Лабораторна робота 7:

Использование инструментов Data Mining Client для Excel 2007 для создания модели интеллектуального анализа данных

Рассмотренные в лабораторных работах ["Надстройки интеллектуального анализа данных для MicrosoftOffice"](https://intuit.ru/studies/courses/2312/612/lecture/13263)-["Использование инструментов "Prediction Calculator" и "ShoppingbasketAnalysis""](https://intuit.ru/studies/courses/2312/612/lecture/13268)"Средства анализа таблиц для *Excel*" (**TableAnalysisTools**) для конечного пользователя во многом представляются "черным ящиком", выполняющим *анализ*,но не дающим информации о том, как получен результат. Если такое решение не устраивает, можно перейти с вкладки **Analyze** на вкладку **DataMining** и воспользоваться инструментами **DataMiningClient** для *Excel* ([рис. 14.1](https://intuit.ru/studies/courses/2312/612/lecture/13274?page=1#image.14.1)).



**Рис. 14.1.**Группа инструментов DataModeling

В ["Использование инструментов Data Mining Client для Excel 2007 для подготовки данных"](https://intuit.ru/studies/courses/2312/612/lecture/13273)мы рассмотрели инструменты, позволяющие подготовить данные для анализа. Следующая *группа* показанные на [рис. 14.1](https://intuit.ru/studies/courses/2312/612/lecture/13274?page=1#image.14.1) инструменты **DataModeling**,позволяющие создать модели интеллектуального анализа данных.

|  |  |
| --- | --- |
| **Классификация (Classify)** | создает модель классификации на основе существующих данных таблицы Excel, диапазона Excel или внешнего источника данных (**AnalysisServicesDataSource**). На основе обрабатываемых данных формируются шаблоны, которые при использовании позволяют отнести рассматриваемый пример к одному из возможных классов. По умолчанию используется алгоритм **DecisionTrees**, но также доступны **LogisticRegression**, **NaiveBayes**, **NeuralNetworks**. |
| **Оценка (Estimate)** | позволяет создать модель оценки значения целевого параметра (он должен быть числовым) на основе данных из таблицы или диапазона ячеек Excel либо внешнего источника данных. По умолчанию используется алгоритм **Decision Trees**, также доступны **Linear Regression**, **Logistic Regression**, **Neural Networks**. |
| **Кластер (Cluster)** | запускает мастер, позволяющий построить модель кластеризации на основе данных из таблицы или диапазона Excel, либо внешнего источника данных. Модель определяет группы строк со сходными характеристиками,для чего используется алгоритм **MicrosoftClustering**. Данная задача аналогична решаемой средством **DetectCategories** из набора **TableAnalysisTools**. |
| **Поиск взаимосвязей (Associate)** | помогает создать модель, описывающую взаимосвязь объектов (покупаемых товаров и т.д.), затрагиваемых одной транзакцией, для чего используется алгоритм **AssociationRules**. С подобной задачей мы сталкивались, используя инструмент **ShoppingBasketAnalysis** из **TableAnalysisTools**. Для построения модели анализа необходимо, чтобы исходные данные содержали столбец с идентификатором транзакций и были по нему отсортированы. В качестве источника данных может использоваться только таблица или диапазон ячеек Excel. |
| **Прогноз (Forecast)** | Данный мастер позволяет построить модель для прогнозирования новых значений в числовой последовательности, аналогично инструменту **Forecast в TableAnalysisTools**. Используется алгоритм **TimeSeries**, для работы которого требуется, чтобы столбец (или столбцы), в отношении которого будет выполняться прогноз, имели непрерывные числовые значения. Также может присутствовать столбец с отметкой времени (в этом случае, строки в таблице должны быть по нему отсортированы). |
| **Дополнительно (Advanced)** | позволяет создать структуру1 интеллектуального анализа данных или добавить в существующую структуру новую модель (например, для сравнения результатов, выдаваемых разными алгоритмами анализа). |

Используем инструмент **Classify**. В поставляющемся с надстройками наборе данных (*меню* **"Пуск"->"Надстройки интеллектуального анализа данных"->"Образцы данных Excel"**) выберем таблицу **TrainingData**,содержащую случайную выборку 70% данных из таблицы **SourceData**. Запустим мастер **Classify**, в первом окне которого будет комментарий *по* применению инструмента, а второе окно позволит указать *источник данных* для анализа (*таблица* **TrainingData**).Дальше потребуется описать цель анализа.

Пусть нас интересует, сделает ли данный клиент покупку. В целевом столбце указываем *параметр* **BikeBuyer** ([рис. 14.2](https://intuit.ru/studies/courses/2312/612/lecture/13274?page=1#image.14.2), окно слева), сбрасываем в перечне входных столбцов отметки напротив **ID** (порядковый номер клиента в базе никак не влияет на его решение о покупке). Если **ID** оставить среди анализируемых параметров, то итоговая модель может его учесть. В частности, на [рис. 14.3](https://intuit.ru/studies/courses/2312/612/lecture/13274?page=1#image.14.3) показано *дерево* решений, учитывающее *значение* поля **ID** в процессе классификации, что однозначно неправильно.



**Рис. 14.2.**Указание параметров для анализа



**Рис. 14.3.**Неудачный вариант дерева решений

Если требуется более точная настройка, можно открыть окно **Parameters** и явно указать используемый *алгоритм* и его параметры ([рис. 14.2](https://intuit.ru/studies/courses/2312/612/lecture/13274?page=1#image.14.2), окно справа). Далее мастер предложит разделить имеющиеся данные на набор для обучения модели и для ее тестирования ([рис. 14.4](https://intuit.ru/studies/courses/2312/612/lecture/13274?page=1#image.14.4)-1). *По* умолчанию на набор для тестирования выделяется 30 % строк исходного набора.

|  |  |
| --- | --- |
| 1 | 2 |
| Разбиение данных и указание названий модели и структуры**Рис. 14.4.**Разбиение данных и указание названий модели и структуры | https://intuit.ru/EDI/18_07_20_2/1595024415-12778/tutorial/1036/objects/14/files/23_04-2.jpg |

Последний этап работы мастера - указание имени создаваемой структуры и модели ([рис. 14.4](https://intuit.ru/studies/courses/2312/612/lecture/13274?page=1#image.14.4)-2). В нашем примере структура будет назваться **TrainingDataStructure**, а модель **Classify BikeBuyer\_1**. Эти названия нам понадобятся впоследствии для работы с моделью.

Если выполняющий *анализ* *пользователь* не имеет прав администратора в базе Аналитических Служб (эту настройку мы делали в ["Надстройки интеллектуального анализа данных для MicrosoftOffice"](https://intuit.ru/studies/courses/2312/612/lecture/13263)), то создать постоянною модель интеллектуального анализа на сервере он не сможет. В этом случае можно использовать временную модель, для чего отметить *пункт* **Use temporary model**.Временная модель будет автоматически удалена с сервера *по* *завершению сеанса работы* пользователя.

Отмеченная *по* умолчанию настройка **Brose model** указывает на то, что после создания модели будет открыто окно просмотра. Для модели, созданной с использованием алгоритма **DecisionTrees**,отображается построенное *дерево* решений и *диаграмма* зависимостей. Представленное на [рис. 14.5](https://intuit.ru/studies/courses/2312/612/lecture/13274?page=1#image.14.5) *дерево* решений позволяет оценить построенную модель. Расположенные в верней части экрана "ползунок" и выпадающий *список* позволяют установить число отображаемых уровней дерева (на рисунке показаны все пять). Если навести *указатель* мыши на точку ветвления, можно увидеть всплывающую подсказку с указанием того, сколько и каких случаев в обучающем наборе ей соответствует. Для выделенного узла в правой части экрана отображается его описание и *гистограмма* с распределением значений. Кнопкой **Copy to Excel** можно перенести результат из окна просмотра на новый *лист* *Excel* (для дерева решений в *Excel* будет перенесено его растровое изображение).



**Рис. 14.5.**Построенное дерево решений

Щелкнув *по* узлу дерева правой клавишей мыши и выбрав в контекстном *меню* **DrillThroughModelColumns** (можно примерно перевести как "*детализация* использовавшихся моделью данных") мы получим новую таблицу *Excel*, содержащую набор строк из обучающей выборки, которые соответствуют данному узлу ([рис. 14.6](https://intuit.ru/studies/courses/2312/612/lecture/13274?page=1#image.14.6)).



**Рис. 14.6.**Результат выполнения DrillThroughModelColumns

На [рис. 14.7](https://intuit.ru/studies/courses/2312/612/lecture/13274?page=1#image.14.7) представлена *диаграмма* зависимостей, показывающая выявленные взаимосвязи между параметрами.Ее также можно скопировать в *Excel*.Выделяя на диаграмме узел, можно увидеть все влияющие на него.

Закроем окно просмотра модели. Если нужно будет снова просмотреть ее параметры, воспользуйтесь инструментом **Browse**, который находится в группе **ModelUsage**.



**Рис. 14.7.**Диаграмма зависимостей

Для того чтобы управлять имеющимися на сервере структурами и моделями интеллектуального анализа, можно воспользоваться соответствующим мастером, запускаемым *по* нажатию кнопки **ManageModels** на вкладке **DataMining** ([рис. 14.8](https://intuit.ru/studies/courses/2312/612/lecture/13274?page=1#image.14.8)). Он позволяет просмотреть имеющиеся структуры и модели, переименовать их, удалить ненужные, выполнить другие действия на сервере прямо из **DataMiningClient**.



**Рис. 14.8.**Окно управления моделями

**Задание 1.** Создайте модель интеллектуального анализа, аналогичную описанной в лабораторной работе.

**Задание 2.** Воспользуйтесь набором данных в таблице **Associate** и одноименным мастером для создания модели, описывающей взаимосвязи между категориями товаров в одном заказе. При необходимости, воспользуйтесь справочной системой по инструменту. Проанализируйте выявленные правила и диаграмму зависимостей. Сравните с результатами, полученными в ["Использование инструментов "Prediction Calculator" и "ShoppingbasketAnalysis""](https://intuit.ru/studies/courses/2312/612/lecture/13268)(раздел "Анализ покупательского поведения").