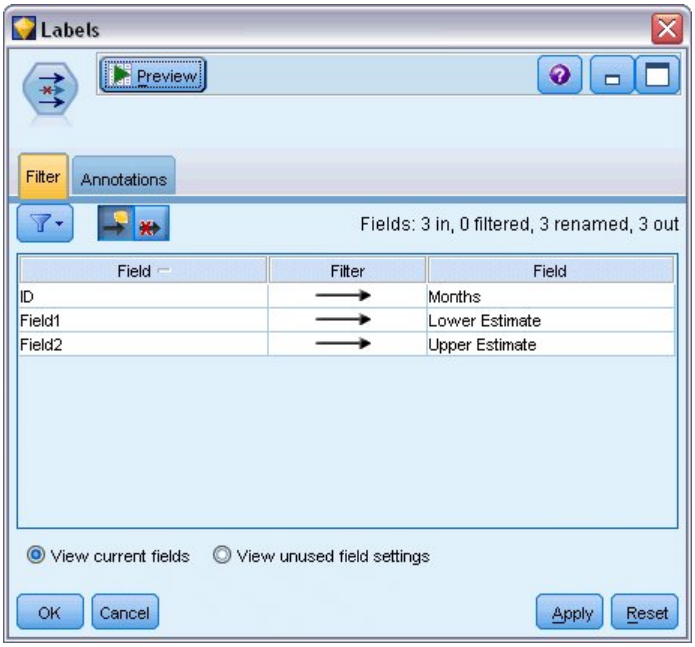


Рисунок 366. Узел Фильтр: Вкладка Фильтр

21. Присоедините узел Фильтр к узлу Транспонирование.

22. На вкладке Параметры узла Фильтр переименуйте *ID* в *Месяцы*, *Field1* в *Нижняя оценка* и *Field2* в *Верхняя оценка*.

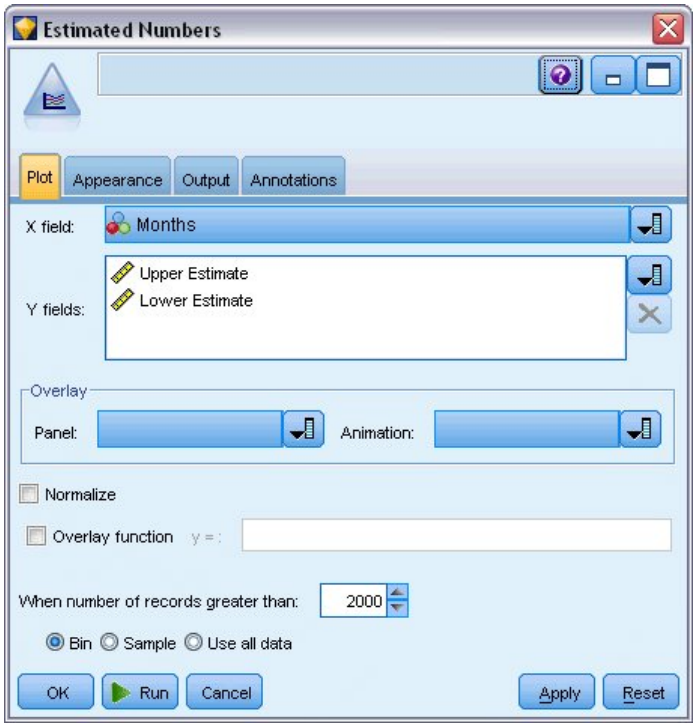


Рисунок 367. Узел Мультиграфик: Вкладка График

23. Присоедините узел Мультиграфик к узлу Фильтр.
24. На вкладке График задайте *Месяцы* как поле X, *Нижняя оценка* и *Верхняя оценка* как поля Y.

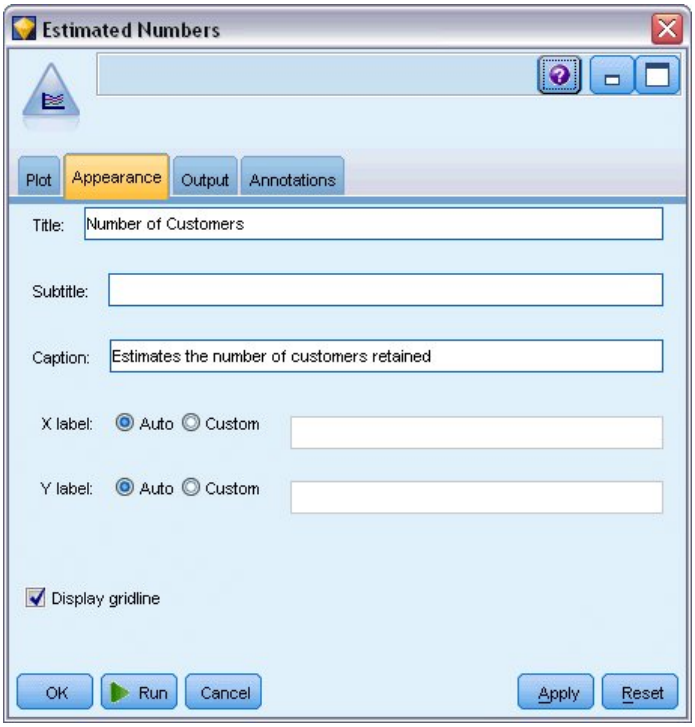


Рисунок 368. Узел Мультиграфик: Вкладка Внешний вид

25. Перейдите на вкладку Внешний вид.
26. Введите `Число клиентов` как заголовок.
27. Введите `Оценивает число сохраненных клиентов` как пояснительный заголовок.
28. Нажмите кнопку **Выполнить**.

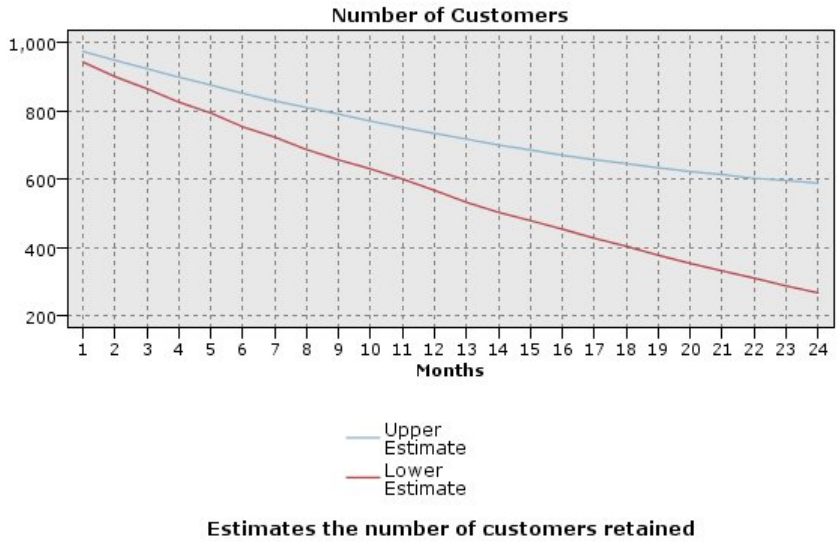


Рисунок 369. Мультиграфик оценки числа сохраненных клиентов

На графике показаны верхние и нижние границы для предполагаемого числа сохраненных клиентов. Разность между двумя линиями равна числу клиентов, оцененных как пустое значение, так что их состояние имеет высокую неопределенность. Со временем число таких клиентов возрастает. После 12 месяцев ожидается сохранение от 601 до 735 первоначальных клиентов в наборе данных; после 24

месяцев - от 288 до 597.

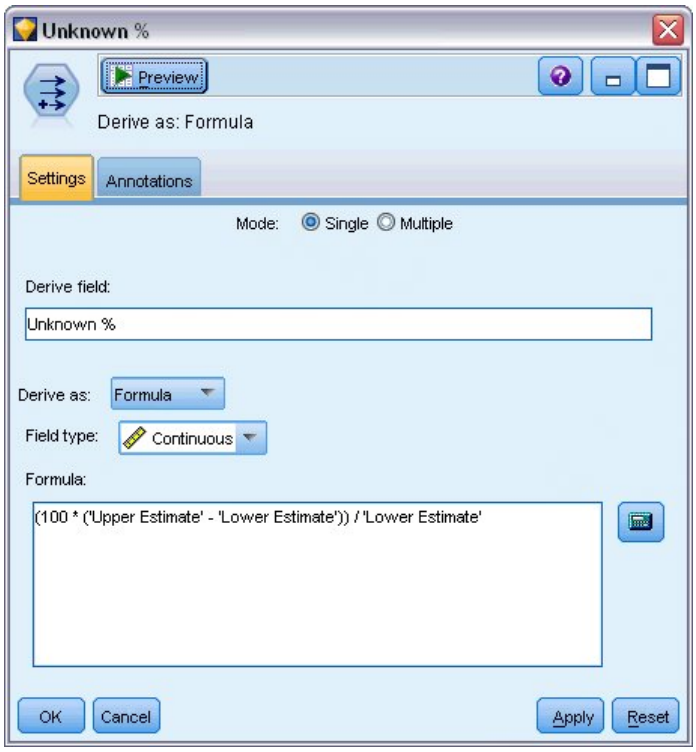


Рисунок 370. Узел извлечения: Вкладка Параметры

29. Чтобы взглянуть с другой стороны на неопределенность оценки числа сохраненных клиентов, подсоедините к узлу Фильтр узел вычислений.
30. На вкладке Параметры узла вычислений введите *Unknown %* (% неизвестных) как производное поле.
31. Выберите тип поля **Непрерывный**.
32. Введите `(100 * ('Upper Estimate' - 'Lower Estimate')) / 'Lower Estimate'` как формулу. *Unknown %* - это число "сомнительных" клиентов как процент от нижней оценки.
33. Щелкните по **OK**.

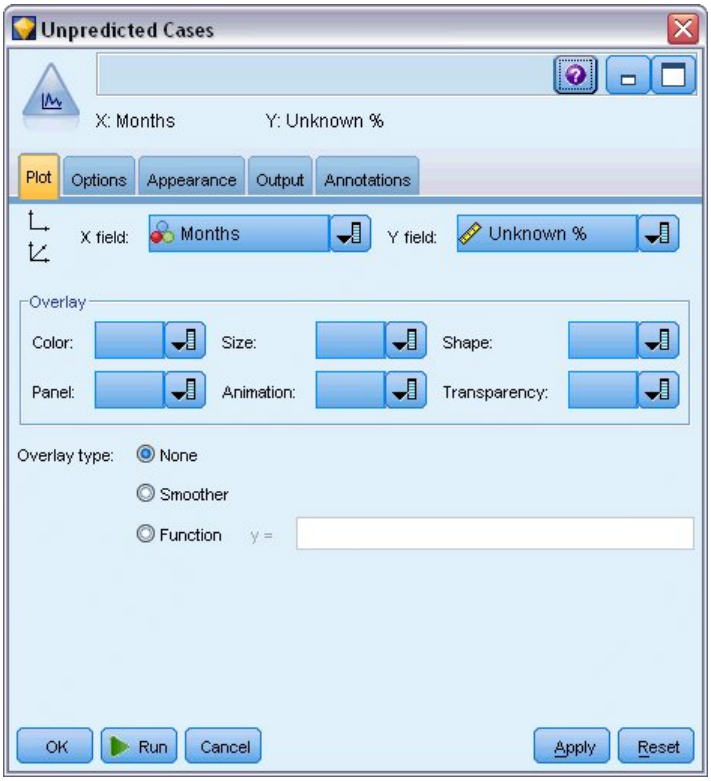


Рисунок 371. Узел График: Вкладка График

34. Подсоедините к узлу вычислений узел График.
35. На вкладке График узла График выберите *Месяцы* как поле X и *Unknown %* как поле Y.
36. Щелкните по вкладке **Внешний вид**.

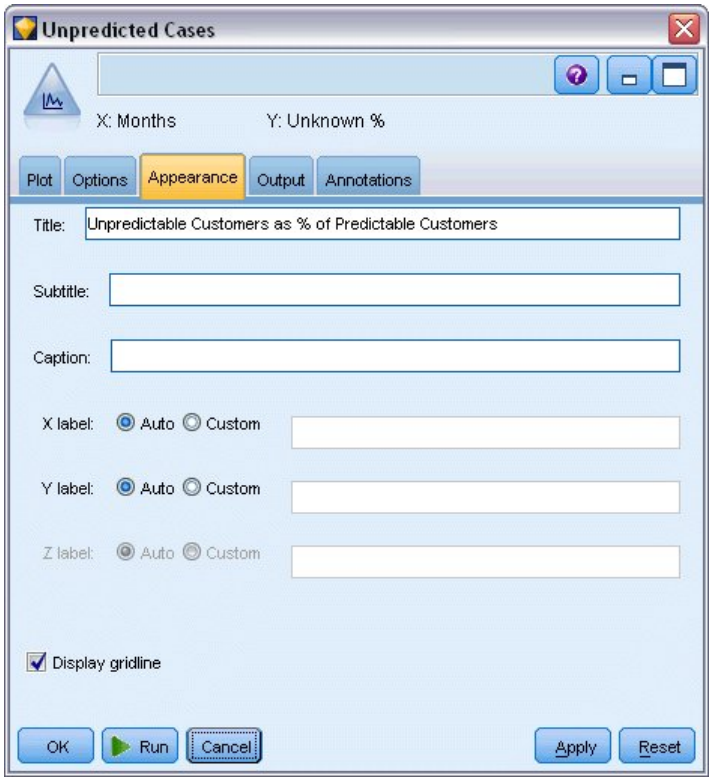


Рисунок 372. Узел График: Вкладка Внешний вид

37. Введите как заголовок `Число непредсказуемых клиентов как % от числа предсказуемых клиентов`.
38. Выполните узел.

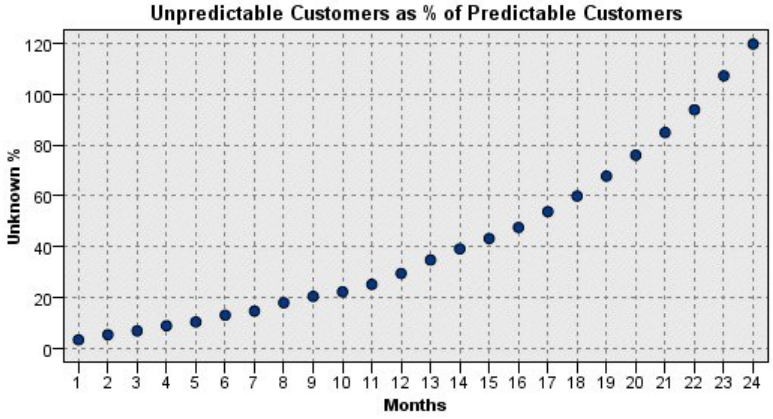


Рисунок 373. График непредсказуемых покупателей

В течение первого года процент непредсказуемых клиентов возрастает довольно близко к линейному закону, но рост резко ускоряется в течение второго года, а к месяцу 23 число клиентов с пустыми значениями превосходит ожидаемое число сохраненных клиентов.

Скоринг

Получив удовлетворительную модель, вы хотите оценить клиентов, чтобы выявить тех, кто с высокой вероятностью перейдет к другим поставщикам в течение кварталов следующего года.

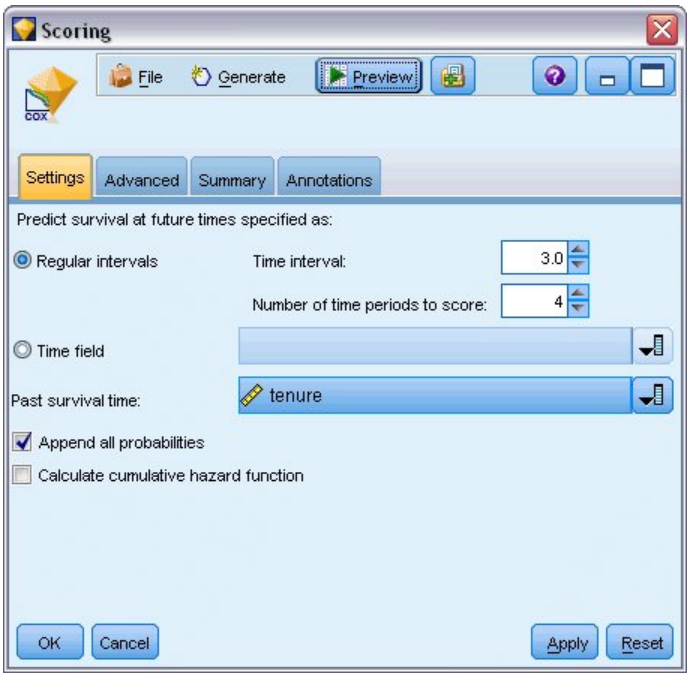


Рисунок 374. Слепок Coxreg: Вкладка Параметры

1. Подсоедините третий слепок модели к узлу Источник и откройте слепок модели.
2. Убедитесь, что включен переключатель **Определенные интервалы**, и укажите временной интервал `3,0` и число оцениваемых периодов `4`. Тем самым вы зададите, что каждая запись будет оценена для каждого из последующих четырех кварталов.
3. Выберите *tenure* (длительность предоставления услуг) как поле, задающее прошедшее время выживания. Алгоритм для оценки учитывает продолжительность времени каждого заказчика как заказчика компании.
4. Выберите **Добавить все вероятности**. Эти дополнительные поля облегчат сортировку записей при просмотре в таблице.

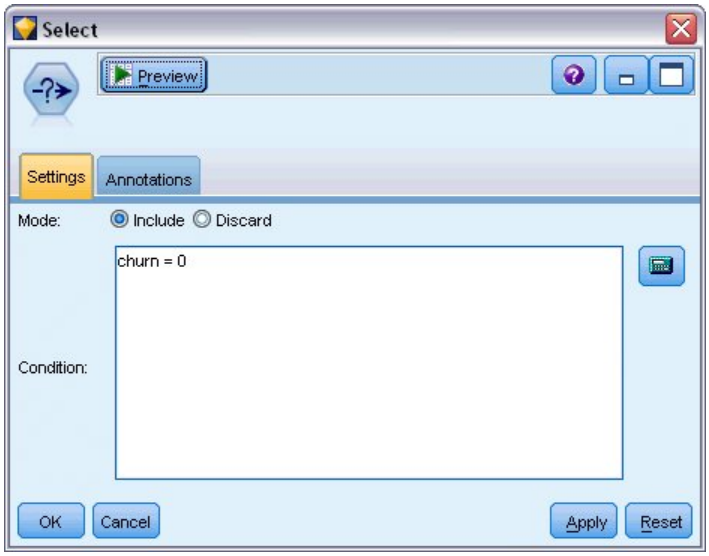


Рисунок 375. Узел Выбор: Вкладка Параметры

5. Подсоедините узел выбора к слепку модели; на вкладке Параметры введите условие `churn=0`. Этим из таблицы результатов будут удалены клиенты, которые уже перешли к другим поставщикам.

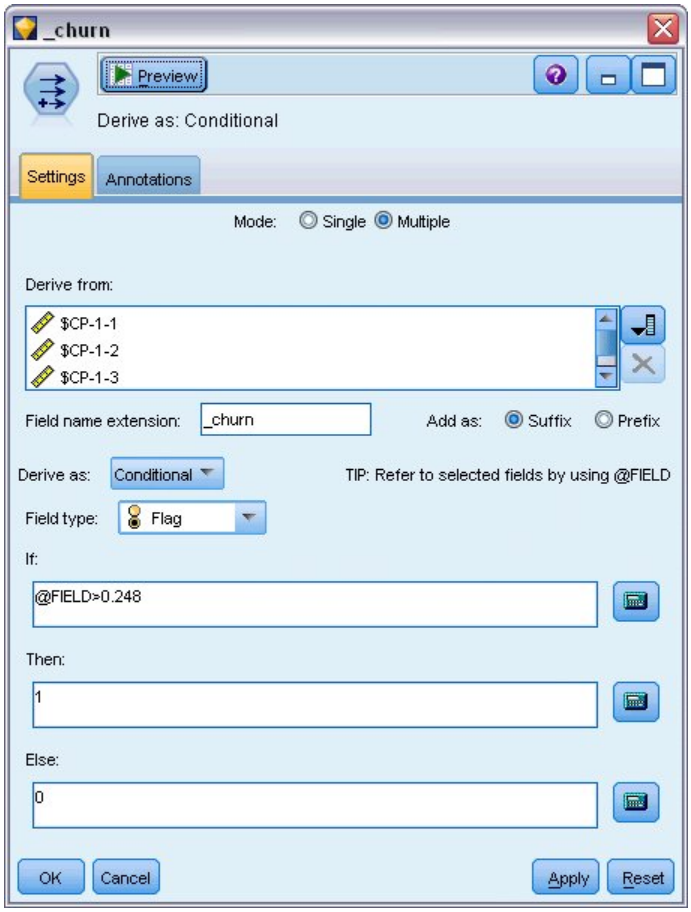


Рисунок 376. Узел извлечения: Вкладка Параметры

6. Подсоедините узел вычислений к узлу выбора; на вкладке Параметры выберите режим **Несколько**.

7. Выберите для вычислений поля с *$CP-1-1* по *$CP-1-4* из полей вида *$CP-1-n* и введите добавляемый суффикс `_churn`. Это легче всего сделать, если в диалоговом окне Выбрать поля отсортировать поля по имени (то есть в алфавитном порядке).
8. Выберите, что производное поле будет **Условное**.
9. Выберите тип измерений **Флаг**.
10. Введите `@FIELD>0.248` как условие **If**. Напомним, что таков был порог отсечения классификации, который мы выявили во время Оценки.
11. Введите `1` как выражение **Then**.
12. Введите `0` как выражение **Else**.
13. Щелкните по **OK**.

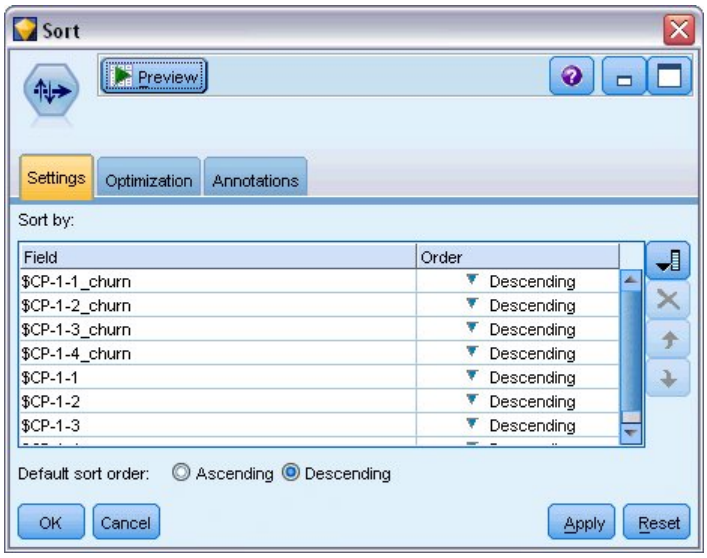
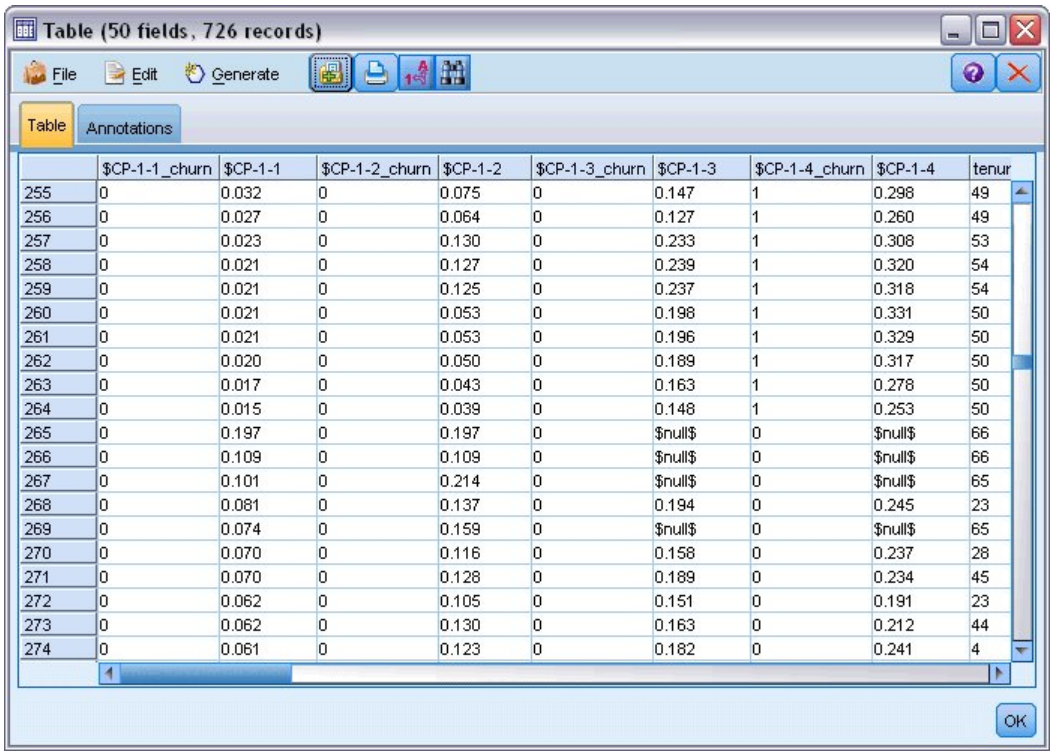


Рисунок 377. Узел Сортировка: Вкладка Параметры

14. Подсоедините узел сортировки к узлу вычислений; на вкладке Параметры выберите сортировку по полям с *$CP-1-1_churn* по *$CP-1-4-churn* и затем по полям с *$CP-1-1* по *$CP-1-4*, всюду в убывающем порядке. Клиенты, которые по прогнозу попадут в отток, окажутся в начале списка.

Рисунок 378. Узел переупорядочения полей: вкладка Переупорядочить

15. Подсоедините узел переупорядочения полей к узлу сортировки; на вкладке Переупорядочить выберите, что поля с *$CP-1-1_churn* по *$CP-1-4* идут перед остальными полями. Это только для удобства чтения таблицы результатов; этот шаг не обязателен. Чтобы переместить поля так, как показано на рисунке, используйте кнопки.



	$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	tenur
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

Рисунок 379. Таблица, содержащая оценки клиентов

16. Подсоедините узел Таблица к узлу переупорядочения полей и выполните его.

Ожидается, что 264 клиентов перейдут к другим поставщикам к концу года, 184 к концу третьего квартала, 103 к концу второго и 31 за первый квартал. Обратите внимание на то, что для данных двух клиентов тот, у которого выше склонность к оттоку в первом квартале, не обязательно имеет более высокую склонность к оттоку в последующих кварталах; пример - записи 256 и 260. Обычно причина в форме функции риска для месяцев, следующих после текущего периода предоставления услуг данному клиенту; например, если клиенты пришли благодаря рекламной кампании, они и уйдут с большей вероятностью, чем клиенты, которые пришли по профессиональной рекомендации, но если такие клиенты не ушли, они могут проявить высокую лояльность в следующем периоде предоставления услуг. Есть смысл пересортировать клиентов, чтобы посмотреть с другой стороны на клиентов с высокой вероятностью оттока.

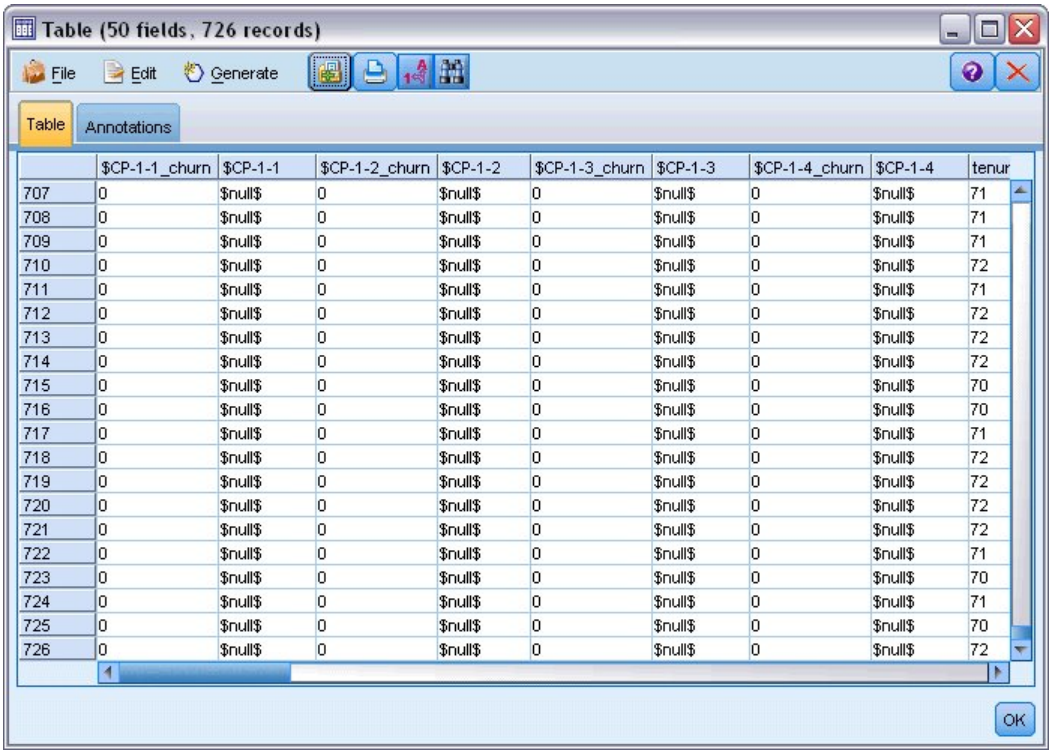


Table (50 fields, 726 records)

File Edit Generate

Table Annotations

	$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	tenur
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

OK

Рисунок 380. Таблица, содержащая клиентов с пустыми значениями

В нижней части расположены клиенты с пустым значением прогноза. Это клиенты, у которых общая длительность предоставления услуг (будущее время + *длительность предоставления услуг*) выпадает за пределы диапазона времени выживания в данных, использованных для обучения модели.

Итог

Используя регрессию Кокса, вы нашли приемлемую модель для времени до оттока, построили график ожидаемого числа оставшихся клиентов на следующие два года и выявили отдельных клиентов, которые с наибольшей вероятностью перейдут к другим поставщикам в следующем году. Имейте в виду, однако, что эта приемлемая модель - не обязательно лучшая. Хорошо бы сравнить эту модель, полученную прямым шаговым методом, с созданной обратным шаговым методом.

Объяснение математических основ методов моделирования, используемых в IBM SPSS Modeler, смотрите в публикации *IBM SPSS Modeler: Руководство по алгоритмам*.

Глава 27. Анализ покупательской корзины (вывод правила/C5.0)

В этом примере рассматриваются фиктивные данные, описывающие содержание покупательских корзин в супермаркете (то есть собрания приобретаемых совместно товаров), а также связанные персональные данные покупателей, которые можно получить, используя информацию карточек покупателей. Цель этого - обнаружение групп покупателей, которые приобретают сходные товары и могут характеризоваться демографическими показателями (возраст, доход и так далее).

Этот пример иллюстрирует две фазы исследования данных:

- Моделирование правил связывания и выявление связей между приобретенными товарами с помощью вывода сетевого графа
- Создание профилей покупателей указанных групп товаров с помощью вывода правила C5.0

Примечание: Эта прикладная программа непосредственно не использует предсказательное моделирование, поэтому в ней нет измерений точности для получающихся моделей и связанного разделения на стадии обучения/проверки в процессе анализа данных.

Этот пример использует поток *baskrule*, в котором используется файл данных *BASKETS1n*. Эти файлы находятся в каталоге *Demos* любой установки IBM SPSS Modeler. Доступ к ним можно получить из группы программ IBM SPSS Modeler в меню Пуск Windows. Файл *baskrule* находится в каталоге *streams*.

Доступ к данным

При помощи узла файла переменных соединитесь с набором данных *BASKETS1n*, выбрав имена полей для чтения из файла. Соедините узел типа с источником данных, а затем соедините этот узел с узлом Таблица. Задайте для уровня измерений поля *cardid* значение *Без типа* (так как каждый ID карточки покупателя встречается в наборе данных только однажды и поэтому не может использоваться в моделировании). Выберите значение *Номинальное* как уровень измерения для поля *пол* (это нужно, чтобы алгоритм априорного моделирования не рассматривал поле *пол* как флаг).

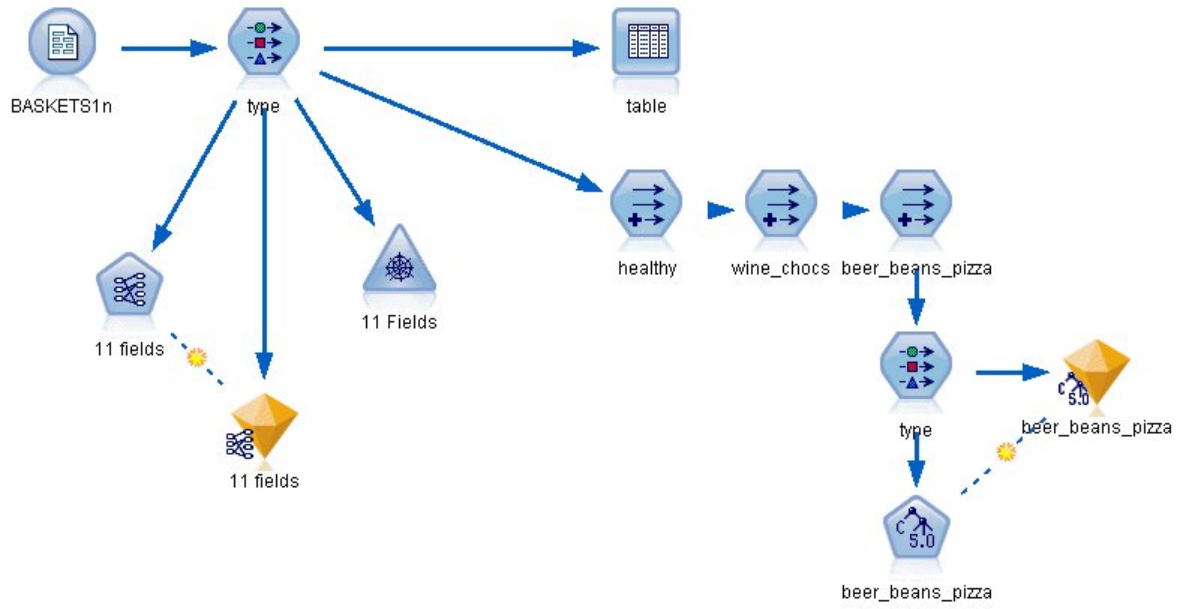


Рисунок 381. Поток baskrule

Теперь запустите поток, чтобы инстанциировать узел Тип и вывести таблицу. Набор данных содержит 18 полей, в которых каждая запись представляет корзину.

Эти 18 полей представлены следующими заголовками.

Сводка корзины:

- *cardid*. Идентификатор карточки покупателя, который приобретает данную корзину.
- *value*. Суммарная цена покупки для корзины.
- *pmethod*. Способ оплаты корзины.

Персональные подробности владельца карточки:

- *sex* (пол)
- *homeown* (домовладелец). Домовладелец ли держатель карточки или нет.
- *income* (доход)
- *age* (возраст)

Флаги содержимого корзины, обозначающие наличие товаров по категориям:

- *fruitveg* (овощи и фрукты)
- *freshmeat* (свежее мясо)
- *dairy* (молочные продукты)
- *cannedveg* (овощные консервы)
- *cannedmeat* (мясные консервы)
- *frozenmeal* (замороженная еда)
- *beer* (пиво)
- *wine* (вино)
- *softdrink* (безалкогольные напитки)
- *fish* (рыба)
- *confectionery* (кондитерские изделия)

Обнаружение аффинитетов в содержимом корзин

Во-первых, вам нужно получить общую картину об аффинитетах (взаимосвязях) в содержимом корзин, используя априорную модель для создания правил связывания. Выберите поля, которые будут использоваться в этом процессе моделирования, изменив узел Тип и задав для роли всех товарных категорий значение *Оба*, а для всех других ролей значение *Нет*. (*Оба* означает, что поле может быть и входным, и выходным полем полученной в результате модели.)

Примечание: Опции для нескольких полей можно задать щелчком мыши при нажатой клавише Shift, чтобы выбрать поля, прежде чем задавать опцию из столбцов.

Рисунок 382. Выбор полей для моделирования

После указания полей для моделирования присоедините узел Априори к узлу Тип, измените его, выберите опцию **Только значения true для флагов** и щелкните по Выполнить в узле Априори. Результат, то есть модель на вкладке Модели в правом верхнем углу окна менеджеров, будет содержать правила связывания, которые можно просмотреть, используя контекстное меню и выбрав опцию **Обзор**.

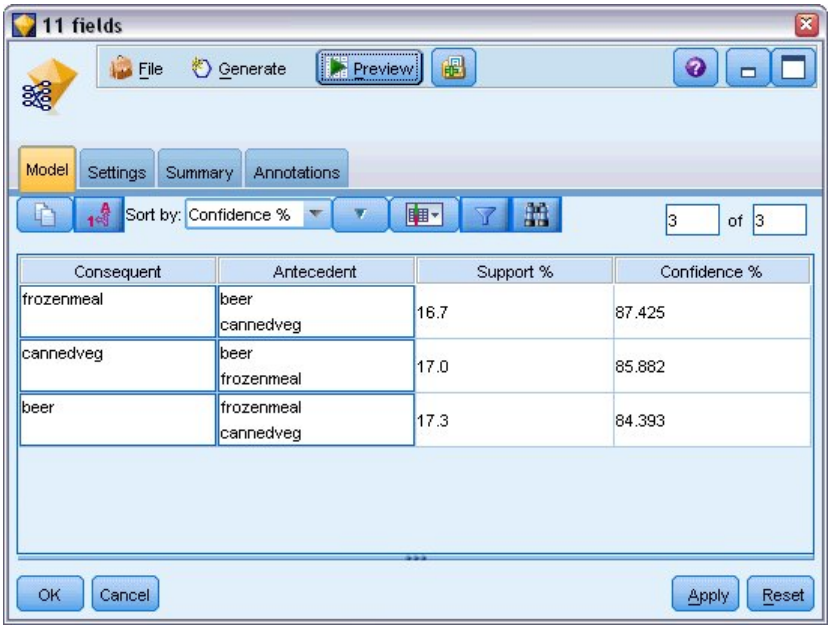


Рисунок 383. Правила связывания

Эти правила показывают разнообразные взаимосвязи между замороженными продуктами, консервированными овощами и пивом. Наличие двусторонних правил связывания, таких как:

```
замороженные продукты -> пиво
пиво -> замороженные продукты
```

показывает, что использование вывода сетевого графа (показывающего только двусторонние взаимосвязи) может выделить некоторые из паттернов в этих данных.

Присоедините узел Веб к узлу Тип, измените узел Веб, выберите все поля содержимого корзины, выберите опцию **Показывать только флаги true** и щелкните по Выполнить в узле Веб.

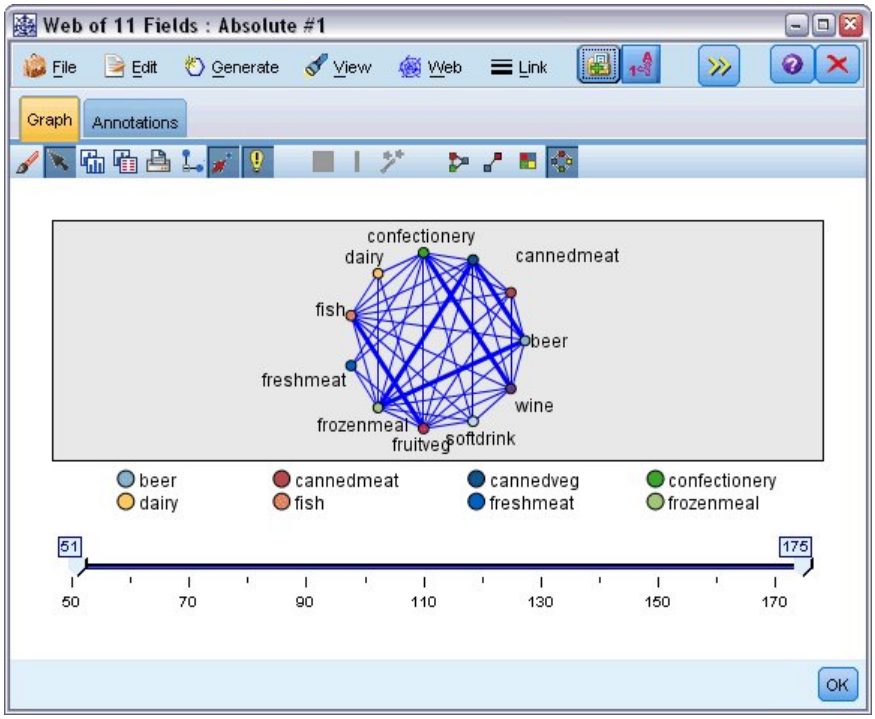


Рисунок 384. Вывод сетевого графа взаимосвязей товаров

Так как большинство комбинаций товарных категорий встречается в нескольких корзинах, сильные связи этого графа слишком многочисленны, чтобы показать предложенные моделью группы покупателей.

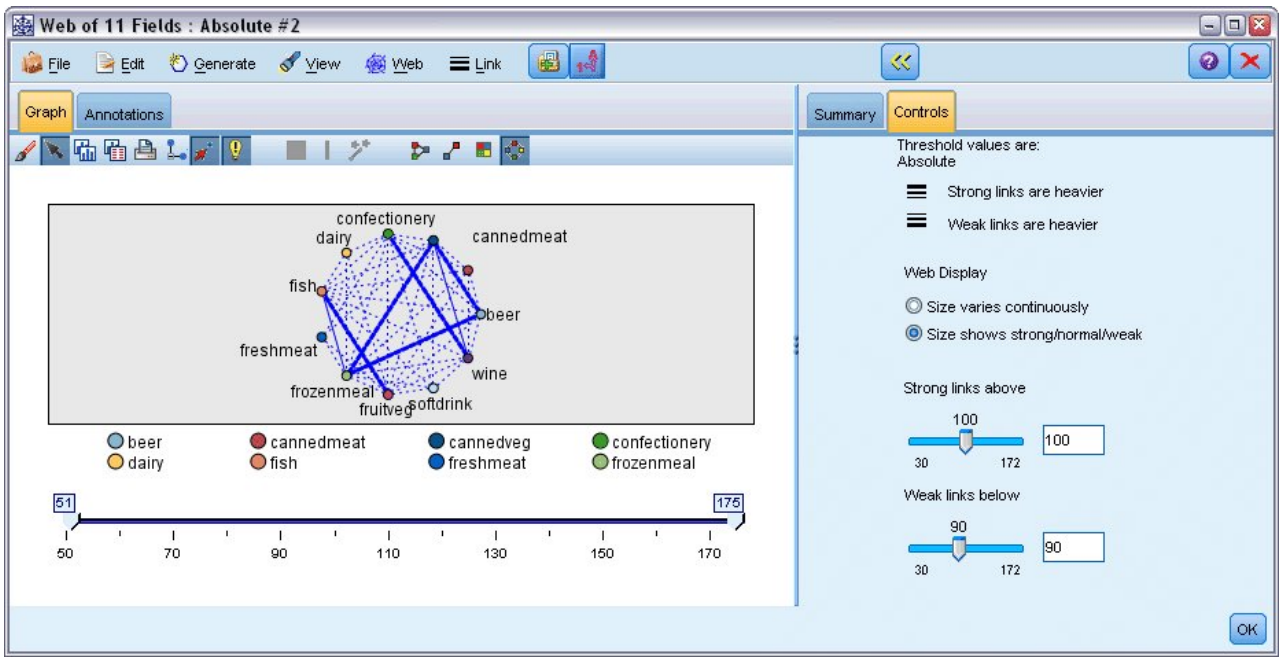


Рисунок 385. Ограниченный вывод сетевого графа

1. Для указания слабых и сильных соединений нажмите кнопку с двойной желтой стрелкой на панели инструментов. После этого раскроется диалоговое окно со сводкой вывода графа и элементами управления.
2. Выберите **Размер отражает сильную/нормальную/слабую**.

3. Задайте для слабых связей значения меньше 90.
4. Задайте для сильных связей значения больше 100.

На полученном выводе будут выделены три группы покупателей:

- Покупатели рыбы, фруктов и овощей; их можно отнести к группе "здорового питания"
- Покупатели вина и кондитерских изделий
- Покупатели пива, замороженной еды и консервированных овощей ("пиво, бобы и пицца")

Создание профилей для групп покупателей

Теперь вы идентифицировали три группы покупателей на основании типов приобретаемых ими товаров, но хотели бы также знать, кто эти покупатели, то есть их демографический профиль. Этого можно достичь, отметив каждого покупателя из этих групп флагом и используя вывод правила (C5.0), чтобы построить профили этих флагов на основании правила.

Сначала необходимо получить флаг для каждой группы. Его можно сгенерировать автоматически, используя уже созданный вывод сетевого графа. Щелкните правой кнопкой мыши по связи между полями *fruitveg* (овощи и фрукты) и *fish* (рыба), чтобы ее выделить, затем еще раз щелкните правой кнопкой мыши и выберите опцию **Сгенерировать узел извлечения для связи**.

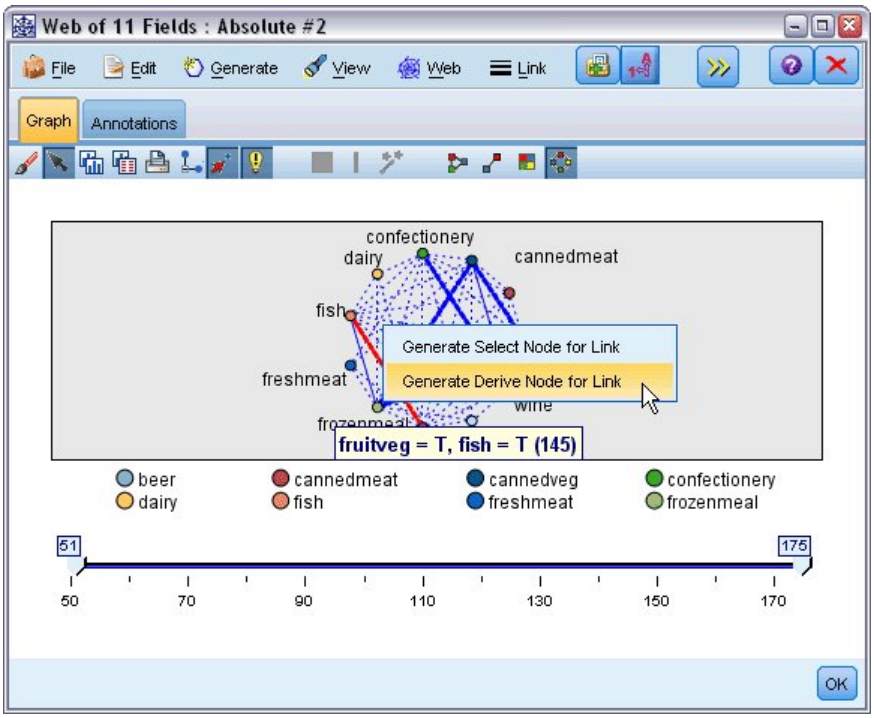


Рисунок 386. Получение флага для каждой группы покупателей

В полученном узле извлечения замените имя поля Извлечь на *healthy* (здоровое питание) Повторите то же самое для связи между полями *wine* (вино) и *confectionery* (кондитерские изделия), назвав получающееся производное поле *вино_шоколад*.

Для третьей группы (в которой есть три связи) сначала убедитесь, что нет выбранных связей. Затем выберите все три связи в треугольнике *cannedveg* (овощные консервы), *beer* (пиво) и *frozenmeal* (замороженная еда), удерживая нажатой клавишу Shift при щелчках левой кнопкой мыши. (Убедитесь, что включен интерактивный режим, а не режим редактирования.) Затем в меню вывода сетевого графа выберите:

Сгенерировать > Узел извлечения ("And")

Измените имя получающегося поля Извлечь на *beer_beans_pizza* (пиво_бобы_пицца).

Чтобы создать профили этих групп покупателей, соедините существующий узел Тип этими тремя узлами извлечения подряд, а затем подсоедините другой узел Тип. В новом узле Тип задайте для ролей всех полей значения *Нет*, кроме полей *value*, *pmethod*, *sex*, *homeown*, *income* и *age* (значение, метод платежа, пол, домовладелец, доход и возраст), которые должны быть указаны как *Входные*, и соответствующих полей групп покупателей (например, *beer_beans_pizza*), для которых должно быть задано *Назначение*. Присоедините узел C5.0, задайте для типа вывода значение **Набор правил** и щелкните по Выполнить в этом узле. Получающаяся модель (для группы *beer_beans_pizza*) будет содержать ясный демографический профиль этой группы покупателей:

```
Rule 1 for T:
if sex = M
and income <= 16,900
then T
```

Тот же способ можно применить к другим флагам групп покупателей, выбрав их для полей вывода во втором узле Тип. Более широкий диапазон альтернативных профилей можно сгенерировать, используя в этом контексте вместо C5.0 априорную модель; априорная модель может использоваться также для профилирования всех флагов групп покупателей одновременно, так как она не ограничена одним выходным полем.

Итог

В этом примере показано, как можно использовать IBM SPSS Modeler для обнаружения аффинитетов, или связей, в базе данных при моделировании (с использованием априорной модели) и визуализации (с использованием вывода сетевого графа). Эти связи соответствуют группировке наблюдений в данных, причем эти группы могут подробно исследоваться и профилироваться при моделировании (с использованием наборов правил C5.0).

В домене розничной торговли такое объединение покупателей в группы может использоваться, например, для определения назначения специальных предложений, чтобы повысить уровень откликов на прямые почтовые рассылки или уточнить диапазон товаров на складе по торговым маркам, чтобы они соответствовали спросу покупателей с соответствующей демографической базой.

Глава 28. Оценка новых предложений транспортных средств (KNN)

Анализ ближайшего сходства представляет собой метод классификации наблюдений на основе сходства наблюдений. Этот метод машинного обучения был разработан в качестве способа распознавания структуры данных при неточном соответствии имеющих структур или наблюдений. Подобные наблюдения близки друг к другу, а непохожие наблюдения, наоборот, удалены друг от друга. Таким образом, дистанция между двумя наблюдениями является критерием их различия.

Близкие друг к другу наблюдения называются “соседи”. Когда представляется новое наблюдение, обозначенное знаком вопроса, вычисляется его расстояние от всех других наблюдений в модели. Определяется классификация наиболее похожих наблюдений (ближайшее сходство) и новое наблюдение помещается в категорию, в которой содержится наибольшее количество ближайшего сходства.

Вы можете указать количество анализируемых ближайших соседей; это значение обозначается *k*. На рисунках ниже показано, каким образом новое наблюдение будет классифицироваться с использованием двух различных значений *k*. Если *k* = 5, новое наблюдение помещается в категорию *1*, поскольку большинство ближайших соседей принадлежит категории *1*. Однако если *k* = 9, новое наблюдение помещается в категорию *0*, поскольку большинство ближайших соседей принадлежит категории *0* .

Анализ ближайшего сходства также может использоваться для вычисления значений для непрерывного целевого объекта. В этой ситуации среднее целевое значение ближайшего сходства используется для получения предсказанного значения для нового наблюдения.

Производитель автомобилей разработал прототипы для двух новых транспортных средств, легкового автомобиля и грузовика. Прежде чем вводить новые модели в свою линейку предложений, производитель хочет определить, какие из существующих моделей на рынке наиболее сходны с этими прототипам, то есть какие машины будут их "ближайшими соседями" и, тем самым, теми моделями, с которыми они будут конкурировать.

Производитель собрал данные о существующих моделях под многими категориями и добавил подробности своих прототипов. В число категорий, по которыми будут сравниваться модели, входят цена в тысячах долларов (*price*), объем двигателя (*engine_s*), мощность (*horsepow*), колесная база (*wheelbas*), ширина (*width*), длина (*длина*), собственный вес (*curb_wgt*), объем топливного бака (*fuel_cap*) и расход топлива (*mpg*).

В этом примере используется поток с именем *car_sales_knn.str*, доступный в папке *Demos* в подпапке *streams*. Файл данных - это *car_sales_knn_mod.sav*. Дополнительную информацию смотрите в разделе “Папка demos” на стр. 4.

Создание потока

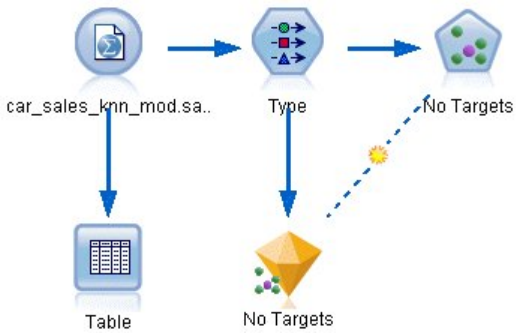
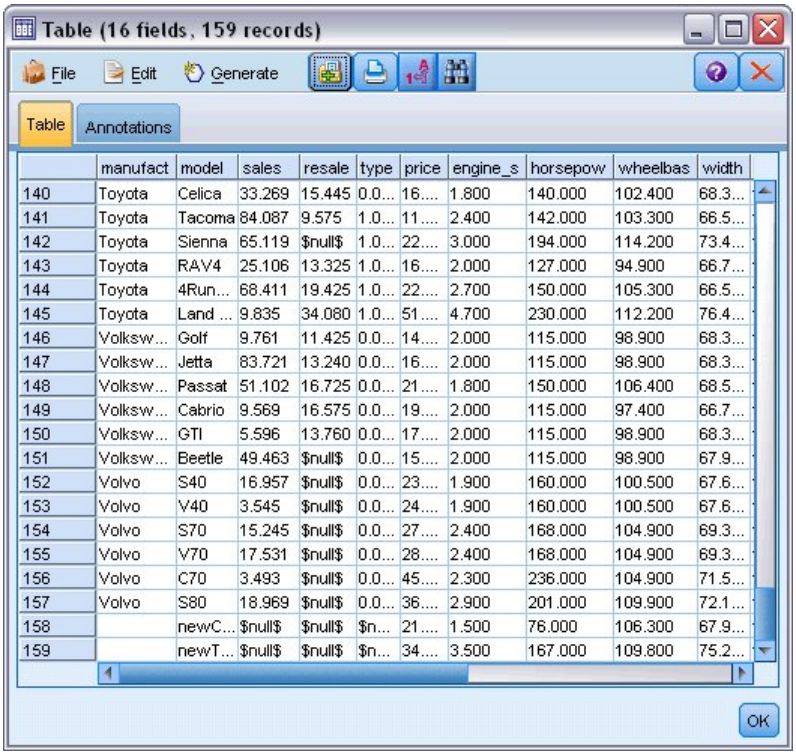


Рисунок 387. Поток примера для моделирования KNN

Создайте новый поток и добавьте исходный узел Файл статистики, указывающий на файл *car_sales_knn_mod.sav* в папке *Demos* вашей установки IBM SPSS Modeler.

Сначала давайте посмотрим, какие данные собрал производитель.

1. Присоедините узел Таблица к исходному узлу Файл статистики.
2. Откройте узел Таблица и щелкните по **Выполнить**.



	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...

Рисунок 388. Исходные данные для легковых и грузовых автомобилей

Подробности для двух прототипов, названных *newCar* и *newTruck*, были добавлены в конец этого файла.

Из этих исходных данных мы видим, что производитель использует классификацию "truck" (значение 1 в столбце *type*) достаточно нестрого для всех транспортных средств, кроме легковых автомобилей.

Последний столбец, *partition*, необходим для того, чтобы эти два прототипа можно было обозначить как исключенные, когда мы перейдем к идентификации их ближайших соседей. Таким образом их

данные не будут влиять на вычисления, так как мы хотим рассмотреть остальную часть рынка. Задание в поле *partition* для двух исключаемых записей значения 1, а для всех остальных записей - значения 0 позволяет использовать это поле позже, когда мы перейдем к указанию фокусных записей - то есть записей, для которых мы хотим вычислить самых близких соседей.

Оставим окно вывода таблицы пока открытым, так как мы будем обращаться к нему позже.

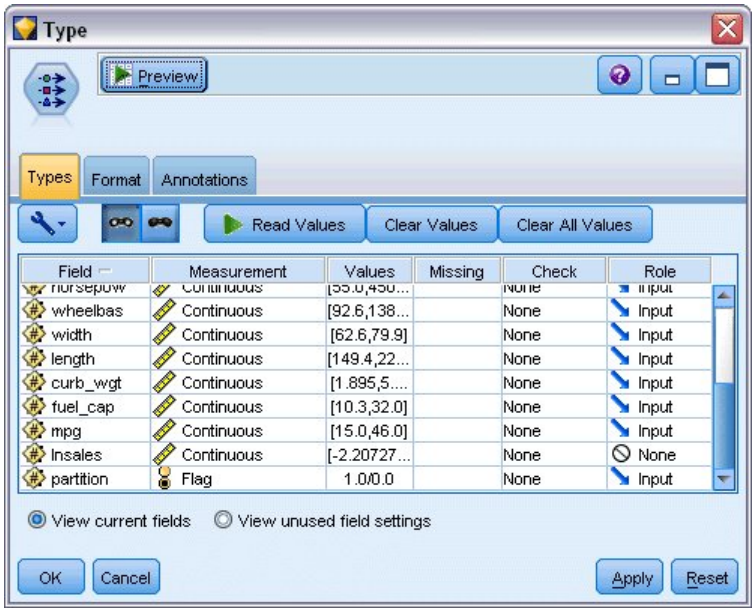


Рисунок 389. Параметры узла Тип

3. Добавьте узел Тип к потоку.
4. Присоедините узел Тип к исходному узлу Файл статистики.
5. Откройте узел Тип.

   Мы хотим выполнить сравнение только по полям от *price* (цена) до *mpg* (расход топлива), так что оставим для всех этих полей в качестве роли **Input** (Входные).
6. Для всех остальных полей (от *manufact* до *type* плюс *lnsales*) зададим роль **None** (Нет).
7. Для последнего поля, *partition*, зададим уровень измерения **Flag** (Флаг). Убедитесь, что в качестве его роли задано **Input** (Входное).
8. Нажмите кнопку **Читать значения**, чтобы прочесть значения данных в поток.
9. Щелкните по **OK**.

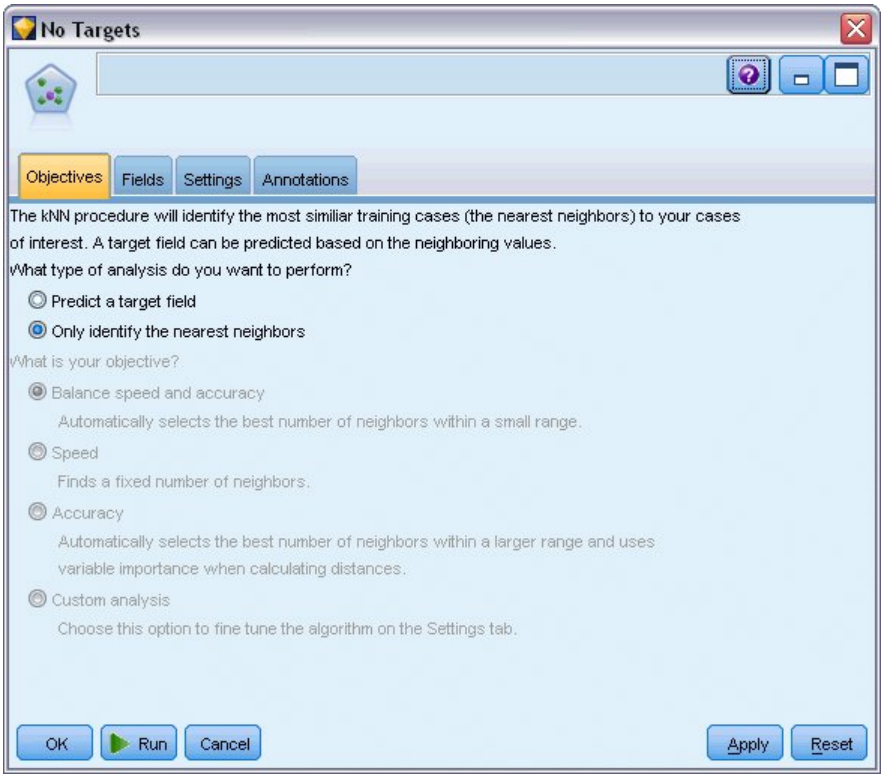


Рисунок 390. Выбор для идентификации ближайших соседей

10. Присоедините узел KNN к узлу Тип.
11. Откройте узел KNN.

    Мы не собираемся предсказывать значение целевого поля в данный момент, мы просто хотим найти ближайших соседей для наших двух прототипов.
12. На вкладке **Цели** выберите **Только определить ближайших соседей**.
13. Щелкните по вкладке **Параметры**.

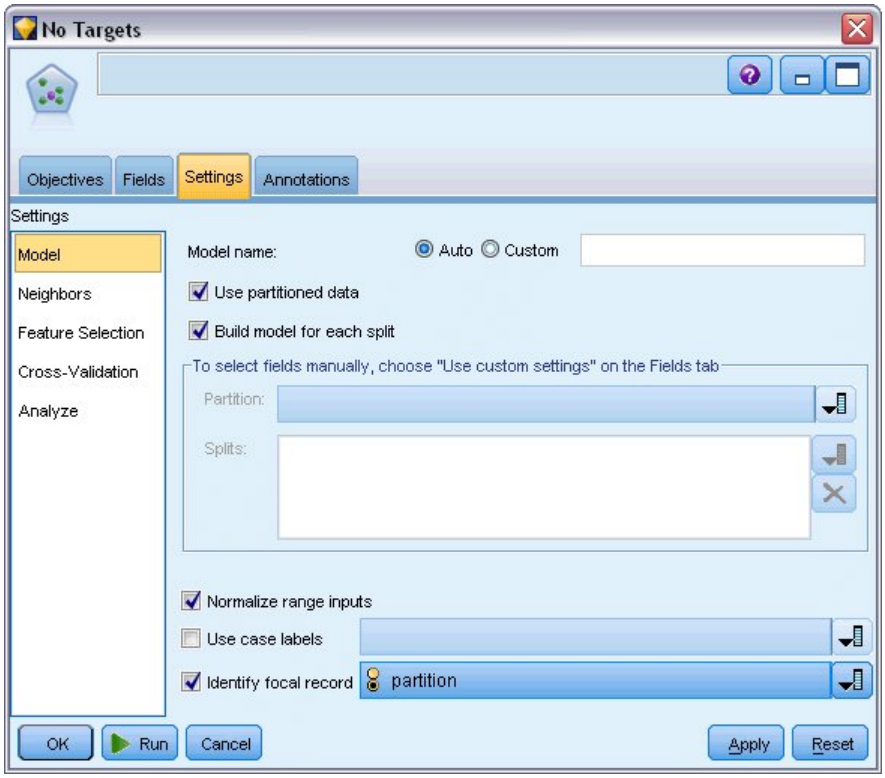


Рисунок 391. Использование поля partition для идентификации фокусных записей

Теперь можно использовать поле *partition* для идентификации фокусных записей - записей, для которых мы хотим определить ближайших соседей. При помощи поля флага мы обеспечиваем, чтобы записи, где значение этого поля - 1, стали нашими фокусными записями.

Как мы видели, единственные записи со значением 1 в этом поле - *newCar* и *newTruck*, таким образом, они будут нашими фокусными записями.

14. На панели **Модель** вкладки **Параметры** включите переключатель **Определить фокусную запись**.
15. Из выпадающего списка для этого поля выберите **partition**.
16. Нажмите кнопку **Выполнить**.

Изучение вывода

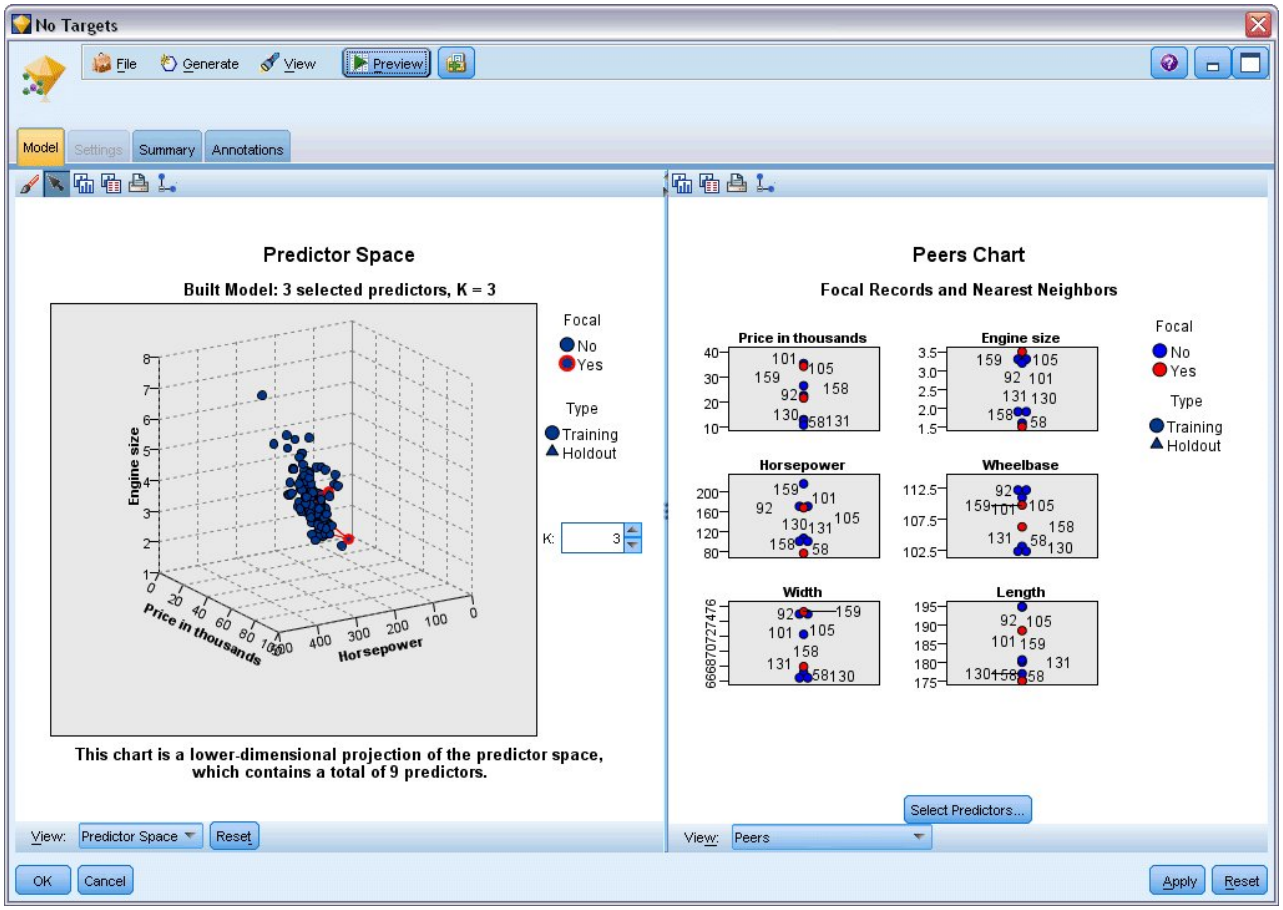


Рисунок 392. Окно Просмотр модели

Слепок модели создается на холсте потока и на палитре Модели. Откройте любой из этих слепков, чтобы открыть окно Просмотр модели, содержащее две панели:

- На первой панели выводится обзорное изображение модели, называемое главным представлением. Главное представление для модели Ближайший сосед называют **Пространством предикторов**.
- Вторая панель выводит изображение одного из двух типов:

  Дополнительное представление модели показывает дополнительную информацию о модели, но не концентрируется на самой модели.

  Связанное представление - это представление, в котором выводятся подробности одной из характеристик модели, когда вы выполняете детализацию часть главного представления.

Пространство предикторов

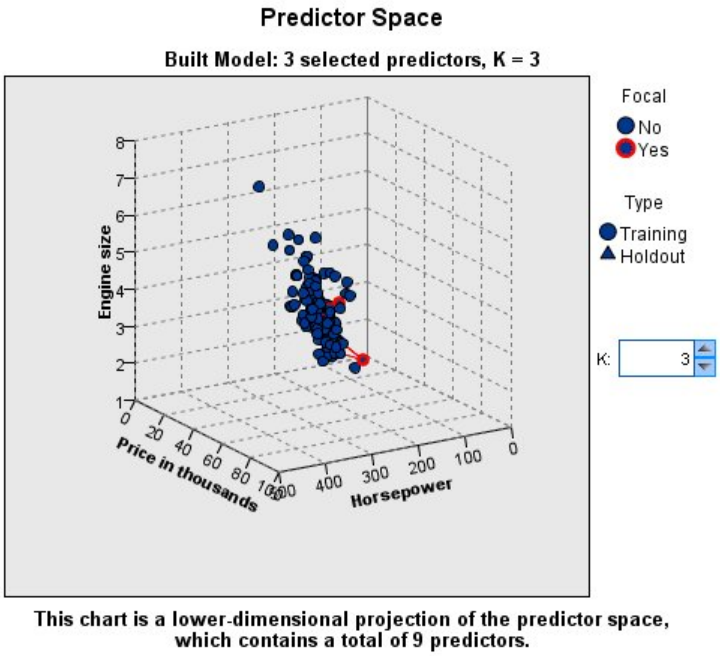


Рисунок 393. Диаграмма пространства предикторов

Диаграмма пространства предикторов - это интерактивная трехмерная диаграмма, на которой размещены точки данных для трех характеристик (фактически для трех первых входных полей в исходных данных), представляющих цену, объем двигателя и мощность.

Наши две фокусных записи выделены красным цветом, линии соединяют их с *k* ближайшими соседями.

Перетаскивая эту диаграмму, можно поворачивать ее, чтобы лучше видеть распределение точек в пространстве предикторов. Нажмите кнопку **Сброс**, чтобы вернуться к представлению по умолчанию.

Диаграмма сходства

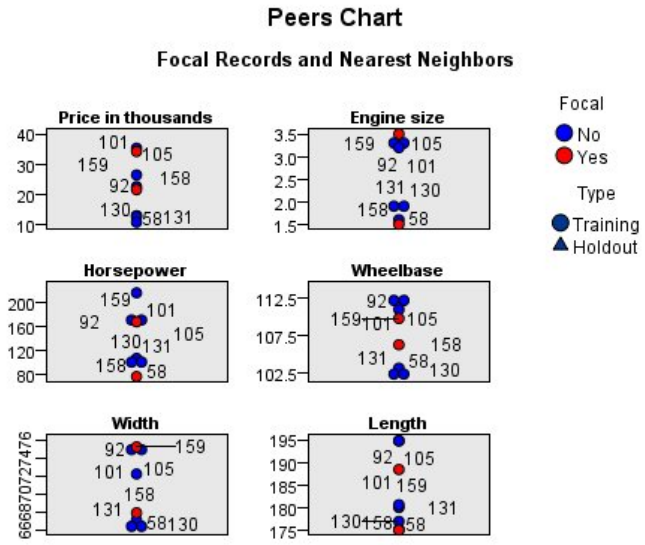


Рисунок 394. Диаграмма сходства

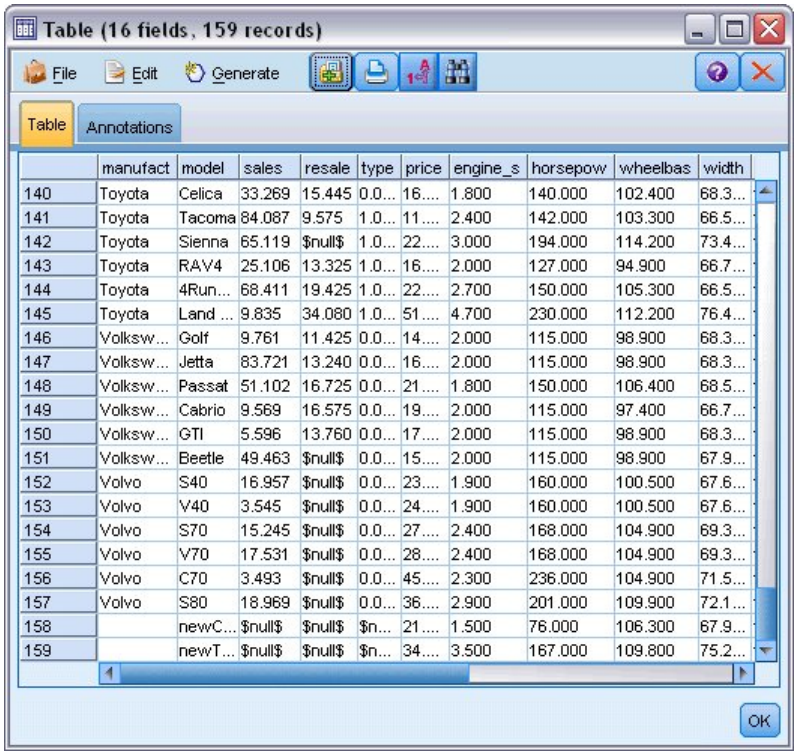
Дополнительное представление по умолчанию - это диаграмма сходства, на которой выделены две фокусные записи, выбранные в пространстве предикторов, и их *k* ближайших соседей по каждой из шести характеристик - первые шесть входных полей в исходных данных.

Транспортные средства представлены их номерами записей в исходных данных. Здесь нам потребуется вывод из узла Таблицы, который поможет идентифицировать их.

Если вывод узла Таблицы все еще доступен:

1. Щелкните по вкладке **Вывод** панели менеджера в верхней правой части главного окна IBM SPSS Modeler.
2. Дважды щелкните по записи **Таблица (16 полей, 159 отчетов)**.

   Если вывод таблицы более не доступен:
3. В главном окне IBM SPSS Modeler откройте узел Таблица.
4. Нажмите кнопку **Выполнить**.



	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...

Рисунок 395. Идентификация записей по номеру записи

Прокрутив таблицу до конца, мы можем видеть *newCar* и *newTruck* - последние две записи в данных с номерами соответственно 158 и 159.

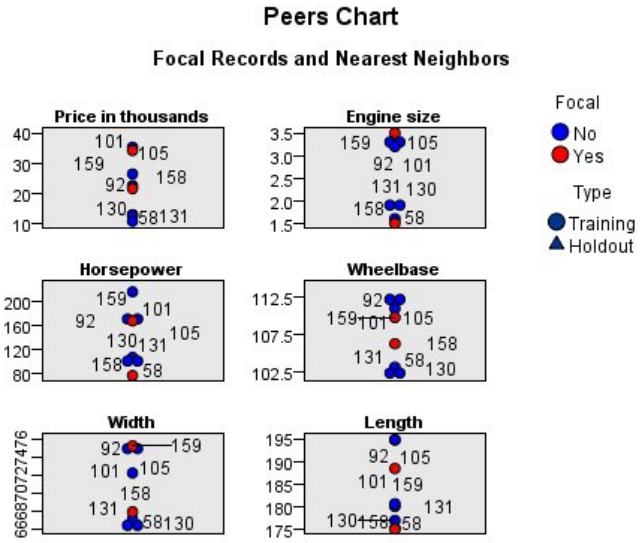


Рисунок 396. Сравнение характеристик на диаграмме сходства

Теперь мы можем посмотреть на диаграмме сходства, например, что *newTruck* (159) превосходит по объему двигателя всех своих ближайших соседей, в то время как *newCar* (158) меньше по объему двигателя *своих* ближайших соседей.

Для каждой из этих шести характеристик можно поместить указатель мыши на отдельные точки, чтобы посмотреть фактическое значения каждой характеристики для этой конкретной точки.

Но какие же транспортные средства будут ближайшими соседями для *newCar* и *newTruck*?

Диаграмма сходства несколько перегружена, так что давайте перейдем к более простому представлению.

5. Щелкните по выпадающему списку **Представление** в нижней части диаграммы сходства (сейчас там указано **Соседи**).
6. Выберите **Таблица соседей и расстояний**.

Таблица соседей и расстояний

k Nearest Neighbors and Distances

Displayed for Initial Focal Records

Focal Record	Nearest Neighbors			Nearest Distar	
	1	2	3	1	2
158	131	130	58	0.979	0.990
159	105	92	101	0.580	0.634

Рисунок 397. Таблица соседей и расстояний

В этой таблице проще разобраться. Теперь мы видим три модели на рынке, к которым каждый из наших двух прототипов ближе всего.

Для *newCar* (фокусная запись 158) это Saturn SC (131), Saturn SL (130) и Honda Civic (58).

Это вполне ожидаемо - все это седаны средних размеров, так что *newCar* должен хорошо соответствовать им, в особенности благодаря отличным показателям расхода топлива.

Для *newTruck* (фокусная запись 159) ближайшие соседи - Nissan Quest (105), Mercury Villager (92) и Mercedes M-Class (101).

Как мы видели ранее, это не обязательно грузовики в традиционном смысле, но просто автомобили, которые не классифицируются как легковые. Глядя на ближайших соседей в выводе узла Таблица, можно увидеть, что *newTruck* относительно дорог, а также будет самым тяжелым в своем типе. Однако расход топлива снова лучше, чем у его ближайших соседей, так что это следует принять во внимание.

Итог

Мы посмотрели, как можно использовать анализ ближайших соседей для сравнения широкого набора характеристик для конкретного набора данных. Мы также вычислили для двух совершенно разных фокусных записей записи, сильнее всего похожие на эти заданные.

Глава 29. Выявление причинных взаимосвязей в бизнес-показателях (TCM)

Бизнес-предприятие отслеживает разнообразные ключевые показатели эффективности, описывающие финансовое положение бизнеса во времени, а также оно отслеживает число показателей, которыми может управлять. Оно заинтересована в использовании причинными моделями времени для выявления причинных взаимосвязей между управляемыми показателями и ключевыми показателями эффективности. Ему также хотелось бы знать обо всех причинных взаимосвязях ключевых показателей эффективности.

Файл данных `tcm_kpi.sav` содержит недельные данные для ключевых показателей эффективности и управляемых показателей. Данные для ключевых показателей эффективности хранятся в полях с префиксом *KPI* (КПЭ). Данные для управляемых показателей хранятся в полях с префиксом *Lever*.

Создание потока

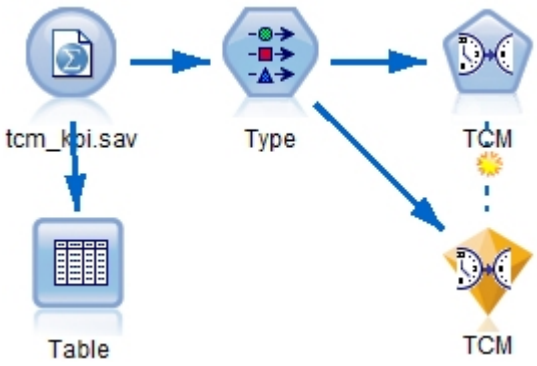
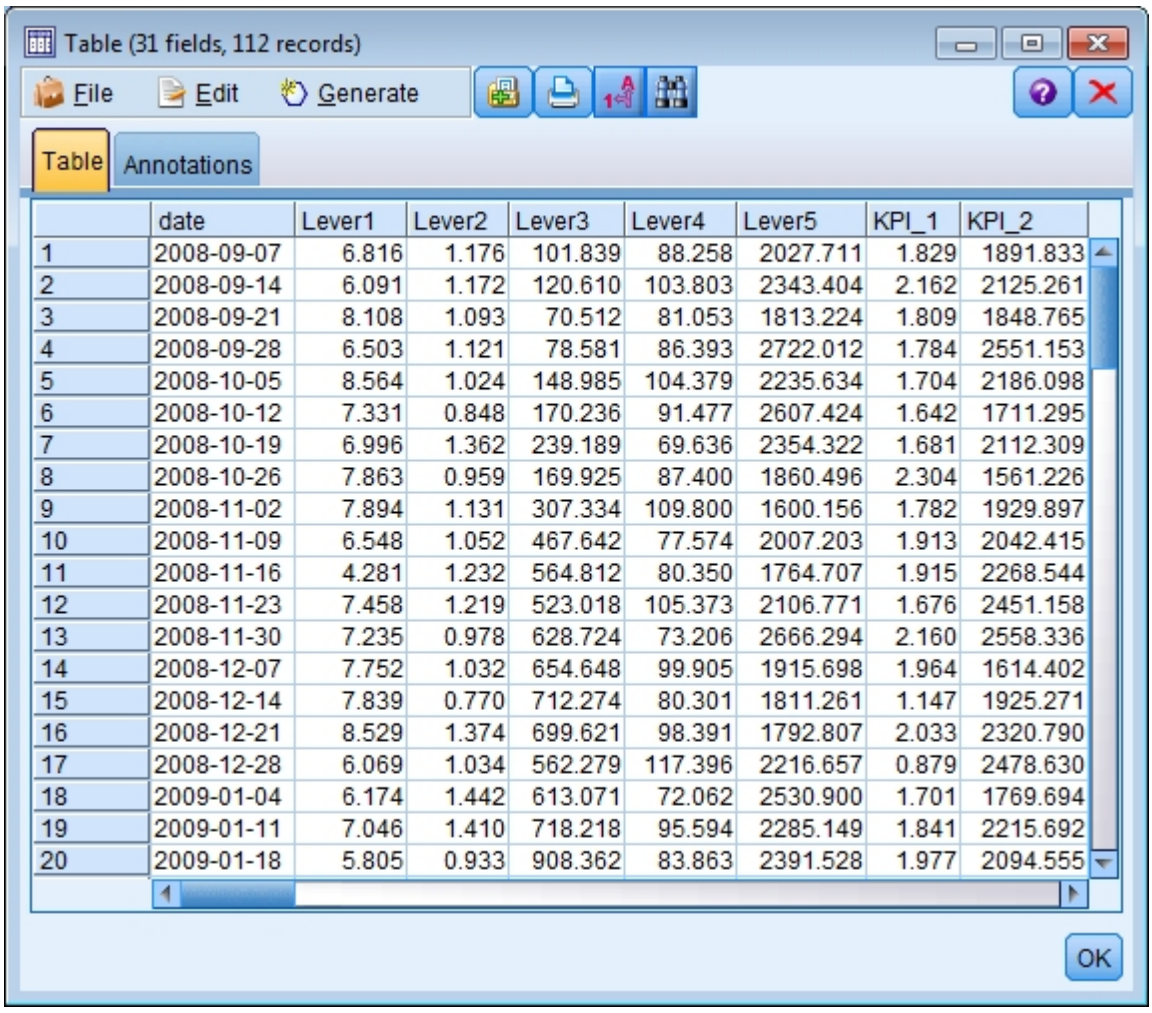


Рисунок 398. Поток примера для моделирования TCM

1. Создайте новый поток и добавьте узел источников файла статистики, указывающий на файл *tcm_kpi.sav* в папке *Demos* вашей установки IBM SPSS Modeler.
2. Присоедините узел Таблица к исходному узлу Файл статистики.
3. Откройте узел Таблица и нажмите кнопку **Выполнить**, чтобы просмотреть данные. Он содержит недельные данные для ключевых показателей эффективности и управляемых показателей. Данные для ключевых показателей эффективности хранятся в полях с префиксом *KPI* (КПЭ), а данные для управляемых показателей - в полях с префиксом *Lever*.



	date	Lever1	Lever2	Lever3	Lever4	Lever5	KPI_1	KPI_2
1	2008-09-07	6.816	1.176	101.839	88.258	2027.711	1.829	1891.833
2	2008-09-14	6.091	1.172	120.610	103.803	2343.404	2.162	2125.261
3	2008-09-21	8.108	1.093	70.512	81.053	1813.224	1.809	1848.765
4	2008-09-28	6.503	1.121	78.581	86.393	2722.012	1.784	2551.153
5	2008-10-05	8.564	1.024	148.985	104.379	2235.634	1.704	2186.098
6	2008-10-12	7.331	0.848	170.236	91.477	2607.424	1.642	1711.295
7	2008-10-19	6.996	1.362	239.189	69.636	2354.322	1.681	2112.309
8	2008-10-26	7.863	0.959	169.925	87.400	1860.496	2.304	1561.226
9	2008-11-02	7.894	1.131	307.334	109.800	1600.156	1.782	1929.897
10	2008-11-09	6.548	1.052	467.642	77.574	2007.203	1.913	2042.415
11	2008-11-16	4.281	1.232	564.812	80.350	1764.707	1.915	2268.544
12	2008-11-23	7.458	1.219	523.018	105.373	2106.771	1.676	2451.158
13	2008-11-30	7.235	0.978	628.724	73.206	2666.294	2.160	2558.336
14	2008-12-07	7.752	1.032	654.648	99.905	1915.698	1.964	1614.402
15	2008-12-14	7.839	0.770	712.274	80.301	1811.261	1.147	1925.271
16	2008-12-21	8.529	1.374	699.621	98.391	1792.807	2.033	2320.790
17	2008-12-28	6.069	1.034	562.279	117.396	2216.657	0.879	2478.630
18	2009-01-04	6.174	1.442	613.071	72.062	2530.900	1.701	1769.694
19	2009-01-11	7.046	1.410	718.218	95.594	2285.149	1.841	2215.692
20	2009-01-18	5.805	0.933	908.362	83.863	2391.528	1.977	2094.555

Рисунок 399. Исходные данные для ключевых показателей эффективности и управляемых показателей

4. Добавьте узел Тип к потоку.
5. Присоедините узел Тип к исходному узлу Файл статистики.

Выполнение анализа

1. Присоединить узел TCM (Temporal Causal Modeling - причинные модели времени) к узлу Тип, затем откройте узел TCM и перейдите к разделу **Наблюдения** вкладки **Поля**.

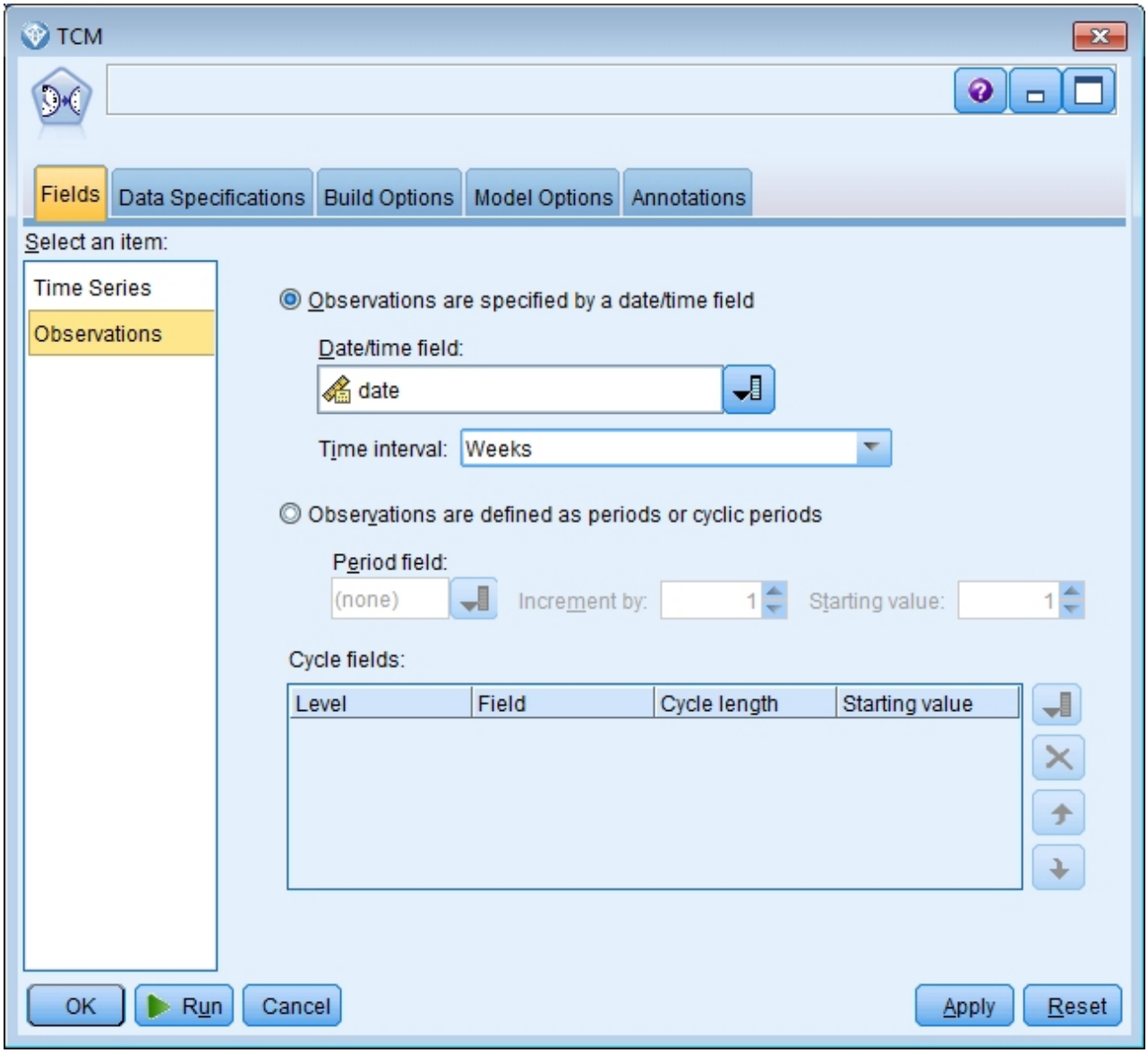


Рисунок 400. Причинные модели времени, наблюдения

2. В поле даты/времени выберите *date*(дата) и в поле временных интервалов выберите *Недели*.
3. Нажмите кнопку **Временные ряды** и выберите **Использовать предопределенные роли**.

   В наборе данных примеров `tcm_kpi.sav` у полей с *Lever1* по *Lever5* роль Входные, а у полей с *KPI_1* по *KPI_25* роль И то, и другое. Если выбрана опция **Использовать предопределенные роли**, поля с ролью Входные обрабатываются и как входные ряды-кандидаты, и как назначения для причинных моделей времени.

   Процедура создания временной причинной модели определяет лучшие входные данные для каждого назначения в наборе входных рядов-кандидатов. В этом примере входные ряды-кандидаты - это поля с *Lever1* по *Lever5* и поля с *KPI_1* по *KPI_25*.
4. Нажмите кнопку **Выполнить**.

Диаграмма Общее качество модели

Диаграмма Общее качество модели выводит элемент вывода, генерируемый по умолчанию, выводит столбчатую диаграмму и связанный точечный график подгонки модели для всех моделей. У каждого ряда назначения есть отдельная модель. Подгонка модели измеряется выбранной статистикой подгонки. В приведенном примере используется статистика подгонки по умолчанию, а именно, R-квадрат.

Элемент Общее качество модели содержит интерактивные возможности. Для включения этих возможностей активируйте указанный элемент, дважды щелкнув мышью по диаграмме Общее качество модели в средстве просмотра.

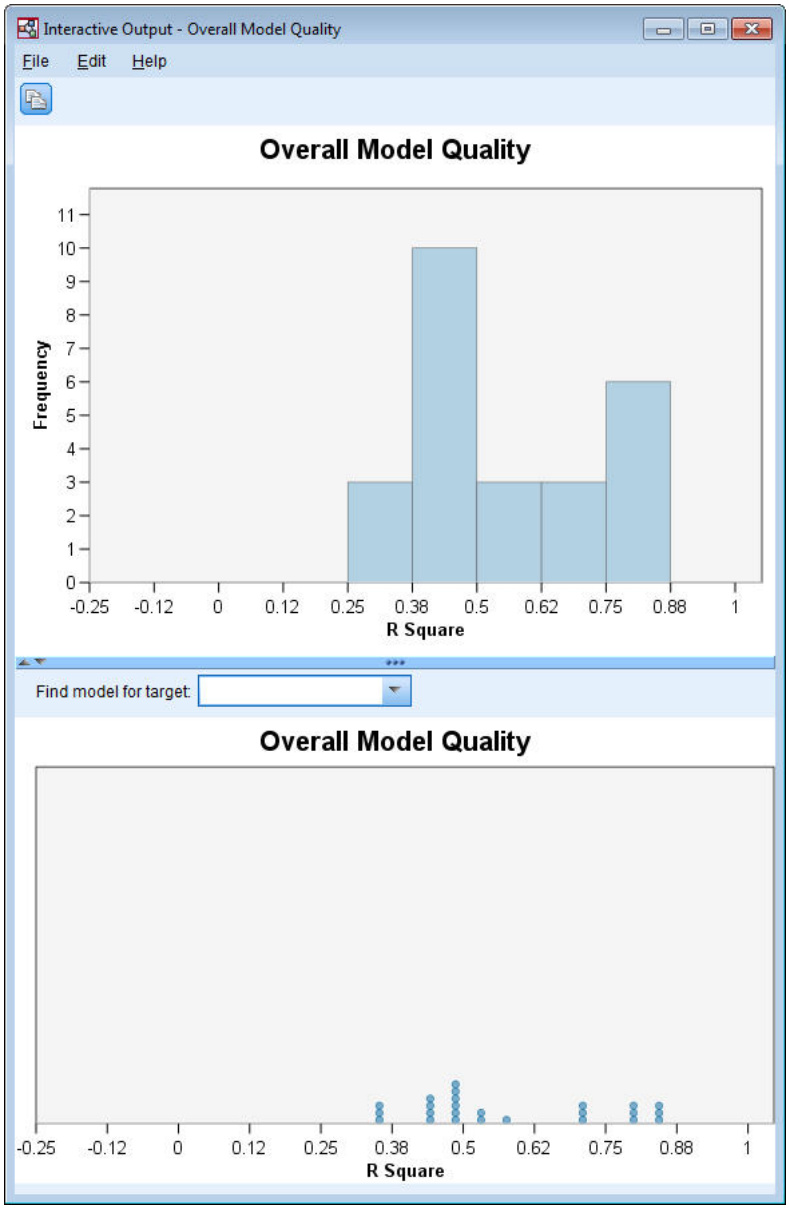


Рисунок 401. Общее качество модели

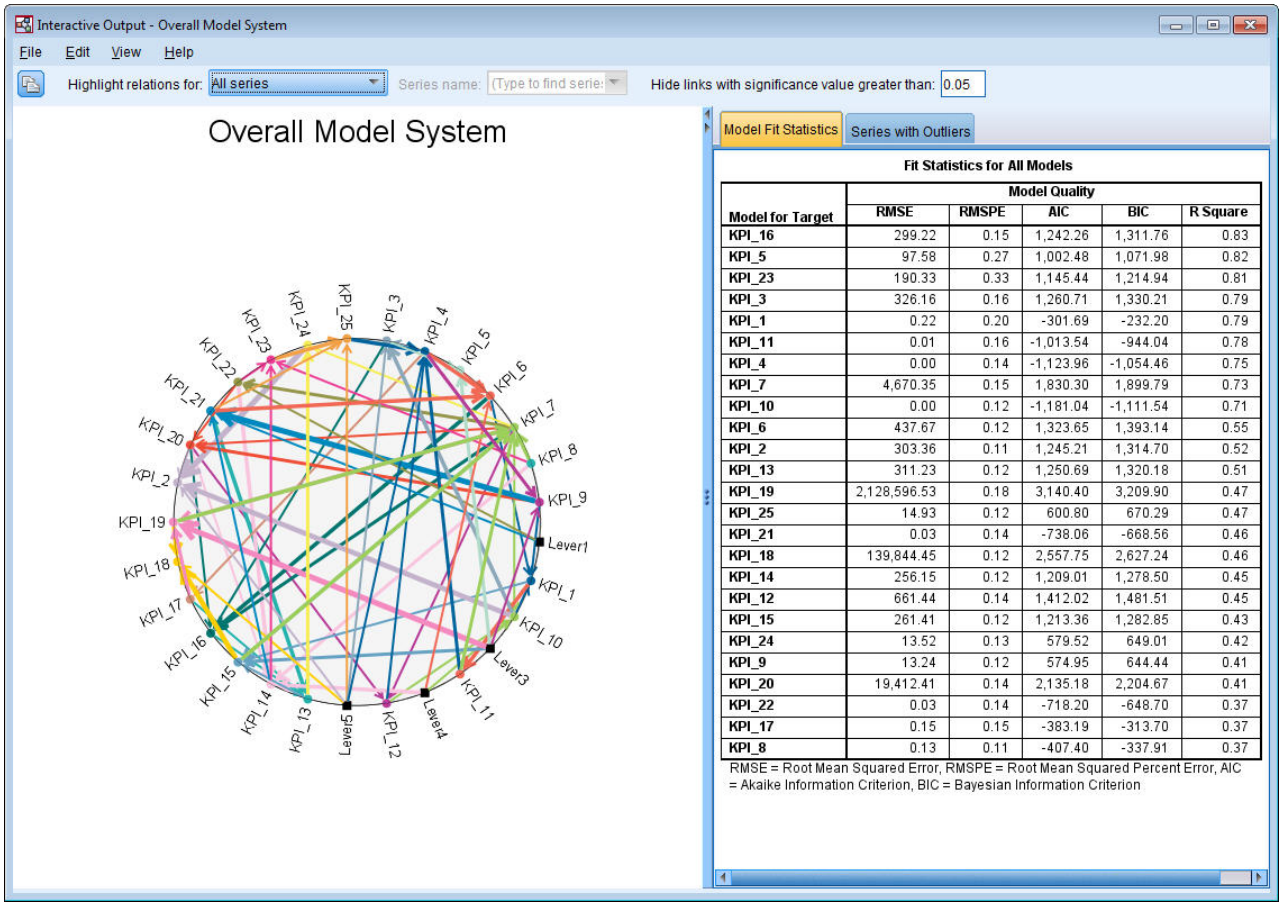
При щелчке по столбцу в столбчатой диаграмме отфильтровывается точечный график, который выводит только модели, связанные с выбранным столбцом. При наведении указателя мыши на точку на точечном

графике всплывает подсказка, содержащая имя связанного ряда и значение статистики подгонки. Модель для конкретного ряда назначения можно найти на точечном графике, указав имя ряда в поле **Найти модель для назначения** .

Общая система модели

Элемент вывода Общая система модели выводит графическое представление причинных взаимосвязей между рядами в системе моделей. По умолчанию выводятся взаимосвязи для 10 лучших моделей, определяемых значением статистики подгонки R-квадрат. Число лучших моделей (называемых также моделями лучшей подгонки) и статистика подгонки задаются в параметрах Ряды для вывода (на вкладке Опции построения) диалогового окна Причинные модели времени.

Элемент Общая система моделей содержит интерактивные возможности. Для включения этих возможностей активируйте указанный элемент, дважды щёлкнув мышью по диаграмме Общая система моделей в средстве просмотра. В этом примере важнее всего увидеть взаимосвязи между всеми рядами в системе. В интерактивном выводе выберите **Все ряды** в выпадающем списке **Выделить взаимосвязи для ряда**.



Model for Target	Model Quality				
	RMSE	RMSPE	AIC	BIC	R Square
KPI_16	299.22	0.15	1,242.26	1,311.76	0.83
KPI_5	97.58	0.27	1,002.48	1,071.98	0.82
KPI_23	190.33	0.33	1,145.44	1,214.94	0.81
KPI_3	326.16	0.16	1,260.71	1,330.21	0.79
KPI_1	0.22	0.20	-301.69	-232.20	0.79
KPI_11	0.01	0.16	-1,013.54	-944.04	0.78
KPI_4	0.00	0.14	-1,123.96	-1,054.46	0.75
KPI_7	4,670.35	0.15	1,830.30	1,899.79	0.73
KPI_10	0.00	0.12	-1,181.04	-1,111.54	0.71
KPI_6	437.67	0.12	1,323.65	1,393.14	0.55
KPI_2	303.36	0.11	1,245.21	1,314.70	0.52
KPI_13	311.23	0.12	1,250.69	1,320.18	0.51
KPI_19	2,128,596.53	0.18	3,140.40	3,209.90	0.47
KPI_25	14.93	0.12	600.80	670.29	0.47
KPI_21	0.03	0.14	-738.06	-668.56	0.46
KPI_18	139,844.45	0.12	2,557.75	2,627.24	0.46
KPI_14	256.15	0.12	1,209.01	1,278.50	0.45
KPI_12	661.44	0.14	1,412.02	1,481.51	0.45
KPI_15	261.41	0.12	1,213.36	1,282.85	0.43
KPI_24	13.52	0.13	579.52	649.01	0.42
KPI_9	13.24	0.12	574.95	644.44	0.41
KPI_20	19,412.41	0.14	2,135.18	2,204.67	0.41
KPI_22	0.03	0.14	-718.20	-648.70	0.37
KPI_17	0.15	0.15	-383.19	-313.70	0.37
KPI_8	0.13	0.11	-407.40	-337.91	0.37

Рисунок 402. Общая система моделей, представление для всех рядов

Все линии, соединяющие отдельное назначение с его входными элементами, одного цвета, а стрелка на каждой линии направлена от входного элемента к назначению этого входного элемента. Например, *Lever3* - это элемент ввода в *KPI_19*.

Толщина каждой линии обозначает важность причинного отношения, причем более толстая линия соответствует более важному отношению. По умолчанию причинные взаимосвязи с уровнем значимости больше 0,05 скрыты. На уровне 0,05 только у *Lever1*, *Lever3*, *Lever4* и *Lever5* будут значимые причинные

взаимосвязи с полями ключевых показателей эффективности. Этот порог уровня значимости можно изменить, введя значение в поле с меткой **Скрыть связи с уровнем значимости больше чем**.

Помимо выявления причинных взаимосвязей между полями *Lever* и полями ключевых показателей эффективности анализ выявляет также взаимосвязи полей ключевых показателей эффективности. Например, элемент *KPI_10* был выбран в качестве элемента ввода в модель для *KPI_2*.

К представлению можно применить фильтр, чтобы посмотреть взаимосвязи только для одного ряда. Например, чтобы просмотреть взаимосвязи только для ряда *KPI_19*, щелкните по метке для *KPI_19*, щелкните правой кнопкой мыши и выберите **Выделить взаимосвязи для ряда**.

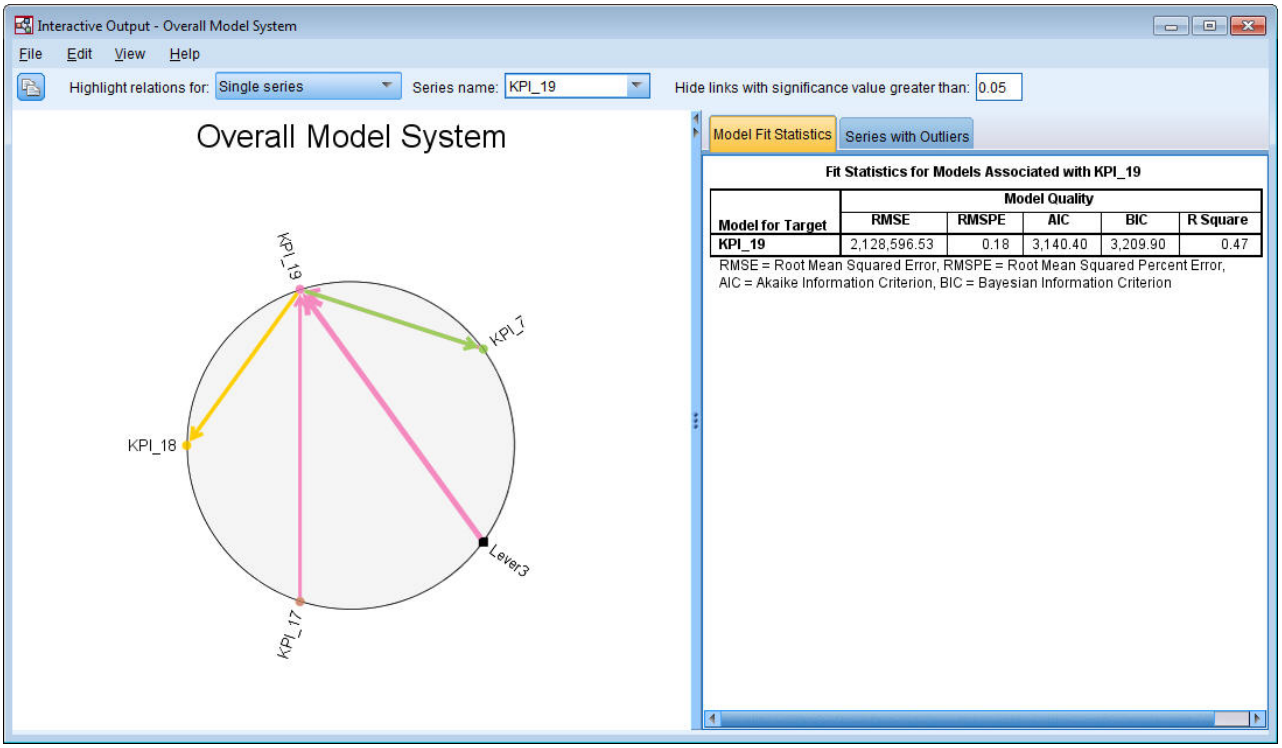


Рисунок 403. Общая система моделей, представление для одного ряда

Это представление показывает элементы ввода в *KPI_19*, уровень значимости которого меньше или равен 0,05. Оно также показывает, что при уровне значимости 0,05 элемент *KPI_19* был выбран в качестве элемента ввода и в *KPI_18*, и в *KPI_7*.

Помимо того, что на экран выводятся взаимосвязей для выбранного ряда, указанный элемент вывода содержит также информацию обо всех обнаруженных для ряда выбросах. Щелкните по вкладке **Ряды с выбросами**.

Series with Outliers for KPI_19

Series	Time	Observed Value
KPI_19	2008-10-12	7,358,201.68
	2009-04-05	2.10E+007
	2010-09-19	6,492,157.97

Рисунок 404. Выбросы для KPI_19

Для *KPI_19* было обнаружено три выброса. Принимая во внимание систему моделей, содержащую все обнаруженные соединения, можно пойти дальше обнаружения выбросов и определить ряд, с наибольшей вероятностью вызывающий конкретный выброс. Анализ этого типа называется анализом первопричин выбросов и рассматривается в этом исследовании случаев в теме ниже.

Диаграммы воздействия

Можно получить полное представление обо всех зависимостях, связанных с конкретным рядом, сгенерировав диаграмму воздействия. Щелкните по метке для поля *KPI_19* на диаграмме Общая система моделей, щелкните правой кнопкой мыши и выберите **Создать диаграмму воздействия**.

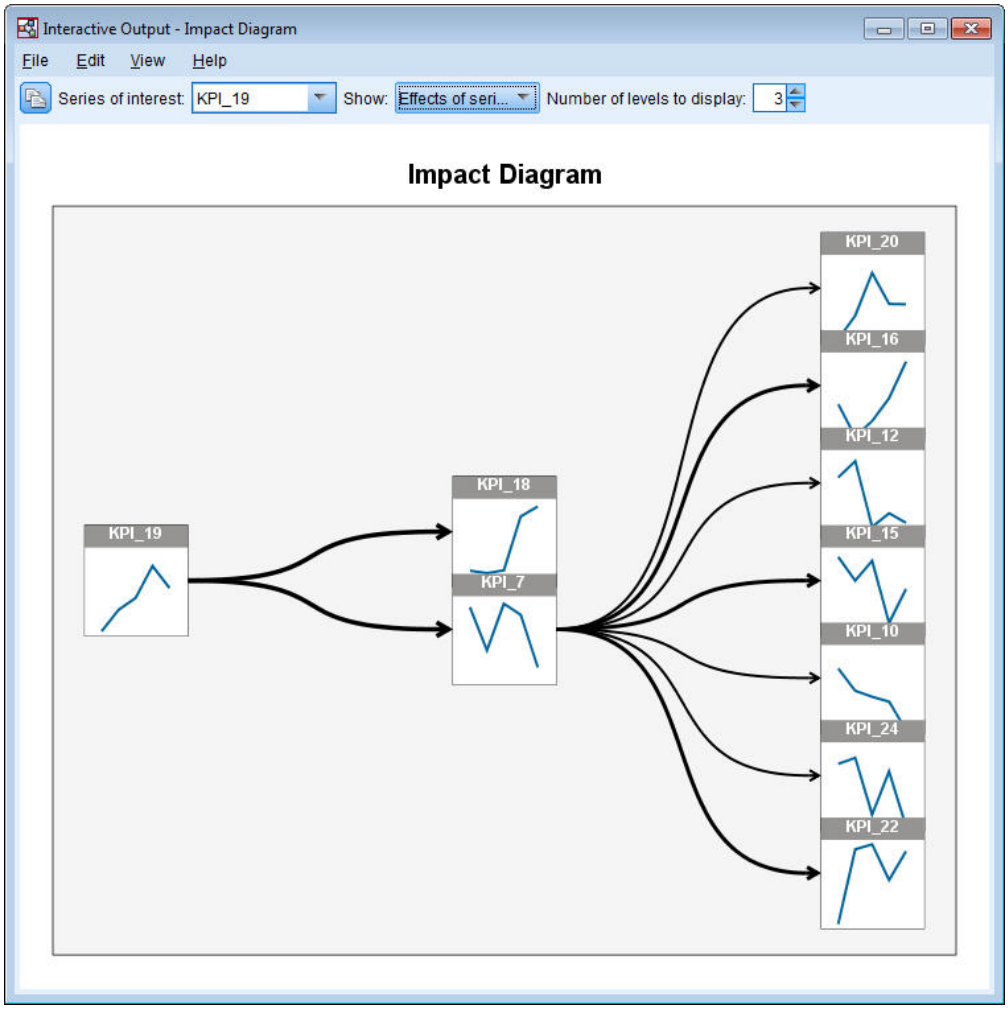


Рисунок 405. Диаграмма воздействия эффектов

После создания диаграммы воздействия эффектов из общей системы моделей (как в этом примере) первоначально она показывает ряды, на которые воздействует выбранный ряд. По умолчанию диаграммы воздействия показывают три уровня эффектов, где первый уровень - это сам рассматриваемый ряд. Каждый дополнительный уровень показывает более неявные эффекты рассматриваемого ряда. Значение **Число выводимых уровней** можно изменить, чтобы выводилось больше или меньше уровней эффектов. Диаграмма воздействия для этого примера показывает, что *KPI_19* - непосредственное поле ввода и в *KPI_18*, и в *KPI_7*, но косвенно оно влияет на несколько рядов посредством воздействия его эффекта на ряд *KPI_7*. Как и в общей системе моделей, толщина линий означает значимость причинных зависимостей.

Диаграмма, выводимая на каждом узле диаграммы воздействия, показывает последние L+1 значений связанного ряда в конце периода оценки и все значения прогноза, где L - это количество шагов задержки, включаемых в каждую модель. Щелкнув один раз по связанному узлу, можно получить подробную диаграмму последовательности.

Двойным щелчком по узлу задается связанный ряд в качестве исследуемого ряда и на основе этого ряда генерируется диаграмма воздействия. В поле **Исследуемый ряд** можно также задать имя ряда, чтобы выбрать другой исследуемый ряд.

Диаграммы воздействия могут также показывать ряды, влияющие на исследуемый ряд. Эти ряды называются *причинами*. Чтобы посмотреть ряды, влияющие на *KPI_19*, в выпадающем списке **Показать** выберите **Причины ряда**.

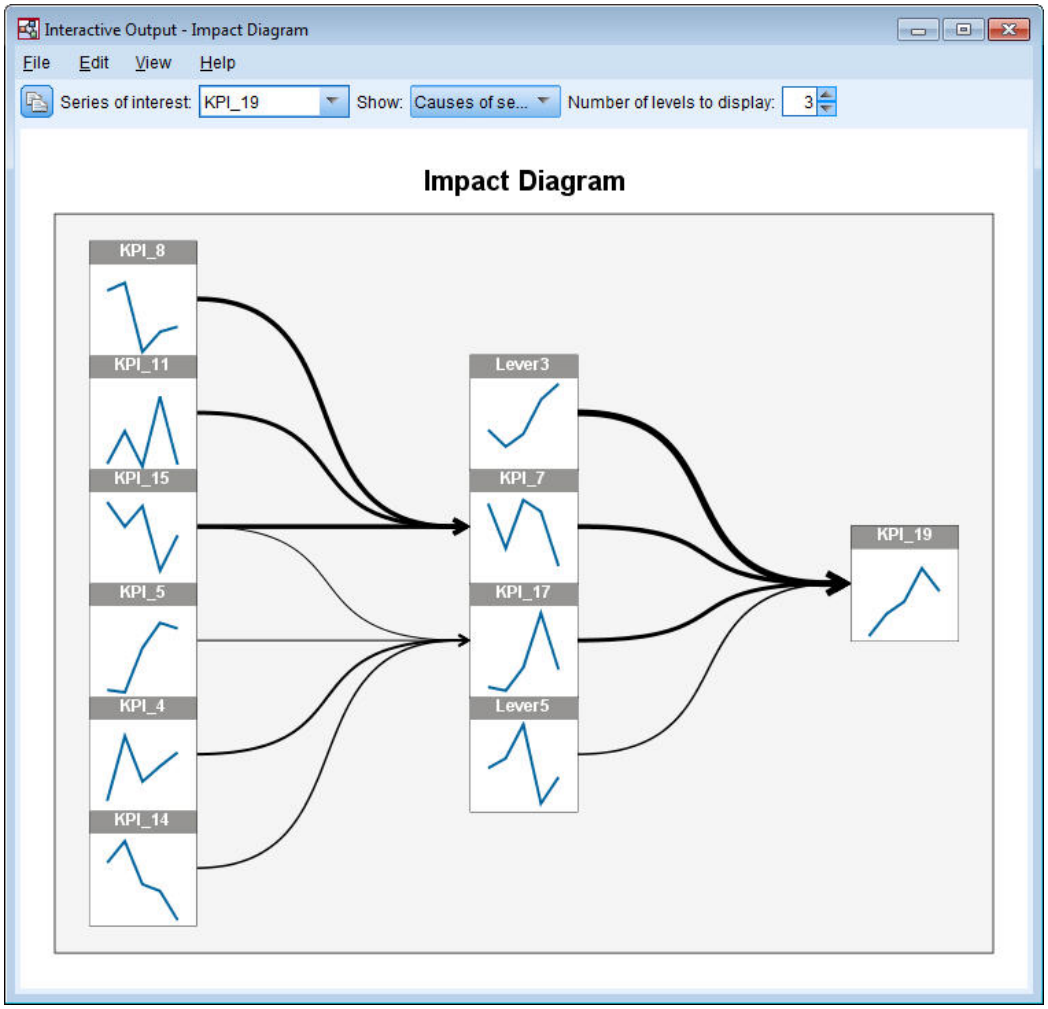


Рисунок 406. Диаграмма воздействия причин

Это представление показывает, что у модели для *KPI_19* четыре входных поля и что у *Lever3* наиболее значимое причинное соединение с *KPI_19*. Оно также показывает ряды, косвенно влияющие на *KPI_19* посредством воздействия их эффектов на *KPI_7* и *KPI_17*. К причинам применима та же концепция уровней, что и обсуждавшаяся для эффектов. Значение **Число выводимых уровней** точно также можно изменить, чтобы выводилось больше или меньше уровней причин.

Определение основных причин выбросов

Принимая во внимание систему причинных моделей времени, можно пойти дальше обнаружения выбросов и определить ряд, с наибольшей вероятностью вызывающий конкретный выброс. Этот процесс называется анализом первопричин выбросов, и его надо запрашивать ряд за рядом. Для анализа требуется система причинных моделей времени и данные, которые использовались для построения этой системы. В этом примере активный набор данных представляет собой данные, использованные для построения системы моделей.

Для выполнения анализа первопричин выбросов:

1. В диалоговом окне TCM (Temporal Causal Model - Причинная модель времени) перейдите на вкладку **Опции построения**, а затем выберите **Ряды для вывода** в списке **Выберите элемент**.

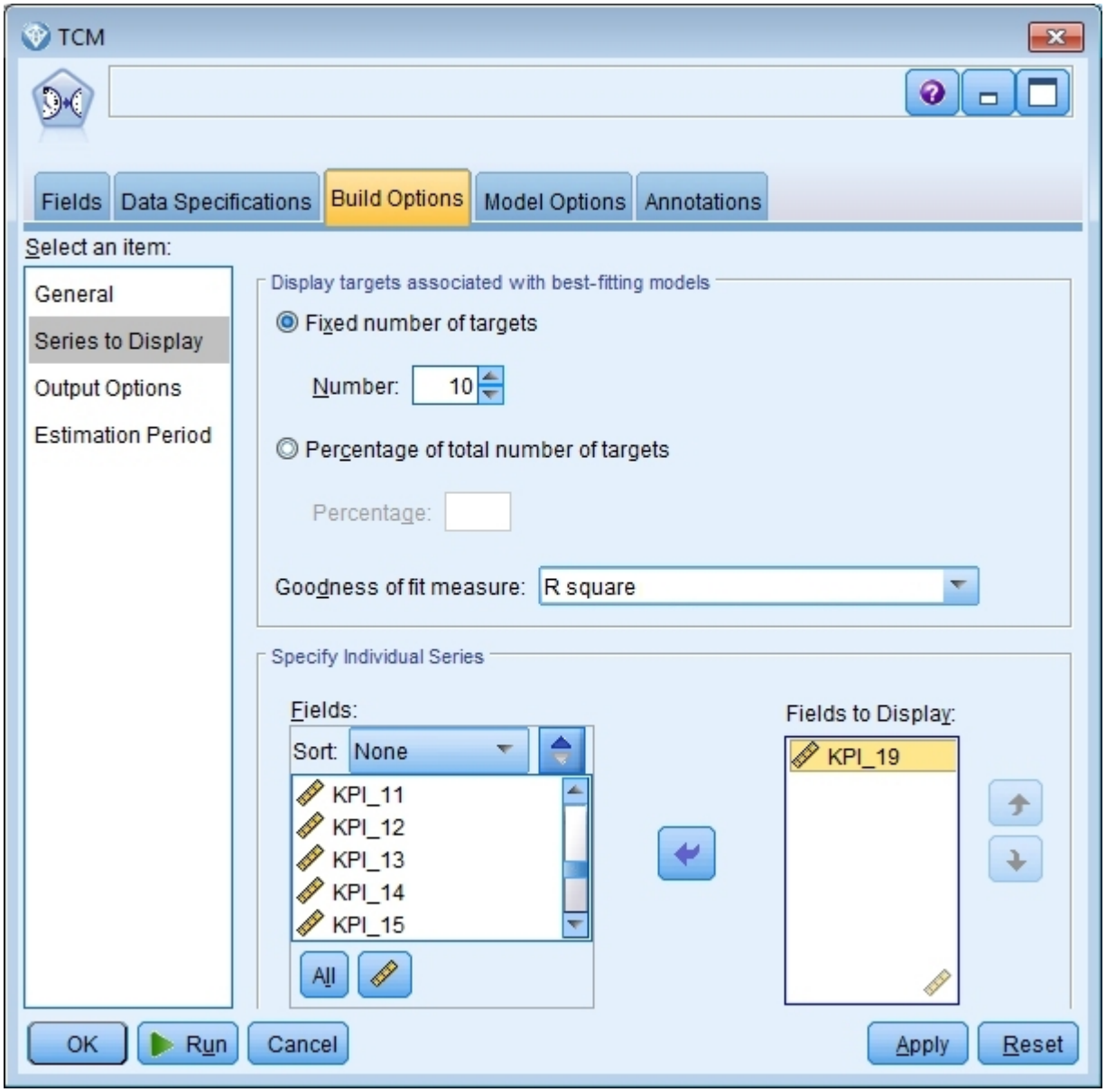


Рисунок 407. Выводимые ряды причинных моделей времени

2. Переместите *KPI_19* в список **Поля для вывода**.

3. На вкладке Опции в списке **Выберите элемент** выберите **Опции вывода**.

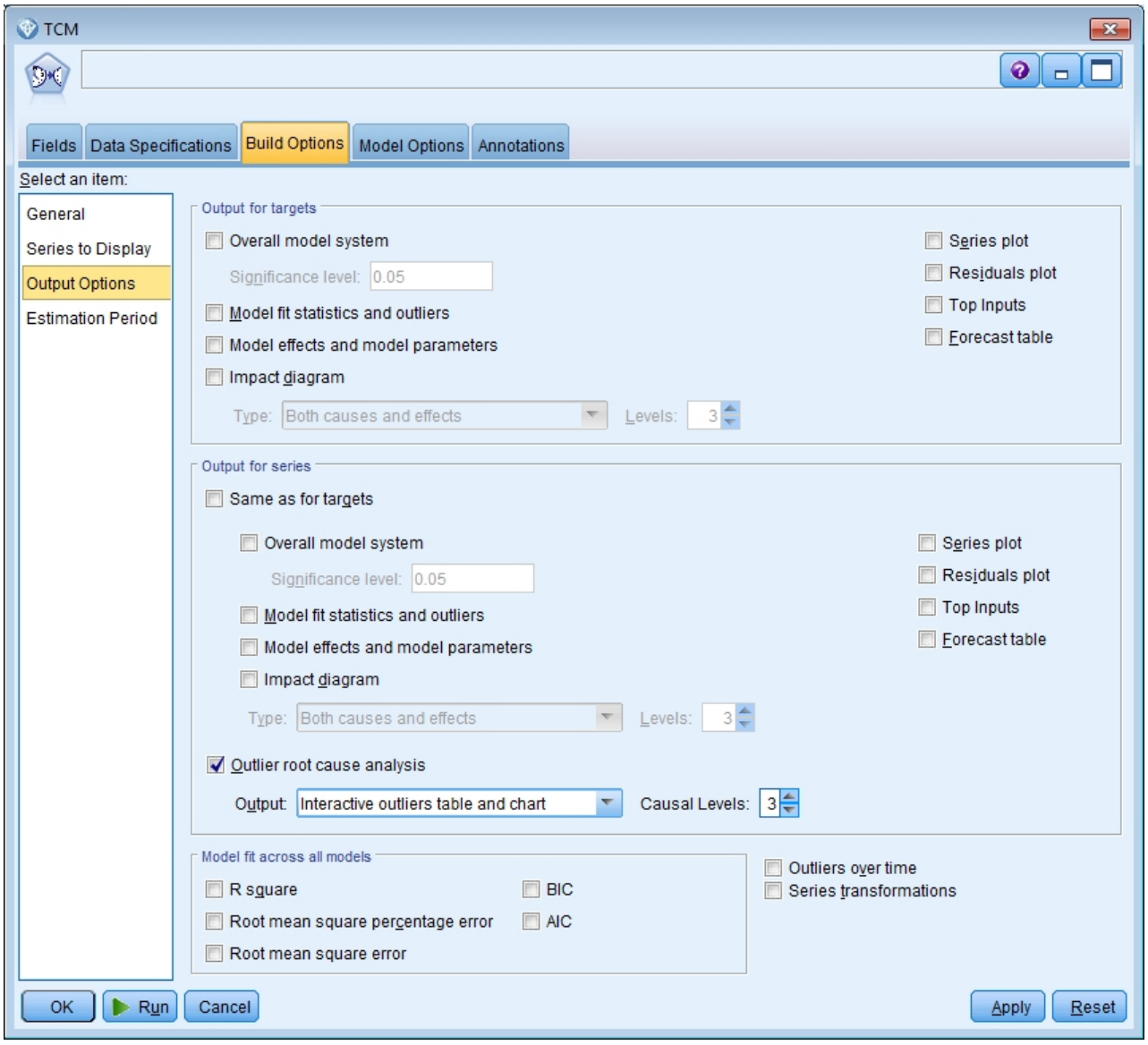
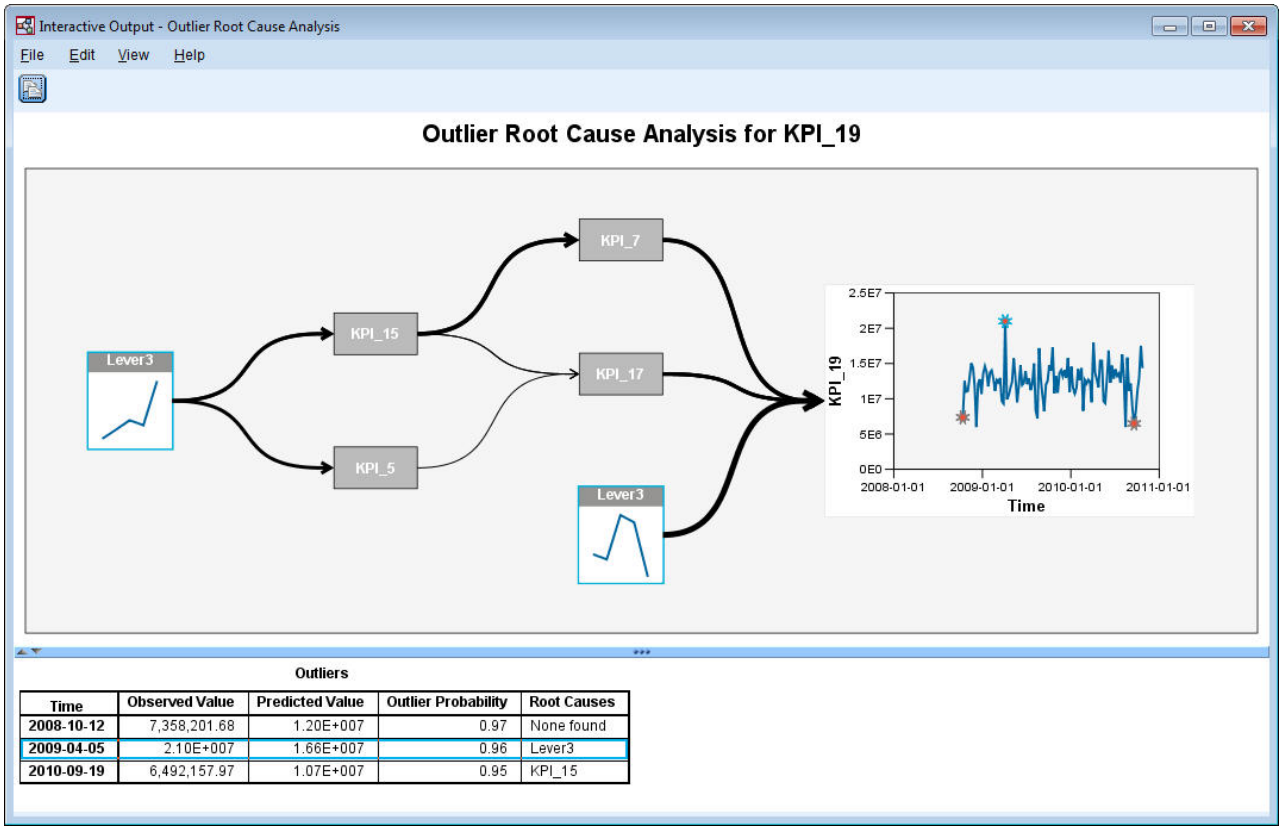


Рисунок 408. Опции вывода причинных моделей времени

4. Выключите опции **Общая система моделей**, **То же, что и для назначений**, **R-квадрат** и **Преобразования рядов**.
5. Выберите **Анализ первопричин выбросов** и сохраните существующие значения опций **Вывод** и **Уровни причин**.
6. Нажмите кнопку **Выполнить**.
7. Щелкните дважды мышью по диаграмме Анализ первопричин выбросов для *KPI_19* в средстве просмотра, чтобы ее активировать.



Time	Observed Value	Predicted Value	Outlier Probability	Root Causes
2008-10-12	7,358,201.68	1.20E+007	0.97	None found
2009-04-05	2.10E+007	1.66E+007	0.96	Lever3
2010-09-19	6,492,157.97	1.07E+007	0.95	KPI_15

Рисунок 409. Анализ первопричин выбросов для KPI_19

Сводка результатов анализа содержится в таблице Выбросы. Эта таблица показывает, что были найдены основные причины для выбросов за 2009-04-05 и 2010-09-19, но основная причина выброса за 2008-10-12 найдена не была. При щелчке мышью по строке в таблице Выбросы выделяется путь к ряду основных причин, как здесь показано для выброса за 2009-04-05. Это действие выделяет также выбранный выброс на диаграмме последовательности. Можно также щелкнуть по значку для выброса непосредственно на диаграмме последовательности, чтобы выделить путь к ряду основной причины для этого выброса.

Для выброса за 2009-04-05, основной причиной является *Lever3*. Диаграмма показывает, что *Lever3* - это непосредственный элемент вывода в *KPI_19*, но он также косвенно влияет на *KPI_19* посредством его эффекта на другие ряды, влияющие на *KPI_19*. Один из конфигурируемых параметров для анализа первопричин выбросов - число уровней причин для поиска основных причин. По умолчанию поиск выполняется по трём уровням. Вхождения ряда основных причин выводятся максимально до заданного числа уровней причин. В этом примере вхождения *Lever3* есть на первом уровне причин и на третьем уровне причин.

Каждый узел в выделенном пути для выброса содержит диаграмму, диапазон времени которой зависит от уровня, на котором находится это узел. Узлам на первом уровне причин соответствует диапазон с T-1 по T-L, где T - это время, в которое произошёл выброс, а L - количество шагов задержки, включаемых в каждую модель. Узлам на втором уровне причин соответствует диапазон с T-2 по T-L-1, а узлам на третьем уровне причин соответствует диапазон с T-3 по T-L-2. Щелкнув один раз по связанному узлу, можно получить подробную диаграмму последовательности.

Выполнение сценариев

Принимая во внимание систему причинных моделей времени, можно выполнять пользовательские сценарии. *Сценарий* определяется временным рядом, который называют *корневым рядом*, и набором пользовательских значений для этого ряда на указанный диапазон времени. Эти заданные значения используются затем для генерирования предсказаний для временных рядов, на которые влияет корневой ряд. Для анализа требуется система причинных моделей времени и данные, которые использовались для построения этой системы. В этом примере активный набор данных представляет собой данные, использованные для построения системы моделей.

Для выполнения сценариев:

1. В диалоговом окне TCM (Temporal Causal Model - Причинная модель времени) нажмите кнопку **Анализ сценариев**.
2. В диалоговом окне Сценарии причинных моделей времени нажмите кнопку **Определить период сценария**.

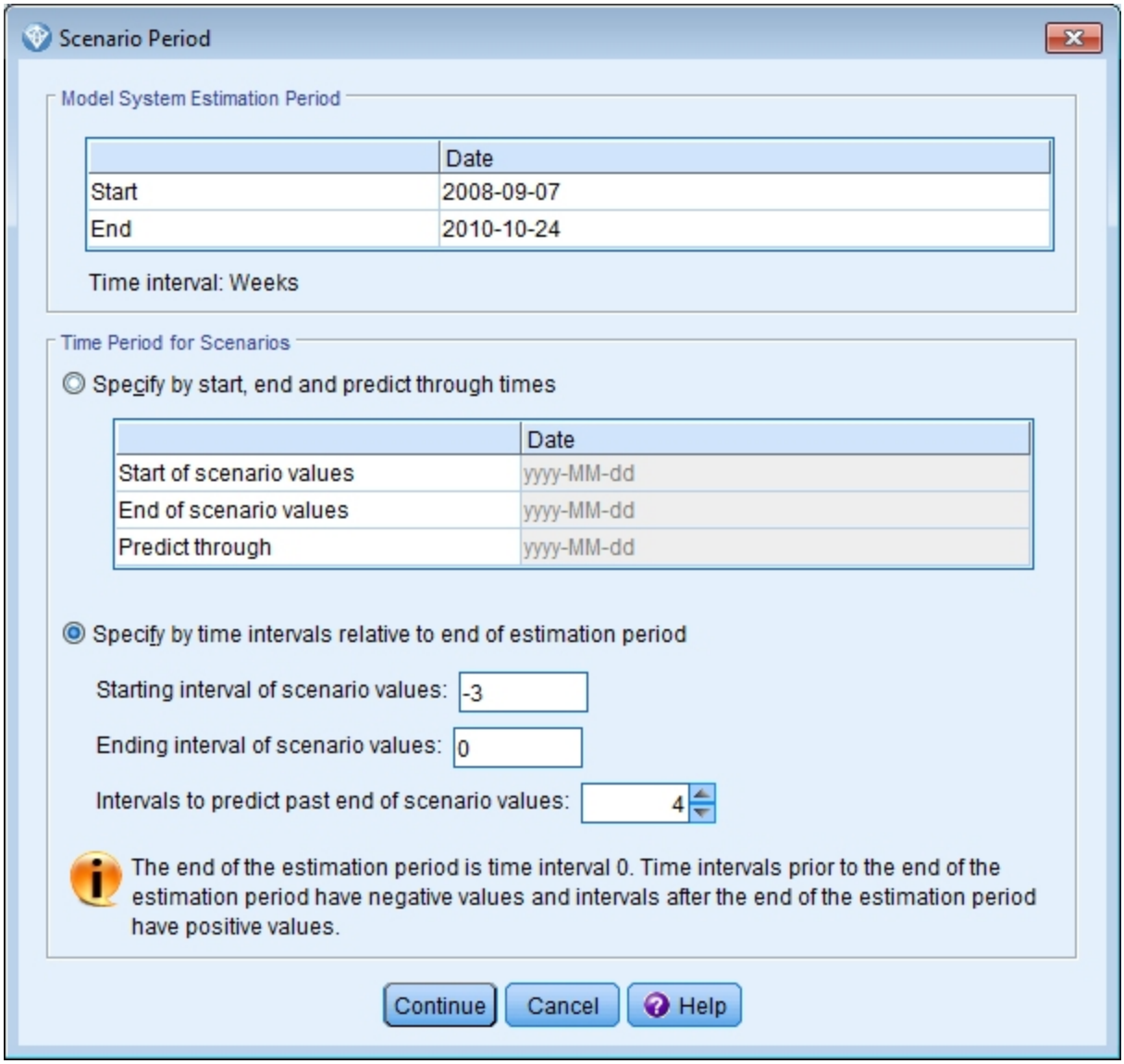


Рисунок 410. Период сценария

3. Выберите **Задать по интервалам времени относительно окончания периода оценки**.

4. Введите `-3` для начального интервала и `0` для интервала окончания.

   Эти параметры задают, что каждый сценарий основан на значениях, задаваемых для последних четырех интервалов времени в период оценки. Для этого примера последние четыре интервала - это последние четыре недели. Диапазон времени, на который задаются значения сценария, называется *периодом сценария*.

5. Введите `4` для интервалов предсказания на время после значений окончания сценария.

   Этот параметр указывает, что будут сгенерированы предсказания для четырех интервалов времени после окончания периода сценария.

6. Нажмите кнопку **Продолжить** .
7. Нажмите кнопку **Добавить сценарий** на вкладке Сценарии.

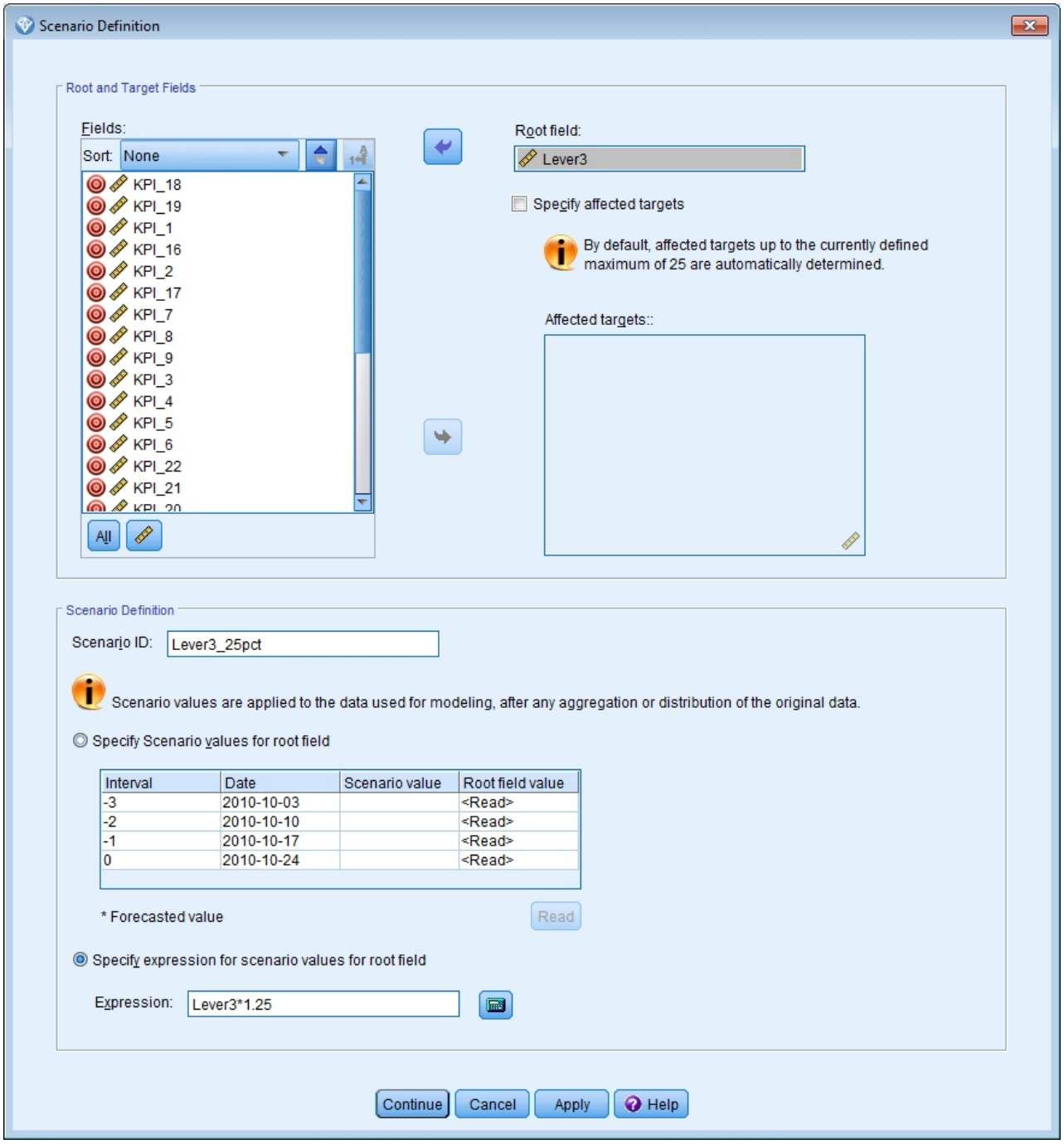


Рисунок 411. Определение сценария

8. Переместите *Lever3* в поле **Корневое поле** чтобы исследовать, как заданные значения *Lever3* в периоде сценария влияют на предсказания других рядов, которые причинно затрагивает *Lever3*.
9. Для ID сценария введите `Lever3_25pct`.
10. Выберите **Задайте выражение для значений сценария корневого поля** и в качестве выражения введите `Lever3*1.25`.

    Этот параметр указывает, что значения для *Lever3* в периоде сценария на 25% больше наблюдаемых значений. Для более сложных выражений можно использовать построитель выражений, щелкнув по значку калькулятора.
11. Нажмите кнопку **Продолжить** .

12. Повторите шаги с 10 по 14, чтобы определить сценарий с *Lever3* для корневого поля, `Lever3_50pct` для ID сценария и `Lever3*1.5` для выражения.

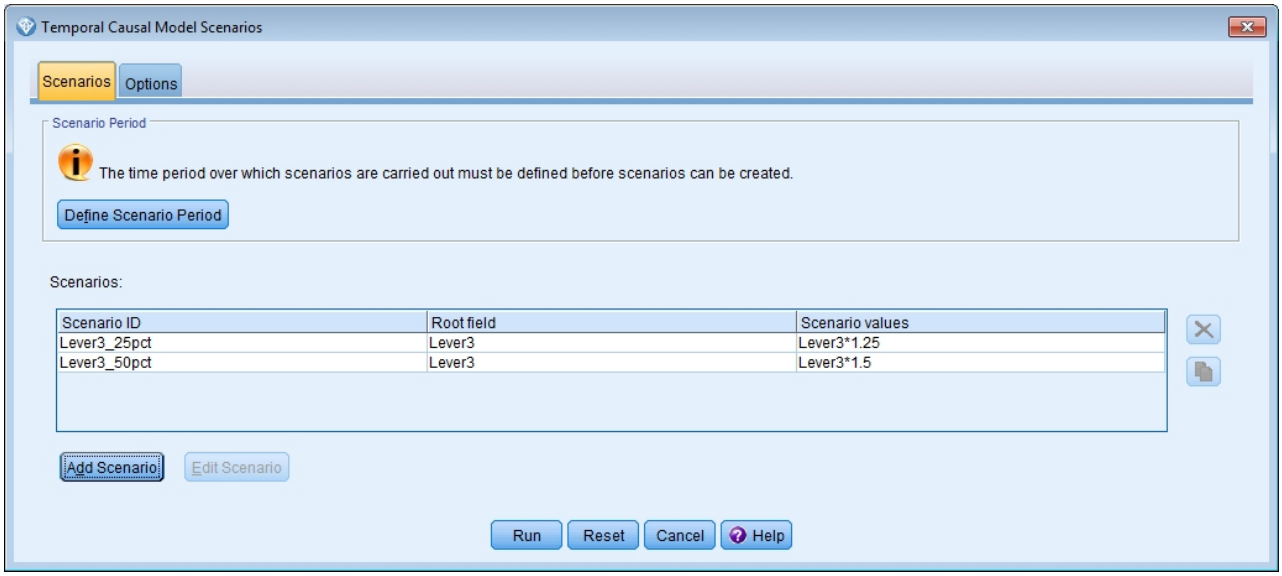
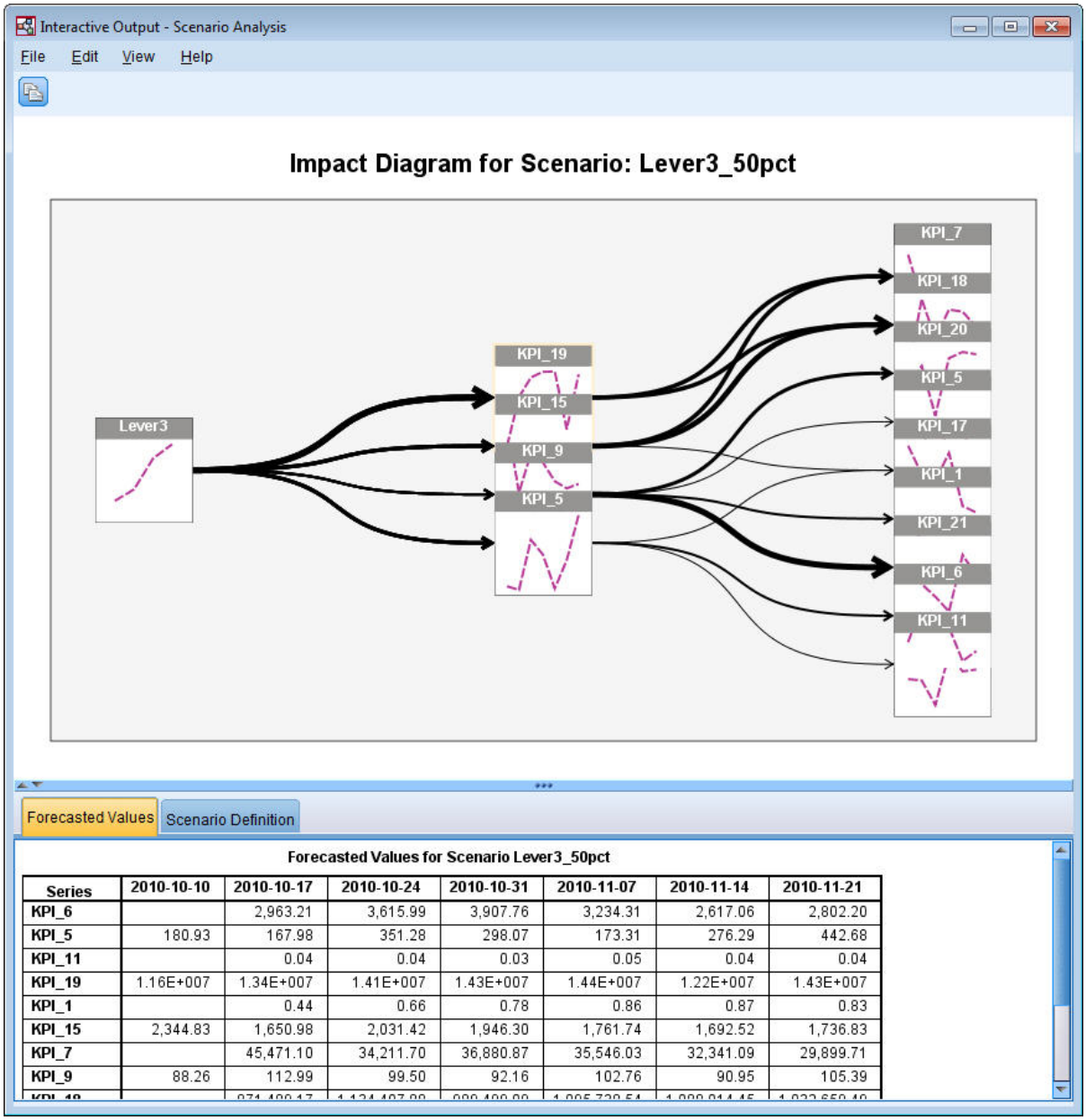


Рисунок 412. Сценарии

13. Щелкните по вкладке **Опции** и введите *2* в качестве максимального уровня для затронутых назначений.
14. Нажмите кнопку **Выполнить**.
15. Щелкните дважды мышью по диаграмме воздействия для *Lever3_50pct* в средстве просмотра, чтобы ее активировать.



Series	2010-10-10	2010-10-17	2010-10-24	2010-10-31	2010-11-07	2010-11-14	2010-11-21
KPI_6		2,963.21	3,615.99	3,907.76	3,234.31	2,617.06	2,802.20
KPI_5	180.93	167.98	351.28	298.07	173.31	276.29	442.68
KPI_11		0.04	0.04	0.03	0.05	0.04	0.04
KPI_19	1.16E+007	1.34E+007	1.41E+007	1.43E+007	1.44E+007	1.22E+007	1.43E+007
KPI_1		0.44	0.66	0.78	0.86	0.87	0.83
KPI_15	2,344.83	1,650.98	2,031.42	1,946.30	1,761.74	1,692.52	1,736.83
KPI_7		45,471.10	34,211.70	36,880.87	35,546.03	32,341.09	29,899.71
KPI_9	88.26	112.99	99.50	92.16	102.76	90.95	105.39

Рисунок 413. Диаграмма воздействия для сценария: Lever3_50pct

Диаграмма воздействия показывает ряды, затронутые корневым рядом *Lever3*. Выводятся два уровня эффектов, поскольку в качестве максимального уровня для затронутых назначений вы задали *2*.

Таблица Прогнозные значения содержит предсказания для всех рядов, затронутых *Lever3*, вплоть до второго уровня эффектов. Предсказания для рядов назначения на первом уровне эффектов начинаются в первый период времени после начала периода сценария. В этом примере предсказания для рядов назначения на первом уровне эффектов начинаются 2010-10-10. Предсказания для рядов назначения на втором уровне эффектов начинаются во второй период времени после начала периода сценария. В этом примере предсказания для рядов назначения на втором уровне эффектов начинаются 2010-10-17. Ступенчатый характер предсказаний отражает тот факт, что модели временных рядов основаны на значениях входных данных, сдвинутых относительно друг друга.

16. Щелкните по узлу для *KPI_5*, чтобы сгенерировать подробную диаграмму последовательности.

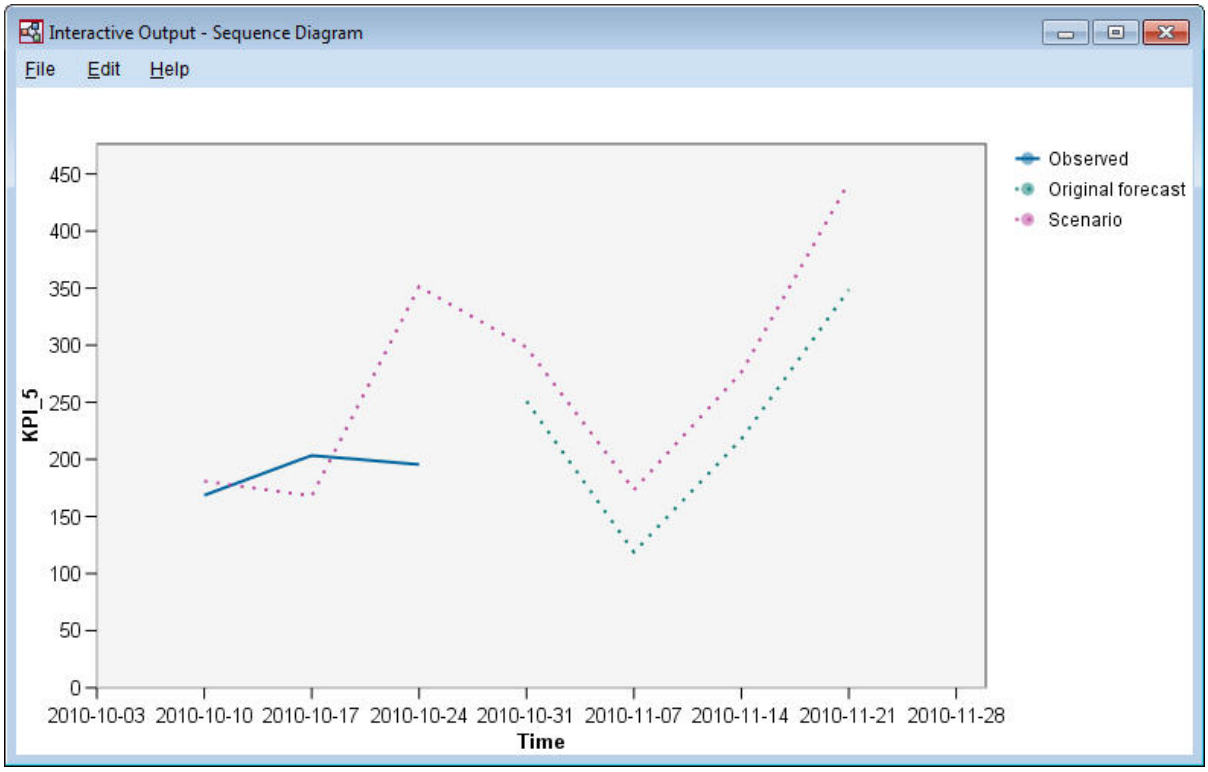


Рисунок 414. Диаграмма последовательности для KPI_5

Диаграмма последовательности показывает предсказанные значения из сценария, а также показывает значения ряда в отсутствие сценария. Если период сценария содержит моменты времени в периоде оценки, выводятся наблюдаемые значения ряда. Для моментов времени после окончания периода оценки выводятся исходные прогнозы.

Уведомления

Эта информация относится к продуктам и сервису, предлагаемым в США. Этот материал может быть доступен от IBM на других языках. Однако для его получения может понадобиться приобрести продукт или версию продукта на нужном языке.

IBM может не предоставлять в других странах продукты, услуги и аппаратные средства, описанные в данном документе. За информацией о продуктах и услугах, предоставляемых в вашей стране, обращайтесь к местному представителю IBM. Ссылки на продукты, программы или услуги IBM не означают и не предполагают, что можно использовать только указанные продукты, программы или услуги IBM. Разрешается использовать любые функционально эквивалентные продукты, программы или услуги, если при этом не нарушаются права IBM на интеллектуальную собственность. Однако ответственность за оценку и проверку работы любого продукта, программы или сервиса, не произведенного корпорацией IBM, лежит на пользователе.

IBM может располагать патентами или рассматриваемыми заявками на патенты, относящимися к предмету данного документа. Предъявление данного документа не предоставляет какую-либо лицензию на эти патенты. Вы можете послать письменный запрос о лицензии по адресу:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

По поводу лицензий, связанных с использованием наборов двухбайтных символов (DBCS), обращайтесь в отдел интеллектуальной собственности IBM в вашей стране или направьте запрос в письменной форме по адресу:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

INTERNATIONAL BUSINESS MACHINES CORPORATION ПРЕДСТАВЛЯЕТ ДАННУЮ ПУБЛИКАЦИЮ "КАК ЕСТЬ", БЕЗ КАКИХ-ЛИБО ГАРАНТИЙ, КАК ЯВНЫХ, ТАК И ПОДРАЗУМЕВАЕМЫХ, ВКЛЮЧАЯ, НО НЕ ОГРАНИЧИВАЯСЬ ТАКОВЫМИ, ПРЕДПОЛАГАЕМЫЕ ГАРАНТИИ СОБЛЮДЕНИЯ ЧЬИХ-ЛИБО АВТОРСКИХ ПРАВ, ВОЗМОЖНОСТИ КОММЕРЧЕСКОГО ИСПОЛЬЗОВАНИЯ ИЛИ ПРИГОДНОСТИ ДЛЯ КАКИХ-ЛИБО ЦЕЛЕЙ И СООТВЕТСТВИЯ ОПРЕДЕЛЕННОЙ ЦЕЛИ. В некоторых странах для ряда сделок не допускается отказ от явных или предполагаемых гарантий; в таком случае данное положение к вам не относится.

Эта информация может содержать технические неточности и типографские ошибки. В представленную здесь информацию периодически вносятся изменения; эти изменения будут включаться в новые издания данной публикации. Фирма IBM может в любое время без уведомления вносить изменения и усовершенствования в продукты и программы, описанные в этой публикации.

Любые ссылки в этой публикации на сайты, не принадлежащие IBM, приведены только для удобства и никоим образом не означают их поддержки. Материалы на этих сайтах не входят в число материалов по данному продукту IBM, и весь риск пользования этими сайтами несете вы сами.

Любую предоставленную вами информацию IBM может использовать или распространять любым способом, какой сочтет нужным, не беря на себя никаких обязательств по отношению к вам.

Если обладателю лицензии на данную программу понадобятся сведения о возможности: (i) обмена данными между независимо разработанными программами и другими программами (включая данную) и (ii) совместного использования таких данных, он может обратиться по адресу:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

Такая информация может быть доступна при соответствующих условиях и соглашениях, включая в некоторых случаях взимание платы.

Описанную в данном документе лицензионную программу и все прилагаемые к ней лицензированные материалы IBM предоставляет на основе положений Соглашения между IBM и Заказчиком, Международного Соглашения о Лицензиях на Программы IBM или любого эквивалентного соглашения между IBM и заказчиком.

Данные производительности и примеры клиентов представлены только для иллюстрации. Фактическая производительность зависит от конкретной конфигурации и условий работы.

Информация о продуктах других компаний (не IBM) получена от поставщиков этих продуктов, из их опубликованных объявлений или из иных общедоступных источников. IBM не производила тестирование этих продуктов и никак не может подтвердить информацию о их точности работы и совместимости, а также прочие заявления относительно продуктов других компаний (не IBM). Вопросы о возможностях продуктов других компаний (не IBM) следует направлять поставщикам этих продуктов.

Все утверждения о будущих планах и намерениях IBM могут быть изменены или отменены без уведомлений, и описывают исключительно цели фирмы.

Эти сведения содержат примеры данных и отчетов, используемых в повседневных деловых операциях. Чтобы проиллюстрировать их настолько полно, насколько это возможно, данные примеры включают имена индивидуумов, названия компаний, брендов и продуктов. Все эти имена являются вымышленными и любое их сходство с реальными именами и адресами предприятий является случайным.

Товарные знаки

IBM, логотип IBM, и ibm.com являются товарными знаками или зарегистрированными товарными знаками компании International Business Machines Corp., зарегистрированными во многих странах мира. Прочие наименования продуктов и услуг могут быть товарными знаками, принадлежащими IBM или другим компаниям. Текущий список товарных знаков IBM смотрите на веб-сайте "Copyright and trademark information" (Информация об авторских правах и товарных знаках) по адресу www.ibm.com/legal/copytrade.shtml.

Adobe, логотип Adobe, PostScript и логотип PostScript являются либо зарегистрированными товарными знаками, либо товарными знаками корпорации Adobe Systems в Соединенных Штатах и/или других странах.

Intel, логотип Intel, Intel Inside, логотип Intel Inside, Intel Centrino, логотип Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium и Pentium являются товарными знаками или зарегистрированными товарными знаками компании Intel или ее дочерних компаний в Соединенных Штатах и других странах.

Linux является зарегистрированным товарным знаком Linus Torvalds в Соединенных Штатах и других странах.

Microsoft, Windows, Windows NT и логотип Windows являются товарными знаками корпорации Microsoft в Соединенных Штатах и других странах.

UNIX является зарегистрированным товарным знаком The Open Group в Соединенных Штатах и других странах.

Java и все основанные на Java товарные знаки и логотипы - товарные знаки или зарегистрированные товарные знаки Oracle и/или его филиалов.

Правила и условия для документации продукта

Разрешения для использования этих публикаций предоставляются на следующих условиях.

Применимость

Данные правила и условия являются дополнением к правилам использования для сайта IBM.

Персональное использование

Вы можете воспроизводить эти публикации для персонального некоммерческого использования при условии сохранения всех замечаний о правах собственности. Вам запрещается распространять эти публикации, полностью или по частям, демонстрировать их или создавать из них производные продукты без явного на то согласия от IBM.

Коммерческое использование

Вам предоставляется право воспроизводить эти публикации исключительно в пределах своего предприятия при условии, что будут воспроизведены все замечания об авторских правах. За пределами вашего предприятия вам запрещается распространять эти публикации, полностью или по частям, демонстрировать их или создавать из них производные продукты без явного на то согласия от IBM.

Права

За исключением прав, явным образом предоставляемых настоящим разрешением, никаких иных разрешений, лицензий и прав, ни явных, ни подразумеваемых, в отношении публикаций и любой содержащейся в них информации, данных, программ или иной интеллектуальной собственности, не предоставляется.

IBM оставляет за собой право отозвать разрешения, предоставленные этим документом, если, по мнению IBM, использование публикаций наносит ущерб IBM или, как это установлено IBM, вышеприведенные инструкции не соблюдаются должным образом.

Запрещается загружать, экспортировать или реэкспортировать эту информацию, если при этом не будут полностью соблюдаться все применимые законы и постановления, включая все законы и постановления США, касающиеся экспорта.

IBM НЕ ДАЕТ НИКАКИХ ГАРАНТИЙ ОТНОСИТЕЛЬНО СОДЕРЖАНИЯ ЭТИХ ПУБЛИКАЦИЙ. ПУБЛИКАЦИИ ПРЕДСТАВЛЯЮТСЯ "КАК ЕСТЬ", БЕЗ КАКИХ-ЛИБО ГАРАНТИЙ, ЯВНЫХ ИЛИ ПОДРАЗУМЕВАЕМЫХ, ВКЛЮЧАЯ (НО НЕ ОГРАНИЧИВАЯСЬ ТАКОВЫМИ) ПРЕДПОЛАГАЕМЫЕ ГАРАНТИИ ОТСУТСТВИЯ НАРУШЕНИЙ, КОММЕРЧЕСКОЙ ПРИГОДНОСТИ ИЛИ СООТВЕТСТВИЯ ОПРЕДЕЛЕННОЙ ЦЕЛИ.

Индекс

Спец. символы

C

E

I

M

U

A

B

Г

Д

E

З

И

К

Л

M

H

Напечатано в Дании