# Курс по интеллектуальному анализу данных (Data Mining) с использованием платформы MS SQL 2005 . Чаcть 1

<http://www.interface.ru/home.asp?artId=21179>

Источник: [businessdataanalytics](http://www.interface.ru/iservices/goto.asp?Url=enc:rn%7Ej%40%255X%7Fisdoiy%5Eknk%5Bt%5Bvs%7Ecii4h%7F%25N%5B%7E%5BWctct%5DYKVIoh%7C%5Fx%28%3A%2A%3F%27%3F%294b%7Eg)

Seth Paul, Jamie MacLennan, Zhaohui Tang, Scott Oveson

**Data Mining** - современная технология анализа информации с целью нахождения в накопленных данных ранее неизвестных, нетривиальных и практически полезных знаний, необходимых для принятия оптимальных решений в различных областях человеческой деятельности.

**Microsoft SQL Server 2005** предоставляет интегрированную среду для создания моделей Data Mining и работы с ними. В курсе описаны четыре сценария (адресная рассылка, прогнозирование, анализ покупательской корзины и анализ последовательности покупок) для демонстрации методов использования алгоритмов Data Mining, а также сопутствующего инструментария, входящего в состав **SQL Server Analysis Services 2005**.

Средства анализа данных **OLAP** и **Data Mining** объединены в две среды разработки: **Business Intelligence Development Studio** и SQL Server Management Studio. В Business Intelligence Development Studio можно создавать отсоединенные от сервера проекты. В этом случае готовый проект можно загрузить на сервер, но ничто не мешает работать с проектом, работающим с сервером напрямую.

Все средства для управления моделями Data Mining доступны в редакторе моделей. С его помощью можно создавать, просматривать, сравнивать разные модели, а также создавать на их базе прогнозы.

После создания модели можно провести ее анализ на предмет выявления интересных для нас шаблонов (паттернов) и правил. В зависимости от применяемого алгоритма в среде разработки используется различные представления просматриваемой модели.

Поскольку зачастую проект содержит несколько моделей Data Mining, предварительно созданных на основе наших предположений, то возникает задача определения наиболее адекватной из них. Для решения этой задачи в редакторе присутствует средство сравнения моделей Mining Accuracy Chart. С использованием этого инструмента можно предсказать точность модели и выбрать лучшую их них.

Для создания прогнозов используется язык **Data Mining Extensions (DMX)**, который является расширением SQL и содержит команды для создания, изменения и осуществления предсказаний на основании различных моделей. Создание таких прогнозов может быть сложной задачей, поэтому в редакторе присутствует инструмент под названием Prediction Query Builder, которых представляет собой визуальное средство создания DMX-запросов.

Кроме инструментария для работы с моделями, не менее значимыми являются и способы создания моделей. Ключевым моментом создания модели является выбор алгоритма обнаружения данных. SQL Server 2005 Analysis Services включает в себя следующие девять алгоритмов:

* Дерево решений (Microsoft Decision Trees)
* Кластеризация (Microsoft Clustering)
* "Наивный" Байес (Microsoft Na?ve Bayes)
* Кластеризация последовательностей (Microsoft Sequence Clustering)
* Временные ряды (Microsoft Time Series)
* Ассоциативные правила (Microsoft Association)
* Нейронная сеть (Microsoft Neural Network)
* Линейная регрессия (Microsoft Linear Regression)
* Логистическая регрессия (Microsoft Logistic Regression)

Используя комбинацию этих алгоритмов можно создавать решения для большей части встречающихся задач по обнаружению скрытых закономерностей в больших объемах данных.

Наиболее важные шаги создания модели Data Mining состоят в получении, объединении, очистке и предобработке данных для дальнейшего использования алгоритмом. В состав SQL Server 2005 входят средства преобразования данных - SQL Server Integration Services (SSIS), которые позволяют объединять, очищать, проверять и проводить предварительную обработку данных.

Для демонстрации возможностей SQL Server мы будем использовать демо-проект AdventureWorksDW. Эта база включена в поставку SQL Server 2005 и спроектирована должным образом для работы с OLAP и моделями Data Mining. Чтобы эта база была доступна следует выбрать опцию установки этого проекта во время инсталляции SQL Server.

# Проект Adventure Works

Проект AdventureWorksDW описывает работу производителя велосипедов - компании "Adventure Works Cycles". Компания занимается производством и реализацией велосипедов из металлических и композиционных материалов на территории Северной Америки, Европы и Азии. Головное производство, которое располагает 500 сотрудниками, находится в городе Bothell, штат Вашингтон. Несколько региональных офисов находятся непосредственно на территории рынков сбыта.

Компания реализует продукцию оптом для специализированных магазинов и в розницу через интернет. Для решения демонстрационных задач мы будем использовать в базе AdventureWorksDW данные об интернет продажах, поскольку они содержат данные, хорошо подходящие для анализа.

### База данных

В базе интернет продаж содержатся сведения о 9 242 клиентах, которые живут в 6 странах, относящихся к трем регионам:

 Серверная Америка (83%)

 Европа (12%)

 Австралия (7%)

Данные относятся к периоду с 2002 по 2004 годы. Вся продукция классифицирована по категориям, моделям и товарам.

## Среда разработки - Business Intelligence Development Studio

Среда разработки "Business Intelligence Development Studio" представляет собой интегрированную среду разработки для создания проектов анализа данных (business intelligence) и включает в себя множество инструментов. Она позволяет работать с моделями в отсоединенном от сервера режиме, а затем развертывать обновленный проект на сервере.

Можно отметить следующие преимущества работы в единой среде:

* Среда является мощным средством управления проектами
* Вы можете объединять несколько проектов в одно решение (solution) и использовать для них общее представление.
* Полная интеграция с системой хранения исходного кода позволяет удобно организовывать совместную работу над проектом.

Проект Analysis Services является центральным в решении business intelligence. Проект включает в себя модели Data Mining и кубы OLAP, а также вспомогательные объекты наполнения аналитической базы данных. Из студии можно создавать проекты анализа данных и развертывать их на множестве серверов.

При работе с ранее созданным проектом, который уже находится на сервере, можно работать с ним напрямую.

## Среда управления - SQL Server Management Studio

Среда SQL Server Management Studio является оболочкой для множества административных функций для управления компонентами SQL Server. Эта среда отличается от Business Intelligence Development Studio в первую очередь тем, что изменения сохраняются на сервер в реальном времени.

После предварительной очистки и предобработки данных большая часть всех дальнейших действий по созданию модели выполняется в среде Business Intelligence Development Studio. В ней выполняются действия по созданию и проверке моделей, а также итеративное тестирование для выявления наиболее адекватной модели. После получения нужной модели ее можно развернуть на сервере. После этого фокус смещается с разработки на поддержку и использование модели для чего подходит SQL Server Management Studio. С ее помощью можно администрировать базу данных и выполнять просмотр и создание предсказаний аналогично тому, как это выполняется в среде Business Intelligence Development Studio

## Сервисы интеграции - Integration Services

Integration Services (сервисы интеграции, SSIS) включают в себя средства получения (Extract), преобразования (Transform) и загрузки (Load) (ETL) данных из различных источников. Они используются для некоторых очень важных задач в проектах извлечения данных, таких как извлечение, объединение, очистка и предобработка данных, предшествующие созданию модели. В проектах извлечения данных и в процессе настройки моделей обычно необходимо проводить неоднократную предобработку и очистку данных. С помощью SSIS можно объединить задачи предобработки данных в единый пакет.

Среда SSIS содержит визуальный редактор, с помощью которого удобно создавать и отлаживать пакеты преобразования данных. Из него можно загрузить пакеты на сервер и настроить планировщик для их выполнения. Это удобно использовать, к примеру, для автоматической предобработки новых поступающих данных.

Алгоритмы Data Mining

**Оглавление**

Алгоритмы Data Mining являются основой для создания моделей. Набор алгоритмов Data Mining, входящих в SQL Server 2005, позволяют производить различные виды анализа данных.

###  Дерево решений - Microsoft Decision Trees

Алгоритм "Дерево решений" предназначен для решения задач классификации и регрессии и хорошо подходит для прогнозирования. В алгоритме Microsoft Decision Trees используются как дискретные, так и непрерывные атрибуты.

В процессе построения модели алгоритм итеративно вычисляет степень влияния каждого входного атрибута модели на значения выходного атрибута и использует атрибут, влияющий на выходную переменную в наибольшей степени для разбиения узла дерева решений. Узел верхнего уровня описывает распределение значений выходного атрибута по всей совокупности данных. Каждый последующий узел описывается распределением выходного атрибута при соблюдении условий на входные атрибуты, соответствующие этому узлу. Модель продолжает расти до тех пор, пока разбиение узла на последующие узлы увеличивает вероятность того, что выходной атрибут будет принимать какое-то определенное значение по сравнению со всеми другими значениями, т.е. разбиение увеличивает качество прогноза. Алгоритм осуществляет поиск атрибутов и их значений, разбиение по которым позволяет с большей вероятностью правильно предсказать значение выходного атрибута.

### Кластеризация - Microsoft Clustering

Алгоритм кластеризации использует итеративный метод группировки записей набора данных в кластеры, обладающие сходными характеристиками. Используя разбиение на кластеры можно выявить в исследуемом массиве данных такие связи, которые невозможно обнаружить простым просмотром этих данных. Кроме того, с помощью алгоритмов кластеризации можно осуществлять прогнозирование. К примеру, объединить в группу людей, которые живут в одном районе, водят одну марку машин, имеют сходные предпочтения в пище и покупают один тип продукции. Такое объединение и есть кластер. Другой кластер может включать в себя людей, посещающих один ресторан, имеющих один уровень дохода и ездящих дважды в год в отпуск в другие страны. Оценивая распределение данных в этих кластерах, можно лучше понять взаимосвязи различных характеристик исследуемых объектов, а также как эти взаимосвязи влияют на значение прогнозируемого атрибута.

В Microsoft Analysis Services 2005 для кластеризации используется модификации алгоритмов максимизации ожидания (Expectation Maximization) и K-ближайших соседей (K-Means).

В первом случае каждый кластер характеризуется своей функцией распределения атрибутов входных данных, которые вычисляются в ходе итеративного процесса. В процессе выполнения алгоритма итеративно максимизируется функция правдоподобия на пространстве параметров функций распределения для каждого кластера. При этом предполагается, что непрерывные атрибуты имеют совместное многомерное нормальное распределение, а дискретные - совместное дискретное распределение.

Во втором случае итеративно минимизируется сумма квадратов расстояний (в различных метриках) от каждого элемента данных до центра соответствующего кластера.

### Наивный алгоритм Байеса - Microsoft Na?ve Bayes

Наивный алгоритм Байеса предназначен для решения задач классификации и прогнозирования. В процессе его реализации вычисляются вероятности состояний входных атрибутов для каждого состояния выходного атрибута. Эти значения используются для вычисления вероятности того, что выходной атрибут принимает то или иное состояние при заданных значениях входных атрибутов. Алгоритм принимает только дискретные или дискретизированные атрибуты модели, а также исходит из предположения о стохастической независимости входных атрибутов ("наивность" алгоритма как раз обусловлена этими предположениями). Алгоритм Microsoft Na?ve Bayes представляет собой простую модель для анализа данных, которую можно рассматривать как средство первоначального разведочного анализа. Вследствие того, что большинство требуемой для построения модели информации вычисляется в процессе обработки соответствующего куба, результаты алгоритма Байеса возвращаются очень быстро. Это служит еще одним преимуществом использования алгоритма в качестве оптимального разведочного механизма для классификации и прогнозирования.

### Временные ряды - Microsoft Time Series

Алгоритм Microsoft Time Series создает модели, предназначенные для прогнозирования значений непрерывных переменных по времени и использованием как OLAP, так и реляционных источников данных. Например, можно использовать этот алгоритм для прогнозирования объема продаж и прибыли по историческим данных в разрезе регионов продаж.

Алгоритм позволяет осуществлять прогнозирование по нескольким непрерывным переменным. Временной ряд должен характеризоваться временной характеристикой по которой осуществляется прогноз, а также может содержать произвольный набор других характеристик, разделяющий ряд на несколько (например, регион, продавца и т.д.).

Алгоритм Microsoft Time Series может анализировать и использовать корреляции между различными прогнозируемыми характеристиками. В этом случае результат прогнозирования будет зависеть не только от исторических значений той же переменной, но и значений других прогнозируемых переменных. Например, продажи в определенном магазине могут зависеть от прошлых продаж в другом магазине.

Алгоритм Microsoft Time Series является разновидностью алгоритмов ART (Autoregressive Trees - деревья авторегрессии). В процессе реализации алгоритма строится дерево решений, листьям которого соответствуют функции линейной регрессии.

### Ассоциативные правила - Microsoft Association

Алгоритм Microsoft Association предназначен в первую очередь для анализа покупательских корзин. Анализируется каждая пара атрибут=значение (например продукт=велосипед) как значение некой логической переменной. Алгоритм сканирует транзакции в базе данных для определения "частых" наборов таких пар. Набор считается "частым" если его support (поддержка, т.е. число транзакций, в которые входит данный набор) превышает определенный порог. Например, частым набором может быть {Пол="мужской", Семейный статус = "женат", Возраст="30-35"}. Каждый набор имеет размер, например, в этом случае, он составляет 3.

Часто транзакции описываются вложенными таблицами, например, строками заказа продажи в заказе. В этом случае ключ вложенной таблицы служит названием атрибута, а факт его наличия - значением. Например {Товар("Велосипед")=existing, Товар("Велосипедная кепка")=existing}.

Алгоритм Microsoft Association помимо выявления частых набор предназначен для выявления правил вида A, B=>C, характеризуемых вероятностью выполнения (confidence), где {A, B}, {C} частые наборы. Символ '=>' означает, что появление набора C в транзакции можно предсказать из факта появления наборов A и B. Например, {Фотокамера=exisiting, Батарейки=existing}=>{Фотопленка=existing}. Порог значения вероятности соблюдения правила является параметром модели, определяющим факт рассмотрения каждого правила.

Ассоциативные правила используются для оптимизации кросс-продаж, целевых рекламных кампаний, определения политики скидок, в оптимизации размещения товаров на товарных полках в супермаркетах и т.д.

### Кластеризация последовательностей действий - Microsoft Sequence Clustering

Алгоритм Microsoft Sequence Clustering анализирует последовательности каких-либо фактов, представляющих собой временные последовательности дискретных переменных. Обычно такие последовательности атрибутов рассматриваются как происхождение событий в определенном порядке (например, порядок просмотра страниц сайта). Алгоритм предназначен для прогнозирования наступления последующих событий на основании уже осуществленного перехода между состояниями.

Алгоритм Microsoft Sequence Clustering является гибридом алгоритма последовательностей действий и алгоритмом кластеризации. В процессе реализации алгоритма выполняется группировка последовательности переходов в типичные кластера, характеризующиеся паттернами переходов. Каждый из этих кластеров анализируется с точки зрения распределения вероятности переходов отдельно. Типичным сценарием использования для этого алгоритма является задача анализа клиентов web-портала. У web-портала имеется набор связанных доменов, таких как новости, погода, финансы, почта, спорт и т.д. Каждый web-посетитель характеризуется последовательностью "кликов" - переходов в рамках каждого из этих доменов. Алгоритм Microsoft Sequence Clustering может сгруппировать этих web-посетителей в более-менее однородные группы на основании их паттернов навигации по порталу. Эти группы могут быть визуализированы, обеспечивая представление шаблонов использования портала каждой группой посетителей.

В алгоритме используется алгоритм Expectation Maximization для выявления кластеров, а также Марковские процессы первого порядка для моделирования переходов между состояниями (событиями).

### Нейронные сети - Microsoft Neural Network

Алгоритм Microsoft Neural Network предназначен для создания моделей классификации и регрессии путем конструирования многослойной нейронной сети перцептронов. Как и в случае алгоритма дерева решений, для каждого состояния выходного атрибута алгоритм вычисляет распределение вероятности входных атрибутов. Обрабатывается полный набор записей, при этом итеративно сравниваются предсказанные значения классификатора с известным значением. Ошибки классификации на первой итерации подаются на вход сети для изменения параметров на следующей итерации и т.д. Впоследствии, полученные вероятности используются для прогнозирования значения выходного атрибута на основании значений входных атрибутов. Одно из самых значительных отличий между этим алгоритмом и алгоритмом дерева решений состоит в том, что процесс обучения заключается в оптимизации параметров сети для минимизации ошибки классификации, в то время как алгоритм дерева решений осуществляет разбиение узлов для максимизации информации (минимизации энтропии). Алгоритм поддерживает как непрерывные так и дискретные типы атрибутов.

Линейная регрессия - Microsoft Linear Regression

Алгоритм Microsoft Linear Regression представляет собой алгоритм регрессии являющийся частным случаем алгоритма Microsoft Decision Trees, получаемый в случае запрета на разбиение узлов в дереве решений. Формула регрессии определена на всем объеме данных, т.е. на корневом узле дерева. Алгоритм предназначен для прогнозирования непрерывных атрибутов.

Логистическая регрессия - Microsoft Logistic Regression

Алгоритм Microsoft Logistic Regression представляет собой алгоритм регрессии являющийся частным случаем алгоритма Microsoft Neural Network, получаемый в случае удаления скрытого слоя нейросети. Алгоритм поддерживает прогнозирование значений как непрерывных, так и дискретных атрибутов.

#  Описание работы с курсом

**Оглавление**

В ходе данного курса вы будете работать в среде разработки Business Intelligence Development Studio (рис.1).



рис. 1 Business Intelligence Studio

Крус разделено на три части: подготовка базы данных SQL Server, подготовка базы данных Analysis Services и построение и работа с моделями Data Mining.

### Подготовка базы данных SQL Server

База данных AdventureWorksDW, которая используется в этом обзорном курсе, устанавливается вместе с SQL Server (по умолчанию эта опция отключена) и уже содержит представления (views), которые будут использованы для создания моделей.

В дополнения к ним необходимо будет импортировать в базу данных таблицу prospective customers (в виде плоского файла). Один из сценариев, который мы рассмотрим, анализирует целевые рассылки с использованием построителя запросов для прогнозов.

#### Импорт таблицы Prospective Customers

Вы будете использовать таблицу prospective customer в сценарии целевой рассылкидля проверки качества прогнозирования модели. Вы можете скачать эту таблицу с сайта Betaplace (www.Betaplace.com). Сначала необходимо импортировать плоский файл в базу данных AdventureWorksDW и назвать колонки этой таблицы так же как и в представлении, служащем источником данных для целевой рассылки.

В приложении А содержится сценарий импорта данных и переименования колонок. Загрузите плоский файл на ваш локальный диск. Запустите SQL Management Studio и откройте окно нового запроса к серверу БД AdventureWorksDW. Скопируйте SQL-запросы из приложения A в окно запросов. Измените значение переменной @data\_path так, чтобы она содержала путь к файлу "Prospect.csv". Выполните запрос.

Будет создана таблица **Prospects**, которая будет содержать ту же структуру столбцов, как и представление vTargetedmail , за исключением следующих столбцов:

* Age
* Bike Buyer
* Region

Так же, ключи клиентов (CustomerKey, CustomerAlternateKey) заменены на ключи перспективных клиентов (ProspectKey, ProspectAlternateKey).

### Подготовка базы данных Analysis Services

Прежде чем вы начнете создавать и работать с моделями Data Mining, вы должны выполнить следующие действия:

1. Создать новый проект Analysis Services.
	1. Создать новый источник данных (data source).
	2. Создать новое представление данных (data source view).

#### Создание проекта Analysis Services

Каждый проект Analysis Services определяет схему объектов для единственной базы данных Analysis Services, которая описывается моделями Data Mining, OLAP-кубами и дополнительными объектами.

1. Откройте Business Intelligence Development Studio.
2. Выберете **New** и **Project** из меню **File**.
3. Выберете проект Analysis Services в качестве типа нового проекта и назовите его **AdventureWorks**.
4. Нажмите **Ok**.

Новый проект откроется в Business Intelligence Development Studio.

#### Создание источника данных

Источник данных - это данные о соединении с внешним источником данных, которые хранятся в проекте и в базе данных Analysis Services. В источнике данных хранится информация об имени сервера, базы данных, где хранится информация, используемая в проекте, а так же другие свойства соединения.

1. Правой кнопкой мыши щелкните по узлу **Data Source** вашего решения и выберете **New Data Source**.
2. На первой странице мастера выберете **Next**.
3. Щелкните **New** чтобы добавить соединение к базе данных **AdventureWorksDW**.
4. В диалоговом окне **Connection Manager** выберете провайдер данных **Microsoft OLE DB Provider for SQL Server**, имя сервера - AdventureWorksDW, введите данные о пользователе, под учетной записью которого будет производиться соединение.
5. Щелкните **OK**.
6. Щелкните **Next**.

По умолчанию источник данных будет назван **Adventure Works DW**. Щелкните **Finish**

Новый источник данных **Adventure Works DW** появится в папке источников данных в дереве решения.

#### Создание представления данных

Представление данных (Data Source View) обеспечивает уровень абстракции для источника данных, позволяя вам менять структуру исходных данных для удобства вашего проекта. Используя представление данных, вы можете выбрать только относящиеся к вашему проекту таблицы и представления, определить связи между таблицами, добавить вычисляемые поля, а также именованные запросы без необходимости вносить модификации в исходные данные.

1. В дереве решения щелкните правой кнопкой мыши на **Data Source View** и выберете **New Data Source View**.
2. На первой странице щелкните **Next**.
3. Выберете источник данных **Adventure Works DW**, который вы создали на прошлом шаге в окне **Relational data sources**. Нажмите **Next**.
4. Если вы хотите создать новый источник данных, нажмите **New Data Source**.
5. Выберете таблицы из списка нажмите на правую стрелку для включения их в представление данных:
	* Prospect
	* vAssocSeqLineItems
	* vAssocSeqOrders
	* vTargetMail
	* vTimeSeries
6. Нажмите Next.
7. По умолчанию представление данных будет названо Adventure Works DW. Нажмите **Finish**.

Откроется окно представления данных как показано на рис. 2. Далее можно производить изменения в представлении источника данных.



Рис. 2 Окно просмотра модели базы Adventure Works DW

#### Редактирование представления источника данных

Используя редактор представления источника данных (Data Source View Editor), вы можете вносить изменения в представление данных. Например, вы можете переименовать объект, если его новое название будет больше подходить к модели. При этом оригинальное название не изменится, но появится возможность обращаться к объекту через новое более удобное имя.

Для создания сценариев анализа рыночной корзины или кластеризации последовательностей необходимо создать новое соединение многие-к-одному между vAssocSeqOrders и vAssocSeqLineItems. Используя эту связь можно сделать vAssocSeqLineItems вложенной таблицей по отношению к vAssocSeqOrders, что необходимо для создания модели.

1. В окне просмотра выбрать поле OrderNumber из таблицы vAssocSeqLineItems
2. Перенести выбранную колонку в таблицу vAssocSeqOrders и поместить ее на колонку OrderNumber

Новая связь многие-к-одному между vAssocSeqOrders и vAssocSeqLineItems создадна.

#  Построение и работа с моделями Data Mining

Редактор Data Mining (показан на рисунке 4) содержит все инструменты и средства отображения для построения и работы с моделями Data Mining. Дополнительная информация по редактору доступна в разделе "Using the Data Mining Tools" в SQL Server Books Online.



Рисунок 4 Редактор Data Mining

На протяжении этого руководства вы будете работать со следующими сценариями:

* Целевая отправка писем (Targeted mailing)
* Прогнозирование (Forecasting)
* Рыночная корзина (Market basket)
* Кластеризация последовательностей (Sequence clustering)

В сценарии Целевая отправка писем вы будете заниматься построением моделей, сравнением их ожидаемых возможностей (при помощи окна Mining Accuracy Chart), а также созданием прогнозов с использованием Prediction Query Builder.

# Адресная рассылка

Отдел маркетинга компании Adventure Works заинтересован в увеличении продаж при помощи проведения почтовой кампании, нацеленной на определённых покупателей. Исследуя их характеристики, они хотят обнаружить некоторый шаблон, применимый к потенциальным клиентам, который мог бы впоследствии использоваться для определения наиболее вероятных покупателей.

Кроме того, отдел маркетинга намеревается выявить логические группы среди уже существующих в их базе данных клиентов. Например, группа может содержать покупателей, объединённых по демографическому признаку, обладающих схожим набором покупок.

Adventure Works располагает списком как бывших, так и потенциальных клиентов.

В процессе решения этой задачи, отделу маркетинга потребуется выполнить следующие действия:

* Установить модели, способные выявить наиболее вероятных клиентов из списка потенциальных покупателей
* Кластеризовать своих существующих клиентов

Для выполнения этого сценария вам потребуется использовать Microsoft Na?ve Bayes, Microsoft Decision Trees, и Microsoft Clustering алгоритмы. Сценарий содержит в себе 5 задач:

# Описание структуры модели Data Mining

Первый шаг заключается в создании новой структуры Data mining модели при помощи Mining Model Wizard. При этом также создаётся начальная модель, основанная на алгоритме Microsoft Decision Trees.

Для того чтобы создать структуру модели:

1. В Solution Explorer кликните правой кнопкой мыши на **Mining Models**, выберите **New Mining Model**. Откроется Mining Model Wizard.
2. Нажмите **Next** на странице приветствия.
3. Выберите **From existing relational database or data warehouse**, затем **Next**.
4. В разделе **Which data mining technique do you want to use**? выберите **Microsoft Decision Trees**. Вы создадите несколько моделей, опирающихся на эту начальную структуру, в основе которой лежит алгоритм Microsoft Decision Trees.
5. Нажмите **Next**. По умолчанию в качестве источника данных здесь выбран Adventure Works DW.
6. Нажмите **Next**.
7. Пометьте флагом **Case** таблицу **vTargetMail**, затем нажмите **Next**.
8. Пометьте флагом **Key** столбец **CustomerKey**. Если в исходной таблице присутствует ключевое поле, Mining Model Wizard автоматически выберет его в качестве ключевого для модели.
9. Пометьте флагом **Input** и **Predictable** столбец **BikeBuyer**. В результате столбец становится доступным для прогнозирования в новых наборах данных. После того как вы помечаете столбец подобным образом, становится доступной кнопка Suggest. Нажатие на эту кнопку вызывает диалоговое окно со списком наиболее связанных с данным полем столбцов, расположенных в порядке корреляции с прогнозируемым. Столбцы со значением больше 0.05 автоматически добавляются в модель. Если вы согласны с предлагаемым вариантом, нажмите кнопку **OK** или определите нужные поля вручную.
10. Пометьте флагом **Input** столбцы, перечисленные в таблице.

|  |  |  |
| --- | --- | --- |
| **Age** | **YearlyIncome** | **Region** |
| CommuteDistance | HouseOwnerFlag | TotalChildren |
| EnglishEducation | LastName |  |
| EnglishOccupation | MaritalStatus |  |
| FirstName | NumberCarsOwned |  |
| Gender | NumberChildrenAtHome |  |

Вы можете выбрать несколько столбцов, удерживая клавишу SHIFT.

1. Нажмите Next.
2. В поле Specify Columns' Content and Data Type нажмите Detect. При этом автоматически определяется, содержат ли столбцы с численными данными дискретные, либо непрерывные значения. Например, столбец может содержать информацию о заработной плате с непосредственными значениями сумм (непрерывные значения) или в нём могут содержаться целые числа, отвечающие неким диапазонам значений (1 = <$25,000; 2 = от $25,000 до $50000 и так далее), то есть дискретные значения.
3. В поле Structure Name введите Targeted Mailing и нажмите Finish
4. Нажмите Finish. При этом откроется редактор Data mining, отображающий структуру Targeted Mailing, которую вы только что создали, см. Рисунок 5.



Рисунок 5 Страница структуры Targeted Mailing

# Редактирование Data Mining моделей

Начальная Data Mining структура содержит единственную модель, основанную на Microsoft Decision Trees. В этой части вы, при помощи страницы Mining Models редактора определите две дополнительные модели: модели Microsoft Na?ve Bayes и Microsoft Clustering.

Чтобы создать модель Microsoft Clustering

1. Нажмите на закладку **Mining Models**.
2. Кликните правой кнопкой по Targeted Mailing и выберите **New Mining Model**.
3. В поле **Model Name** введите *TM\_Clustering*.
4. В поле **Alogithm Name** выберите **Microsoft Clustering**.
5. Нажмите OK.

Новая модель появится на странице Mining Models. Модель Microsoft Clustering позволяет группировать и прогнозировать непрерывные и дискретные атрибуты. Вы можете изменить используемые столбцы и свойства для новой модели.

Установка столбца как Predict (прогнозируемого) не оказывает никакого эффекта на этапе обучения модели; но позволяет вам указать этот столбец в запросе на прогнозирование PREDICTION JOIN. При создании кластеров алгоритм игнорирует столбцы, помеченные как **PredictOnly**. Статистика для **PredictOnly** столбцов в кластерной модели определяется на финальном шаге, после того как операция кластеризации завершена. Это имеет смысл, если вы хотите видеть распределение атрибута по кластерам, созданным на основании других атрибутов и позволяет выявить более глубокие взаимосвязи.

Чтобы создать модель Microsoft Na?ve Bayes

1. Кликните правой кнопкой по Targeted Mailing и выберите New Mining Model.
2. В поле Model Name введите TM\_NaiveBayes.
3. В поле Algorithm Name выберите Microsoft Na?ve Bayes. После чего появится диалоговое окно с текстом, объясняющим, что алгоритм Microsoft Na?ve Bayes не поддерживает работу с столбцами Age, Geography Key и Yearly Income, содержащими непрерывные значения, и они будут проигнорированы.
4. Нажмите **Yes**.
5. Нажмите **OK**.
6. После чего, новая модель появится на странице Mining Models. Хотя вы можете менять используемые столбцы и свойства для всех моделей на этой странице, в данном случае оставьте всё без изменений.

**Обработка Data Mining моделей**

Теперь, после того как структура и параметры моделей определены, вы можете осуществить развёртывание и обработку моделей.

Для того чтобы развернуть проект и обработать модели

* Нажмите F5.
Теперь база данных Analysis Services выложена на сервер и модели обработаны. Если база данных уже была на сервере, вы можете осуществить только обработку моделей, следуя следующему алгоритму.

Для обработки data mining моделей

* В меню Mining Model выберите пункт Process All Models.Откроется диалоговое окно Process Targeted Mailing.
* Нажмите Run.
* Откроется диалоговое окно Process Progress, отображающее информацию об обработке модели. В зависимости от производительности вашего компьютера, процесс может занять некоторое время.
* После того как завершится обработка, закройте оба диалоговых окна.

[Читать 2 часть](http://www.interface.ru/home.asp?artId=21206)

**ССЫЛКИ ПО ТЕМЕ**

* Подписаться на рассылку "Вопросы и ответы по MS SQL Server "
* [Обратиться в "Интерфейс" за дополнительной информацией/по вопросу приобретения продуктов](http://www.interface.ru/iservices/mailfrompage.asp?To=mail@interface.ru&final=close&purl=%2Fhome%2Easp%3FartId%3D21179&pname=%CA%F3%F0%F1+%EF%EE+%E8%ED%F2%E5%EB%EB%E5%EA%F2%F3%E0%EB%FC%ED%EE%EC%F3+%E0%ED%E0%EB%E8%E7%F3+%E4%E0%ED%ED%FB%F5+%28Data+Mining%29+%F1+%E8%F1%EF%EE%EB%FC%E7%EE%E2%E0%ED%E8%E5%EC+%EF%EB%E0%F2%F4%EE%F0%EC%FB+MS+SQL+2005+%2E+%D7%E0c%F2%FC+1)

# Курс по интеллектуальному анализу данных (Data Mining) с использованием платформы MS SQL 2005 . Чаcть 2

Источник: [businessdataanalytics](http://www.interface.ru/iservices/goto.asp?Url=enc:rn%7Ej%40%255X%7Fisdoiy%5Eknk%5Bt%5Bvs%7Ecii4h%7F%25N%5B%7E%5BWctct%5DYKVIoh%7C%5Fx%28%3A%2A%3F%27%3F%294b%7Eg)

# Исследование моделей

После того как модели обработаны, вы можете просмотреть их, используя закладку Mining Model Viewer в редакторе Data Mining. При помощи выпадающего списка **Mining Model** в верхней части закладки можно исследовать модели, входящие в структуру.

###### Модель Microsoft Decision Trees

Mining Model Viewer по умолчанию открывает модель Targeted Mailing первой в структуре. Tree viewer содержит две закладки, **Decision Tree** (дерево решений) и **Dependency Network** (сеть зависимостей).

###### Decision Tree (Дерево решений)

На странице **Decision Tree** вы можете просмотреть все модели деревьев, составляющие модель Targeted\_Mailing. Для каждого прогнозируемого атрибута модели существует одно дерево, если только не задействовано feature selection. Так как ваша модель содержит всего один прогнозируемый атрибут, Bike Buyer, то и дерево тоже одно. Если бы их было несколько, можно было бы выбрать нужное при помощи Tree box.

Tree viewer по умолчанию показывает только первые три уровня дерева. Чтобы увидеть больше используйте Show Level slider или Default Expansion box. Для более подробной информации по настройке Tree viewer, обратитесь к разделу "Viewing with Tree Viewer" в SQL Server Books Online.

Чтобы создать дерево, показанное на рисунке 6

1. Установите Show Level 5
2. В списке Background нажмите 1.
Здесь задаётся состояние прогнозируемого атрибута. Изменяя это свойство, вы можете быстро узнать число случаев попадания покупателя велосипеда в каждый узел. Чем темнее узел, тем больше случаев попадания он содержит.



Рисунок 6 Страница Decision Tree для модели Targeted Mailing

Каждый узел в дереве решений несёт в себе следующую информацию:

* Условие, по которому попадаем в этот узел из предыдущего. Полный путь к узлу можно узнать из легенды или из подсказки.
* Гистограмма, описывающая распределение состояний для прогнозируемого столбца. Число состояний, отображаемых на гистограмме, регулируется при помощи Histogram control.
* Концетрацию случаев попадания, если состояние предсказываемого значения определено в контроле Background.

Если включен режим drillthrough ("проваливание" в исходные данные), вы можете просмотреть варианты для узла, кликнув по нему правой кнопкой мыши и выбрав Drillthrough.

###### Сеть зависимостей (Dependency Network)

На странице Dependency Network показаны отношения между атрибутами, влияющими на прогнозирующую способность модели. Сеть зависимостей для модели Targeted Mailing представлена на рисунке 7.



Рисунок 7 Страница Dependency Network модели Targeted Mailing.

Центральный узел на рисунке 7, Bike Buyer, представляет прогнозируемый атрибут модели. Каждый из узлов вокруг него отображает атрибут, влияющий на прогнозируемый. При помощи ползунка на левой стороне страницы можно указать силу показываемых связей. При перемещении ползунка вниз будут показываться только сильные связи.

По легенде цветов внизу окна можно узнать, какой узел является прогнозируемым и какие узлы его определяют.

###### Модель Microsoft Clustering

Используйте combo box Mining Model для перехода к модели TM\_Clustering. Окно этой модели, Cluster viewer, содержит четыре закладки: Cluster Diagram, Cluster Profiles, Cluster Characteristics, и Cluster Discrimination. По умолчанию при первом открытии отображается страница Cluster Diagram.

Для более подробной информации по использованию Cluster viewer, обратитесь к разделу " Viewing with Cluster Viewer " в SQL Server Books Online.

###### Кластерная диаграмма (Cluster Diagram)

На странице Cluster Diagram вы можете изучить отношения между кластерами, обнаруженными алгоритмом. Длина линий, соединяющих кластеры, отображает "близость", а их интенсивность показывает, насколько они схожи. Цвет кластера отвечает за частоту появления переменной и состояния (выбранные в полях Shading Variable и State вверху экрана) в каждом кластере. Переменная по умолчанию population, но вы можете поменять её на любую другую из модели и найти нужные вам кластеры. Используя полосу прокрутки в левой части экрана, вы можете отфильтровать слабые связи и найти кластеры, расположенные наиболее близко к друг другу.

Например, установите поле Bike Buyer в качестве Shading Variable и State в 1. Как показано на рисунке 8, кластер №5 содержит наибольшее число покупателей велосипедов. А самая сильная связь - между 4 и 7 кластерами.



Рисунок 8 Кластерная диаграмма модели TM\_Clustering.

###### Профили кластеров (Cluster Profiles)

Страница Cluster Profiles предоставляет общий вид модели TM\_Clustering. Как показано на рисунке 9, каждому кластеру модели на этой странице соответствует столбец с описывающими его данными. В первом столбце находится список атрибутов, с которыми связан по крайней мере один кластер. Все оставшиеся столбцы отображают распределение состояний этих атрибутов для каждого кластера. Распределение дискретной переменной показывается цветным столбцом, максимальное число отображаемых цветов-состояний определяется в поле Bars per histogram. Непрерывные значения отображаются diamond диаграммой, показывающей среднее и стандартное отклонение в каждом кластере.



Рисунок 9 Страница Cluster Profiles модели TM\_Clustering.

###### Характеристики кластеров (Cluster Characteristics)

Cтраница Cluster Characteristics позволяет более детально рассмотреть характеристики кластера. Для примера, на рисунке 10, вы можете видеть, что люди из Кластера №5 (покупатели велосипедов) обладают следующими характеристиками: они передвигаются на короткие дистанции (0-1 миля), у них нет машины и они женаты.



Рисунок 10 Окно характеристик кластера модели TM\_Clustering.

###### Cluster Discrimination

На странице Cluster Discrimination вы можете посмотреть характеристики, отличающие один кластер от другого. После выбора кластеров (поля First cluster и Second cluster) система определяет различия и показывает их, сортируя по атрибутам, наиболее сильно отличающим кластеры.

На рисунке 11 приведено сравнение Кластера №5 и Кластера №7 модели TM\_Clustering. В кластере №5 наиболее высокая плотность покупателей велосипедов, а в №10, напротив, самая низкая. Видно, к примеру, что люди, попавшие в Кластер №10 скорее всего из Северной Америки и моложе (23-31), чем жители Европы из Кластера №5.



Рисунок 11 Страница Cluster Discrimination модели TM\_Clustering.

###### Модель Microsoft Na?ve Bayes

Используйте выпадающий список Mining Model для перехода к модели TM\_NaiveBayes. Окно этой модели, Na?ve Bayes viewer, содержит четыре закладки: Dependency Network, Attribute Profiles, Attribute Characteristics, и Attribute Discrimination.

Для более подробной информации по использованию Na?ve Bayes viewer, обратитесь к разделу " Viewing with Na?ve Bayes viewer " в SQL Server Books Online.

###### Сеть зависимостей (Dependency Network)

Страница Dependency Network в Na?ve Bayes viewer организована также как одноимённая страница в Tree viewer. Каждый узел представляет атрибут, а линии, связывающие узлы соответствуют отношениям. На рисунке 12 показаны все атрибуты, влияющие на состояние прогнозируемого атрибута, Bike Buyer.



Рисунок 12 Страница Dependency Network для модели TM\_Na?veBayes.

По мере передвижении ползунка вниз, остаются атрибуты, наиболее сильно влияющие столбец Bike Buyer (являющийся прогнозируемым). Действуя таким образом, вы обнаружите, что самым сильным фактором, определяющим, купит ли человек велосипед, является число принадлежащих ему машин.

###### Профили атрибута (Attribute Profiles)

Страница Attribute Profiles описывает как различные состояния входных атрибутов влияют на значение прогнозируемого атрибута.

В поле Predictable выберите Bike Buyer. Атрибуты, влияющие на состояние прогнозируемого атрибута, перечислены вместе со значениями каждого состояния входных атрибутов и их распределениями для каждого состояния прогнозируемого атрибута.

В поле Predictable выберите Bike Buyer. Атрибуты, влияющие на состояние прогнозируемого атрибута, перечислены для всех значений входных параметров совместно с их распределением для каждого состояния прогнозируемого атрибута.

На рисунке 13 показана страница Attribute Profiles для столбца Bike Buyer для модели TM\_NaiveBayes model.



Рисунок 13 Страница Attribute Profiles для модели TM\_Na?veBayes.

###### Характеристики атрибута (Attribute Characteristics)

Используя страницу Attribute Characteristics, вы можете выбрать атрибут и его значение, чтобы увидеть, значения каких атрибутов меняются при этом в большей или меньшей степени.

В поле атрибут выберите Bike Buyer, в поле значение - 1.

Например, из рисунка 14 видно, что люди, не имеющие детей, покупают наибольшее количество велосипедов.



Рисунок 14 Страница Attribute Characteristics для модели TM\_Na?veBayes.

###### Attribute Discrimination

На странице Attribute Discrimination вы можете исследовать отношение между двумя дискретными значениями выбранного прогнозируемого атрибута и значениями других атрибутов.

Поскольку модель TM\_NaiveBayes имеет всего два состояния, 1 и 0, вам не придётся совершать каких-либо дополнительных действий по настройке окна.

К примеру, из рисунка 15 видно, что те, у кого нет машины более склонны к покупке велосипеда, а люди, имеющие две машины скорее не будут этого делать.



Рисунок 15 Страница Attribute Discrimination для модели TM\_NaiveBayes.

 [Читать 3 часть](http://www.interface.ru/home.asp?artId=21271)

# Курс по интеллектуальному анализу данных (Data Mining) с использованием платформы MS SQL 2005 . Чаcть 3

Источник: [businessdataanalytics](http://www.interface.ru/iservices/goto.asp?Url=enc:rn%7Ej%40%255X%7Fisdoiy%5Eknk%5Bt%5Bvs%7Ecii4h%7F%25N%5B%7E%5BWctct%5DYKVIoh%7C%5Fx%28%3A%2A%3F%27%3F%2E4b%7Eg)

# Тестирование точности Data Mining-моделей

Теперь модели обработаны и исследованы. Но насколько хорошо они предсказывают ситуацию? Работает ли какая-нибудь из моделей лучше, чем другие?

Используя страницу Mining Accuracy Chart, вы можете вычислить, как хорошо каждая из моделей предсказывает результат и сравнить их между собой. Этот метод сравнения также иногда называется диаграммой роста (lift chart). На странице The Mining Accuracy Chart использует данные, отделённые от первоисточника, что позволяет сравнивать прогнозы с известными результатами. Затем результаты сортируются и отображаются на графике вместе с идеальной моделью, показывая, насколько хорошо каждая модель делает прогноз. График идеальной модели соответствует теоретической модели, предсказывающей результат со 100% точностью.

Диаграмма роста является важным инструментом, поскольку она помогает различать модели практически идентичные по структуре, показывая, какая делает лучший прогноз. Подобным образом, из её рассмотрения можно сделать вывод о том, какие типы алгоритмов делают лучшие предсказания в данной ситуации. Для более подробной информации по использованию Mining Accuracy Chart, обратитесь к разделу "Comparing Data Mining Models with the Lift Chart" в SQL Server Books Online.

Страница Mining Accuracy Chart показана на рисунке 16.



Рисунок 16 Страница Mining Accuracy Chart.

Для того чтобы создать новую диаграмму выполните следующие шаги:

1. Свяжите столбцы модели с соответствующими столбцами во входных данных.
2. Установите фильтр на входные данные.
3. Выберите сравниваемые модели, а также поля для которых делается прогнозирование и их значения.

**Примечание**: Для использования mining accuracy chart модели должны быть обработаны.

###### Отображение входных столбцов

На первом шаге необходимо связать столбцы, используемые в модели со столбцами во входных данных. В том случае, если имена столбцов совпадают, связи создадутся автоматически.

Чтобы соотнести входные столбцы со структурой модели:

1. В поле Select Input Table(s) выберите Select case table.
Откроется диалоговое окно выбора таблицы, где вы можете выбрать таблицу, содержащую данные, которые вы хотите использовать для прогнозирования и определения точности моделей.
В этом руководстве используются те же данные, что и для обработки моделей.
2. В диалоговом окне Select table, раскройте базу AdventureWorks и выберите источник данных Adventure Works DW.
3. Нажмите vTargetMail.
Столбцы data mining структуры автоматически будут соответствовать столбцам с теми же именами таблицы с исходными данными, как показано на рисунке 17.



Рисунок 17 Связанные столбцы.

Запрос на прогноз генерируется для каждой модели Mining Structure на основании связанных столбцов. Вы можете удалить связь, выбрав линию, связывающую поля в Mining Structure и Select Input Table(s) и нажав DELETE. Также вы можете создать связь вручную, выбрав столбец в Select Input Table(s) и перетащив его на соответствующий столбец в Mining Structure.

###### Фильтрация входных столбцов

Для установки фильтров на входные данные вы можете использовать таблицу Filter the input data used to generate the lift chart. Вы можете перетащить столбцы из Select Input Table(s) на таблицу с результатами или выбрать значения в выпадающих списках. Например, если необходимо учитывать только те строки в столбце Income, где его значение больше х, выберите Income в поле Field, а в поле Criteria/Argument введите >x.

Для выполнения данного упражнения вам не потребуется фильтровать данные.

###### Выбор моделей, прогнозируемых полей и их значений

Следующий шаг заключается в выборе моделей, которые вы хотите включить в рассмотрение, и прогнозируемого поля, по которому они будут сравниваться. По умолчанию, выбраны все модели.

Вы можете создавать два типа диаграмм. Если указано значение прогнозируемого атрибута, вы увидите график вроде того, что показан на рисунке 18, из которого видно какая из моделей показывает лучший результат (lift chart). В том случае, если значение не указано, то диаграмма отобразит точность модели (см. рисунок 19).

Чтобы показать диаграмму lift chart:

1. Для каждой модели в поле **Predictable Column Name** выберите **Bike Buyer**.
2. Для каждой модели в поле **Predict Value** выберите 1.

Чтобы показать диаграмму точности моделей:

* В поле Predictable Column Name, выберите Bike Buyer.
Оставьте поле Predict Value пустым.

Если установлен флаг Synchronize Prediction Columns and Values, прогнозируемый столбец синхронизируется для каждой модели в структуре Data Mining.

**Примечание**: Список столбцов модели Predictable Column Name, ограничен полями с типом Predict и Predict Only, содержащими дискретные или дискретизированные данные.

В некоторых более сложных случаях, вам может понадобиться создать диаграмму роста для двух моделей анализа, но в которых значения прогнозируемого атрибута берутся из разных столбцов структуры Data Mining. Если выключить флаг Synchronize Prediction Columns and Values, можно выбрать любой доступный прогнозируемый столбец и значение. Результаты отображаются вместе, вне зависимости от того как они получены.

###### Диаграмма роста

Кликните на закладку **Lift Chart** чтобы просмотреть диаграмму роста. При выполнении этого действия на сервере выполняется запрос на прогнозирование. Полученные результаты прогноза сравниваются с известными значениями и сортируются по вероятности, после чего отображаются на графике. Для более подробной информации по использованию диаграммы, обратитесь к разделу "Lift Chart" в SQL Server Books Online.

При условии указанного значения прогнозируемого столбца, диаграмма роста имеет вид показанный на рисунке 18.



Рисунок 18 Рост для каждой модели в сравнении с идеальной моделью.

Если значение прогнозируемого атрибута не было указано, диаграмма роста показывает точность предсказаний моделей (см. рисунок 19).



Рисунок 19 Точность каждой модели в сравнении с идеальной моделью.

[Читать 4 часть](http://www.interface.ru/home.asp?artId=21296)

# Курс по интеллектуальному анализу данных (Data Mining) с использованием платформы MS SQL 2005 . Чаcть 4

Источник: [businessdataanalytics](http://www.interface.ru/iservices/goto.asp?Url=enc:rn%7Ej%40%255X%7Fisdoiy%5Eknk%5Bt%5Bvs%7Ecii4h%7F%25N%5B%7E%5BWctct%5DYKVIoh%7C%5Fx%28%3A%2A%3F%27%3F%2F4b%7Eg)

# Создание прогнозов

Теперь, после того как вы закончили с моделями анализа, можете перейти к созданию DMX запросов, используя Prediction Query Builder. Prediction Query Builder аналогичен Access Query Builder, где вы можете использовать drag-and-drop для построения запросов. Инструмент содержит три окна:

* Design
* Query
* Result



Рисунок 20 Prediction Query Builder - вид по умолчанию.

Используя окна Design и Query, вы можете создавать и просматривать запросы. Результаты выполнения отображаются на странице Result.

###### Создание запросов

Первым шагом при создании запроса является выбор модели анализа и таблицы с исходными данными.

Чтобы выбрать модель и таблицу с исходными данными

1. В окне Mining Model нажмите Select model.
Откроется диалоговое окно Select Mining Model. По умолчанию выбрана первая модель в структуре.
2. Переместитесь по дереву к пункту Targeted Mailing и нажмите на него.
3. В поле Select Input Table(s) нажмите Select case table.
Откроется диалоговое окно Select Table.
4. Спуститесь по дереву к таблице Prospect из data source AdventureWorksDW.

После того, как вы указали таблицу с исходными данными, Prediction Query Builder создаст связи по умолчанию между моделью и таблицей, основанные на совпадении имён столбцов, как показано на рисунке 21.



Рисунок 21 Соответствие столбцов на странице Mining Model Prediction.

Чтобы построить запрос на прогноз

1. В поле Source, нажмите на ячейку в первой пустой строке, затем кликните на Prospect table.
2. В поле Field, рядом с записью, созданной на первом шаге, кликните ProspectKey.
При этом добавится уникальный идентификатор к запросу, так что вы сможете определить кто склонен или наоборот, несклонен к покупке велосипеда.
3. Нажмите на следующей ячейке в столбце Source, затем кликните на Targeted Mailing.
4. В ячейке Field, выберите Bike Buyer.
Таким образом, определяется, что для создания прогнозов будет использована модель Microsoft Clustering из структуры Targeted Mailing.
5. Нажмите на следующей ячейке в столбце Source, затем нажмите Prediction Function.
6. Рядом с Prediction Function, в поле Field, нажмите PredictProbability.
Функции прогнозирования дают информацию о том, о том как делается прогноз. Функция PredictProbability предоставляет информацию о вероятности правильного прогноза. Вы можете определить параметры функции в столбце Criteria/Argument.
7. В поле Criteria/Argument, наберите [Targeted Mailing].[Bike Buyer].
Этим действием назначается целевой столбец для функции PredictProbability.

Для более подробной информации о функциях, обратитесь к разделу " DMX Function Reference " в SQL Server Books Online.

На вашем экране должно отобразиться следующее (см. рисунок 22).



Рисунок 22 Prediction Query Builder на странице Mining Model Prediction.

Нажав на иконку в левом верхнем углу, можно переключиться в Query view и просмотреть DMX код, созданный Prediction Query Builder. Кроме того, вы можете исполнить запрос, изменить и затем исполнить его, однако изменённый запрос не отобразится при переходе назад в Design View.

###### Просмотр результатов

Вы можете исполнить запрос, нажав на стрелку рядом с иконкой левом верхнем углу страницы, затем нажмите Result. На рисунке 23 показаны результаты запроса.



Рисунок 23 Страница результатов.

Столбцы ProspectKey, BikeBuyer и Expression отвечают соответственно идентификатору потенциальных покупателей, купят ли они велосипед (1 или 0) и вероятности правильности прогноза. Вы можете использовать эти результаты, чтобы определить, кому следует разослать письма с предложением.

Конец.