**Самостоятельная работа 3:**

**Использование инструментов "FillFromExample" и "Forecast"**

**Аннотация:**В данной лабораторной работе будет рассмотрено использование инструментов "Заполнение по примеру" ("FillFromExample") и "Прогноз" ("Forecast"), относящихся к компоненту "Средства анализа таблиц для Excel" пакета надстроек интеллектуального анализа данных для MicrosoftOffice 2007.

**Ключевые слова:**[excel](https://www.intuit.ru/studies/courses/2312/612/lecture/13265?page=1" \l "keyword1), [обучающая выборка](https://www.intuit.ru/studies/courses/2312/612/lecture/13265?page=1#keyword2), [алгоритм](https://www.intuit.ru/studies/courses/2312/612/lecture/13265?page=1#keyword3), [предметной области](https://www.intuit.ru/studies/courses/2312/612/lecture/13265?page=1#keyword4), [анализ](https://www.intuit.ru/studies/courses/2312/612/lecture/13265?page=1#keyword5), [значение](https://www.intuit.ru/studies/courses/2312/612/lecture/13265?page=1#keyword6), [транспонирование](https://www.intuit.ru/studies/courses/2312/612/lecture/13265?page=1#keyword7), [буфер](https://www.intuit.ru/studies/courses/2312/612/lecture/13265?page=1#keyword8), [меню](https://www.intuit.ru/studies/courses/2312/612/lecture/13265?page=1#keyword9), [поле](https://www.intuit.ru/studies/courses/2312/612/lecture/13265?page=1#keyword10), [взаимно независимые](https://www.intuit.ru/studies/courses/2312/612/lecture/13265?page=1#keyword11), [вес](https://www.intuit.ru/studies/courses/2312/612/lecture/13265?page=1#keyword13), [лист](https://www.intuit.ru/studies/courses/2312/612/lecture/13265?page=1#keyword15), [Currency](https://www.intuit.ru/studies/courses/2312/612/lecture/13265?page=2#keyword18), [периодичность](https://www.intuit.ru/studies/courses/2312/612/lecture/13265?page=2#keyword20), [тренд](https://www.intuit.ru/studies/courses/2312/612/lecture/13265?page=2#keyword21), [определение](https://www.intuit.ru/studies/courses/2312/612/lecture/13265?page=2#keyword22)

Оба рассматриваемых инструмента используются для решения задач прогнозирования неизвестных значений параметров. Поэтому в обоих случаях требуется обучающий набор данных, на базе которого строится модель, применяемая для предсказания.

### Заполнение по примеру

В качестве учебного набора данных, как и в прошлой лабораторной будем использовать локализованный пример для *Excel*, взятый с <http://russiandmaddins.codeplex.com/>

Нужные данные находятся на листе **"Заполнение из примера"** ([рис. 6.1](https://www.intuit.ru/studies/courses/2312/612/lecture/13265?page=1#image.6.1)). Здесь описывается ряд клиентов магазина. Для некоторых из них отмечено, является ли данный клиент высокодоходным. Эти строки будут использоваться как *обучающая выборка*. Задачей анализа будет являться оценка остальных клиентов по этому параметру.



[увеличить изображение](https://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/6/files/19_01.jpg)
**Рис. 6.1.**Набор данных для инструмента FillFromExample

Для решения этой задачи используется *алгоритм* MicrosoftLogisticRegression. Необходимо понимать, что для создания модели в обучающей выборке должны быть представлены варианты со всеми возможными значениями целевого столбца. Необходимое число примеров зависит от особенностей *предметной области*. Но во многих случаях справедливо, что чем больше характерных примеров в обучающей выборке, тем более качественно будет обучена модель.

Соответственно, данный инструмент непригоден для задачи предсказания значений параметра, который может принимать непрерывные числовые значения.

Еще одна особенность - *анализ* проводится по столбцам (т.е. предсказывается *значение* столбца). Если ряд, который необходимо заполнить, хранится в виде строки, перед началом анализа надо выполнить *транспонирование* (скопировать в *буфер*, выбрать в контекстном *меню* **"Специальная вставка"** и отметить флажок **"Транспонировать"**).

Запустим инструмент **FillFromExample**. В первом окне будет предложено выбрать столбец, содержащий образцы данных. В нашем случае он автоматически определен верно - **"Высокодоходный клиент"**. Как и в предыдущих случаях, по ссылке **"Choosecolumnstobeusedforanalysis"**, можно выбрать столбцы, учитываемые при анализе. Эвристический механизм определил, что *поле* **ID** учитывать не надо. На практике, рекомендуемые настройки стоит менять только в случае, если точно известно о *взаимной независимости* параметров. После запуска, инструмент формирует отчет об обнаруженных шаблонах ([рис. 6.3](https://www.intuit.ru/studies/courses/2312/612/lecture/13265?page=1#image.6.3)), и добавляет столбец с предсказанными значениями к исходной таблице.



[увеличить изображение](https://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/6/files/19_02.jpg)
**Рис. 6.2.**Настройка инструмента FillFromExample



[увеличить изображение](https://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/6/files/19_03.jpg)
**Рис. 6.3.**Отчет об обнаруженных шаблонах

В отчете описываются выявленные зависимости между значением целевого столбца (в нашем случае **"да"** или **"нет"**) и значениями других столбцов. На [рис. 6.3](https://www.intuit.ru/studies/courses/2312/612/lecture/13265?page=1#image.6.3) видно, что весовой коэффициент для **"Да"**, соответствующий значению "2-5 км" параметра **"Расстояние до работы"**, равен 34. Это *значение* имеет самый большой удельный *вес* при выборе варианта **"Да"**. Это можно интерпретировать, как **"расстояние 2-5 км до работы"** во многом определяет выбор в пользу покупки велосипеда.



[увеличить изображение](https://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/6/files/19_04.jpg)
**Рис. 6.4.**Полученные оценки заносятся в исходную таблицу

Для каждой строки рассчитывается итоговая оценка для каждого варианта (в примере для **"Да"** и **"Нет"**) и делается выбор в пользу значения с наибольшим суммарным удельным весом. Оно заносится в столбец с суффиксом **"\_Extended"**(на рисунке **"Высоко\_доходный клиент\_Extended"**). Для записей, на которых модель обучалась, *значение* этого столбца совпадает с образцом.

Предположим, мы получили дополнительные данные о каких-то клиентах. Можно изменить образец ([рис. 6.4](https://www.intuit.ru/studies/courses/2312/612/lecture/13265?page=1#image.6.4), последняя строка) и снова запустить инструмент. Новые значения будут получены с учетом уточнений в наборе обучающих данных. Подобные итерации позволяют последовательно уточнять производимую оценку значений.

**Задание.** Проведите анализ и опишите полученные результаты.

Измените обучающий набор данных следующим образом. Найдите строку со значением "расстояние до работы 2-5 км", (например, строку с идентификатором 19562, 97-я строка в таблице) и для параметра **"Высоко-доходный клиент"** поставьте значение **"Нет"**.Повторите анализ. Как изменился отчет о шаблонах? Объясните эти изменения.

Для того, чтобы полностью удалить результаты работы инструмента, достаточно удалить *лист* с отчетом и добавленный столбец в таблице с исходными данными.

### Прогноз

Инструмент **Forecast** позволяет построить прогноз значений числового ряда. Ряд должен быть представлен столбцом в таблице (если исследуемые значения организованы в виде строки, требуется, как и в случае инструмента **"FillFromExample"**, выполнить *транспонирование*).

В используемом нами файле *Excel* на листе прогнозирование есть набор данных по суммам продаж велосипедов марки М200 по месяцам в трех разных регионах. Таким образом, для исследования мы имеем три числовые последовательности, возможно связанные между собой ([рис. 6.5](https://www.intuit.ru/studies/courses/2312/612/lecture/13265?page=2#image.6.5)). В процессе работы инструмент строит модель с использованием алгоритма временных рядов (**MicrosoftTimeSeries**). Для его работы необходимо, чтобы в исследуемых столбцах были только числовые значения (пропуски допустимы). Предсказывать можно числовые (непрерывные) или "денежные" (тип *currency*) значения. Инструмент не рассчитан на предсказание дат.



[увеличить изображение](https://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/6/files/19_05.jpg)
**Рис. 6.5.**Образец данных для прогнозирования - продажи по месяцам в разных регионах

Как отмечается в [[1](https://www.intuit.ru/studies/courses/2312/612/literature#literature.1)], инструмент ищет в анализируемой последовательности шаблоны следующих типов:

* **тренд** - тенденцию изменения значений. Тренд может быть восходящим (возрастание значений ряда) или нисходящим (уменьшение значений);
* **периодичность (сезонность)** - событие повторяется через определённые интервалы;
* **взаимная корреляция** - зависимость значений одного ряда от других (например, стоимость акций нефтяных компаний от цен на нефть). Алгоритмы, обнаруживающие взаимную корреляцию, входят в поставку MS SQLServer 2008 версии Enterprise или Developer, а в версии Standard недоступны.

Настройка параметров заключается в выборе анализируемых столбцов, количества предсказываемых значений ряда, указания временной отметки и типа периодичности.



[увеличить изображение](https://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/6/files/19_06.jpg)
**Рис. 6.6.**Настройка параметров инструмента Forecast

В нашем случае в качестве временной отметки логично выбрать *поле* **"Год/месяц"** (инструмент изменил его название на **"Год\_месяц"** для совместимости с требованиями SQLServer) и согласиться с исключением его из списка предсказываемых. Надо отметить, что значения в столбце, используемом в качестве временной метки, должны быть уникальны.

Что касается периодичности, то предлагаемые для выбора варианты определяются следующим образом[[1](https://www.intuit.ru/studies/courses/2312/612/literature#literature.1)]:

* **Hourly** (почасовая) - ищется периодичность 12;
* **Daily** (дневная) -ищется периодичность 5 и 7 (рабочие дни и неделя полностью);
* **Weekly**(недельная) - 4 и 13 (число недель в месяце и квартале);
* **Monthly** (месячная) - 12 (число месяцев в году);
* **Yearly** - инструмент будет автоматически обнаруживать периодичности.

Если *периодичность* неизвестна, то рекомендуется оставить **"detectautomatically"**, чтобы инструмент проверил данные на наличие периодичности разных типов.

Инструмент создает отчет с графиком ([рис. 6.7](https://www.intuit.ru/studies/courses/2312/612/lecture/13265?page=2#image.6.7)), на котором непрерывной линей обозначен "исторический *тренд*", построенный по имеющимся значениям. Пунктирной линией показано предсказываемое продолжение тренда. Обратите внимание, что временные метки для спрогнозированных значений не проставлены.



[увеличить изображение](https://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/6/files/19_07.jpg)
**Рис. 6.7.**Отчет инструмента "Прогноз"

Кроме того, в исходную таблицу добавляются результаты прогноза (столько значений, сколько было указано при запуске - [рис. 6.6](https://www.intuit.ru/studies/courses/2312/612/lecture/13265?page=2#image.6.6)). На [рис. 6.8](https://www.intuit.ru/studies/courses/2312/612/lecture/13265?page=2#image.6.8) они выделены светло-желтым фоном. Чтобы продолжить ряд временных меток, можно выделить несколько последних значений столбца **"Год/месяц"** и незаполненную область в строках с прогнозом, выбрать на панели управления в ленте **"Главная"**кнопку **"Заполнить"** ([рис. 6.8](https://www.intuit.ru/studies/courses/2312/612/lecture/13265?page=2#image.6.8) подчёркнута красным), из выпадающего списка выбрать вариант **"Прогрессия"** и указать автоматическое *определение* шага. Недостающие значения будут добавлены. Теперь ина графике будут автоматически проставлены недостающие временные метки.



[увеличить изображение](https://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/6/files/19_08.jpg)
**Рис. 6.8.**Предсказанные значения и заполнение столбца временных меток

Чтобы убрать результаты работы инструмента, надо удалить *лист* отчета и строки исходной таблицы с предсказанными значениями.

**Задание.** С помощью инструмента постройте прогноз продаж на год (12 значений). Проанализируйте график. На ваш взгляд, какой тип периодичности обнаружил инструмент в исходных данных и использует для предсказания?