**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ**

**СУМСЬКИЙ ДЕРЖАВНИЙ УНІВЕРСИТЕТ**

**КАФЕДРА КОМП’ЮТЕРНИХ НАУК**

**СЕКЦІЯ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ ПРОЕКТУВАННЯ**

ОДЗ з дисципліни

«Методи інтеграції та аналізу»

|  |  |  |
| --- | --- | --- |
| Викладач: |  | Марченко А.В. |
| Студент: |  | Тиченко Я.М. |
| Група: |  | ІТ.мз-62с |
|  |  |  |
|  |  |  |
|  |  |  |

Суми 2017

Содержание

[Вступ 2](#_Toc500584780)

[Самостійна робота 1: Використання інструментів "AnalyzeKeyInfluencers" і "DetectCategories" 3](#_Toc500584779)

[Виявлення кфАналіз ключових факторів вліяніяатегорій 3](#_Toc500584780)

[Виявлення категорій 6](#_Toc500584781)

[Самостійна робота 2: Використання інструментів "FillFromExample" і "Forecast" 10](#_Toc500584782)

[Заповнення за прикладом 10](#_Toc500584783)

[Прогноз 13](#_Toc500584784)

[Самостоятельная работа 3: Использование инструментов "HighlightExceptions" и "ScenarioAnalysis" 17](#_Toc500584785)

[Виділення исключений 17](#_Toc500584786)

[Аналіз сценаріїв 21](#_Toc500584787)

[Самостійна робота 4: Використання інструментів "Prediction Calculator" і "ShoppingbasketAnalysis " 26](#_Toc500584788)

[Розрахунок прогнозу 26](#_Toc500584789)

[Аналіз купівельної корзини 31](#_Toc500584790)

[Самостійна робота 5: Використання інструментів Data Mining Client для Excel 2007 для підготовки даних 35](#_Toc500584791)

[Explore Data 35](#_Toc500584792)

[Clean Data 36](#_Toc500584793)

[SampleData 38](#_Toc500584794)

[Самостійна робота 7: Використання інструментів Data Mining Client для Excel 2007 для створення моделі інтелектуального аналізу даних дх 50](#_Toc500584795)

[Самостійна робота 8: Аналіз точності прогнозу і використання моделі інтелектуального аналізу 55](#_Toc500584796)

[Висновок 58](#_Toc500584780)

# **Вступ**

# Метою інтелектуального аналізу даних (англ. Datamining, інші варіанти перекладу - "видобуток даних", "розкопка даних") є виявлення неявних закономірностей в наборах даних. Як наукова дисципліна вона стла активно розвиватися в 90-х роках XX століття, що було викликано значним поширенням технологій автоматизованої обробки інформації та накопиченням в комп'ютерних системах великих обсягів даних. І хоча існуючі технології дозволяли, наприклад, швидко знайти в базі даних потрібну інформацію, цього в багатьох випадках було вже недостатньо. Виникла потреба пошуку взаємозв'язків між окремими подіями серед великих обсягів даних, для чого знадобилися методи математичної статистики, теорії баз даних, теорії штучного інтелекту та ряду інших областей.

# В ході виконання роботи ми навчимося використанню технології інтелектуального аналізу даних (Data Mining) і її реалізації в Microsoft SQL Server 2008 і пов'язаних програмних продуктах. Такожм, розгляданемо всі алгоритми DM, підтримувані Microsoft SQL Server 2008, надбудови інтелектуального аналізу даних для Microsoft Office, основи мови DMX.

# **Самостійна робота 1: Використання інструментів "AnalyzeKeyInfluencers" і "DetectCategories "**

До складу пакета надбудов для MS Office 2007 входить електронна таблиця із зразками даних. Вона може бути відкрита з меню Пуск> Надбудови інтелектуального аналізу даних. Microsoft SQL Server 2008.

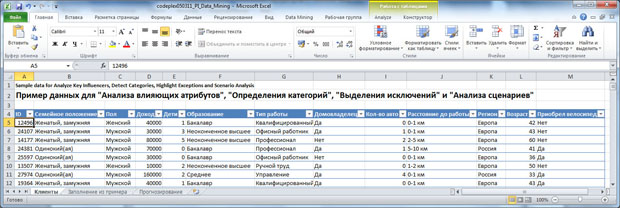
Перейдіть на вкладку Analyze (рис. 1.1). Анализируемая таблиця містить дані фірми, що продає велосипеди. У ній зібрано інформацію про клієнтів (ідентифікатор, сімейний стан, стать і т.д.) і зазначено, придбав клієнт велосипед чи ні.[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/5/files/18_01.jpg)

Рисунок 1.1 —Підготовлений набір даних

**Аналіз ключових факторів впливу**

Інструмент **AnalyzeKeyInfluencers** дозволяє визначити, як залежить цікавий для нас параметр від інших. При цьому важливо правильно визначити, що і від чого може залежати. Власне в цьому почасти й полягає майстерність аналітика, засноване на його знанні предметної області і використовуваних методів **DM**.

У зв'язку з тим, що ми оцінюємо ступінь взаємного впливу різних параметрів один на одного, варто відразу прибрати з розгляду повністю незалежні і навпаки, повністю залежні. Нехай, наприклад, ми хочемо оцінити вплив різних чинників на рівень заробітної плати людини. Якщо у нас є поле, що містить унікальний ідентифікатор (наприклад, порядковий номер запису в таблиці або номер паспорта), його варто прибрати з розгляду, що не впливає на значення досліджуваного параметра. Інший приклад, нехай у нас є значення заробітної плати за місяць і за рік, що розраховується як заробітна плата за місяць, помножена на 12. Ми знаємо, що ці значення завжди пов'язані, шукати залежність одного від іншого засобами DM не має сенсу, а наявна сильна залежність приховає вплив інших факторів, яке ми як раз і хочемо виявити.  
 Тепер визначимо, від чого залежить рішення клієнта про покупку велосипеда. Натискаємо на кнопку **Analyze Key Influencers** і вказуємо в якості цільового стовпчика стовпець **"Приобрел велосипед"** ( [рис. 1.2](http://www.intuit.ru/studies/courses/2312/612/lecture/13264?page=1#image.5.2)). Перейдемо за посиланням **"Choose columns to be used for analysis"**, щоб вказати параметри, вплив яких ми хочемо оцінити (рис. 1.3). Тут скинемо позначку навпроти **"ID"** и **"Приобрел велосипед**.

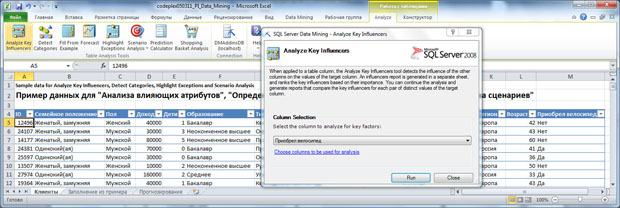
[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/5/files/18_02.jpg)

Рисунок 1.2 —Підготовлений набір даних

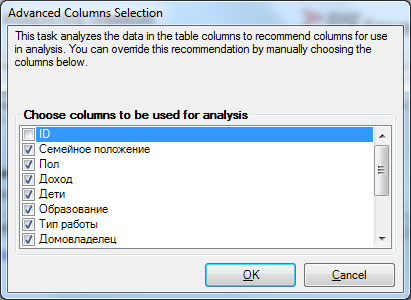


Рисунок 1.3 —Підготовлений набір даних

Вибір параметрів, від яких залежить аналізований

Після запуску процедури аналізу (по кнопці Run, рис. 1.2) буде сформований звіт про фактори впливу та запропоновано формування додаткового порівняльного звіту (рис. 1.5). В основному звіті вказується стовпець (Column), його значення (Value), значення цільового стовпчика, з яким воно зв'язується (Favors) і рівень впливу (Relative Impact), оцінюваний за шкалою від 0 до 100 балів. З представленого на рис. 1.5 звіту видно, що на рішення не купувати велосипед в найбільшою мірою впливає наявність 2-х автомобілів. У той же час не слід сприймати оцінку 100 балів, як ознака того, що в 100% випадків власники 2-х машин велосипед не купували (подивіться набір даних, там є і поєднання "2 машини - велосипед куплений", але їх меншість). Другий за рівнем впливу на відмову від покупки фактор - "Сімейний стан" = "одружений, заміжня".

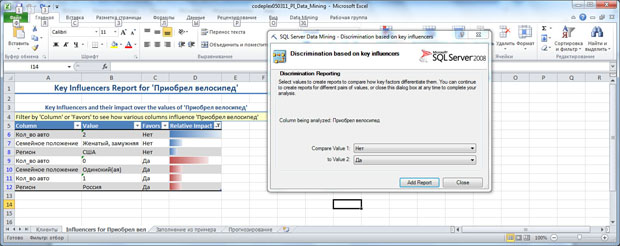
Найбільший вплив на позитивне рішення про придбання велосипеда надає відсутність у клієнта машини.[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/5/files/18_04.jpg)

Рисунок 1.4 —Підготовлений набір даних

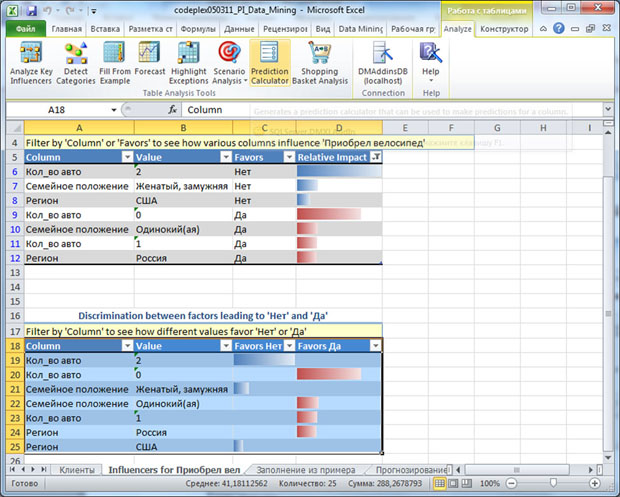
[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/5/files/18_05.jpg)

Рисунок 1.5 —Підготовлений набір даних

Якщо додати порівняльний звіт для двох обраних значень (рис. 1.5, Add Report), можна побачити, чим відрізняється вибір на користь одного значення цільового стовпчика від вибору на користь іншого (рис. 1.5). У нашому прикладі просто відбудеться перегрупування вихідного звіту, тому що можливих значень всього 2. В інших випадках, додатковий звіт дозволяє провести детальне порівняння двох обраних варіантів.

Як зазначається в [1], якщо цільової або інший стовпець, що обробляється інструментом Analyze Key Influencers, містить багато різних числових значень, то проводиться дискретизація. Весь інтервал значень ділиться на кілька діапазонів, кожен з яких розглядається як одне з можливих значень (наприклад, замість точного значення 2,5 ми отримаємо "діапазон від 2 до 3").

**Виявлення категорій**

Інструмент Detect Categories дозволяє вирішити задачу кластеризації, тобто поділу всього безлічі варіантів на "природні" групи, члени яких найбільш близькі за рядом ознак. Подібна задача також називається завданням сегментації.

Отже, в нашому прикладі є опис безлічі клієнтів і потрібно розділити їх на невелику кількість груп (щоб окремим групам сформувати спеціальну пропозицію і т.п.).

У зв'язку з тим, що в процесі роботи інструмент додає дані в вихідну таблицю, рекомендується перед початком роботи зробити її копію (рис. 1.6).

Після цього натискаємо кнопку Detect Categories і налаштовуємо параметри (рис. 1.7). Тут хочеться звернути увагу на атрибут ID, який як було зазначено вище, не можна буде враховувати в ході аналіза.Поетому він автоматично виключений. У нашому випадку, інші атрибути можна залишити. Ще раз хотілося б повторити, що цей вибір кожен раз робиться виходячи з особливостей предметної області

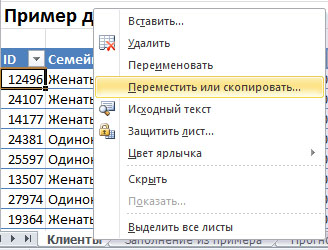
.

Рисунок 1.6 —Detect Categories

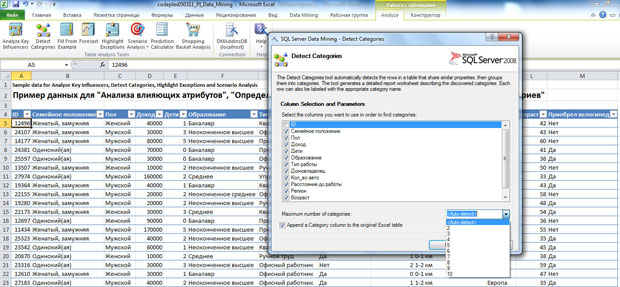
[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/5/files/18_07.jpg)

Рисунок 1.7 —Вибір параметрів, які будуть аналізуватися

Крім вказівки врахованих параметрів, можна явно вказати число категорій (або залишити за замовчуванням автоматичне визначення). Також за замовчуванням поставлений прапорець "Appenda Category column to the original Excel table", який вказує, що до записів у вихідній таблиці буде додано вказівку на категорію.

Сформований звіт містить 3 розділу. У першому вказані певні інструментом категорії і число рядків, що потрапляють в кожну з них (рис. 1.8). Поле з назвою категорії допускає редагування і можна зіставити категорії більш значуще назву. Наприклад, як буде показано нижче, для клієнтів першої категорії характерний низький дохід і її можна так і назвати. Коли ми введемо цю назву, всюди крім діаграми Category Profiles Chat, воно автоматично замінить "Category 1" (щоб назва поміняти і на діаграмі, треба натиснути <Alt> + <Ctrl> + <F5>).

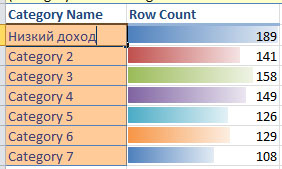


Рисунок 1.8 —виділені категорії

Наступний розділ звіту описує характеристики виділених категорій і ступінь впливу кожного параметра (рис. 1.9). За замовчуванням відображається інформація тільки по одній категорії, але клацанням миші по іконці фільтра на заголовку таблиці можна відображати всі канали категорій або якогось їх поєднання, як це показано на малюнку.

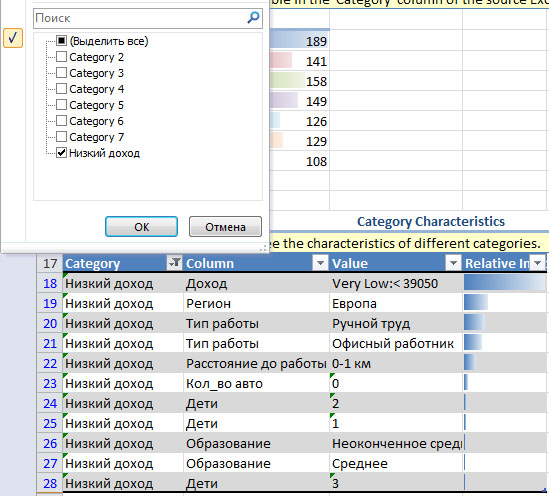
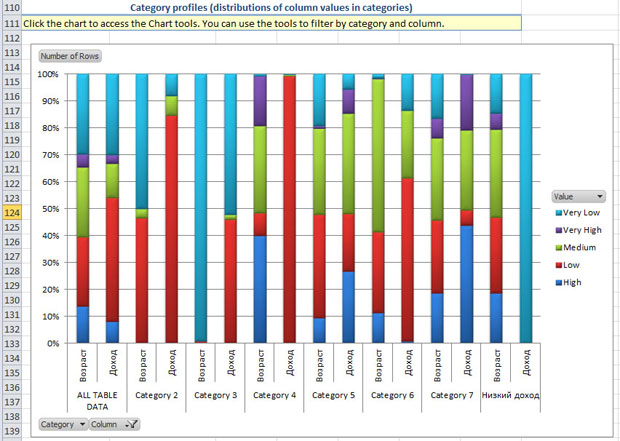
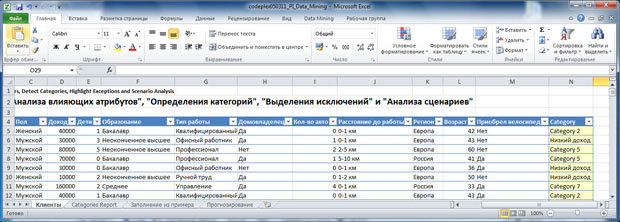


Рисунок 1.9 —опис категорії

Третій розділ звіту - це діаграма профілів категорій. Вона показує кількість рядків даних в кожній категорії з кожним значенням обраних параметрів. За замовчуванням відображається тільки один параметр. Для розглянутого прикладу це вік. Але в нижній частині діаграми є фільтр Column, за допомогою якого можна змінити число параметрів. Наприклад, на рис. 1.10 для кожної категорії відображається розподіл за віком і доходу. З нього видно, що клієнти перейменованої нами категорії "Низький дохід" насправді мають дуже низький дохід. А клієнти категорії 3 в переважній більшості дуже молоді.

[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/5/files/18_10.jpg)

**Рис. 1.10.** Діаграма профілів категорій

[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/5/files/18_11.jpg)

**Рис. 1.11.** Зіставлення категорій записів у вихідній таблиці

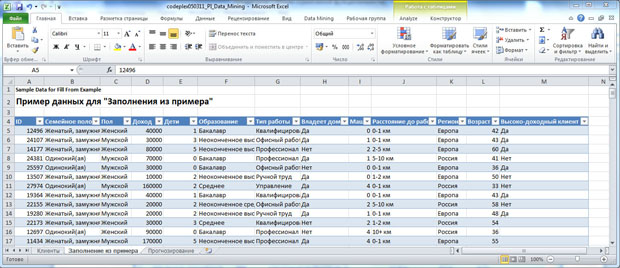
Малюнок 1.11 показує, що всіх записів вихідної таблиці тепер сопоставлена категорія, до якої вони належать. А за допомогою фільтрів можна переглянути записи, які стосуються обраної категорії.

**Самостійна робота 2: Використання інструментів "FillFromExample" і "Forecast"**

Обидва розглянутих інструменту використовуються для вирішення завдань прогнозування невідомих значень параметрів. Тому в обох випадках потрібно навчальний набір даних, на базі якого будується модель, що застосовується для передбачення.

**Заповнення за прикладом**

Потрібні дані знаходяться на аркуші "Заповнення з прикладу" (рис. 2.1). Тут описується ряд клієнтів магазину. Для деяких з них відзначено, чи є даний клієнт високоприбутковим. Ці рядки будуть використовуватися як навчальна вибірка. Завданням аналізу буде оцінка інших клієнтів за цим параметром.

[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/6/files/19_01.jpg)

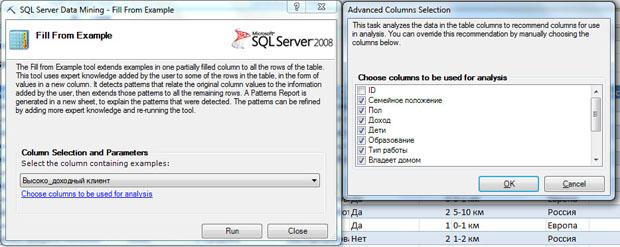
**Рис. 2.1.** Набір даних для інструмента FillFromExample

Для вирішення цього завдання використовується алгоритм MicrosoftLogisticRegression. Необхідно розуміти, що для створення моделі в навчальній вибірці повинні бути представлені варіанти з усіма можливими значеннями цільового стовпчика. Необхідна кількість прикладів залежить від особливостей предметної області. Але в багатьох випадках справедливо, що чим більше характерних прикладів в навчальній вибірці, тим якісніше буде навчена модель.

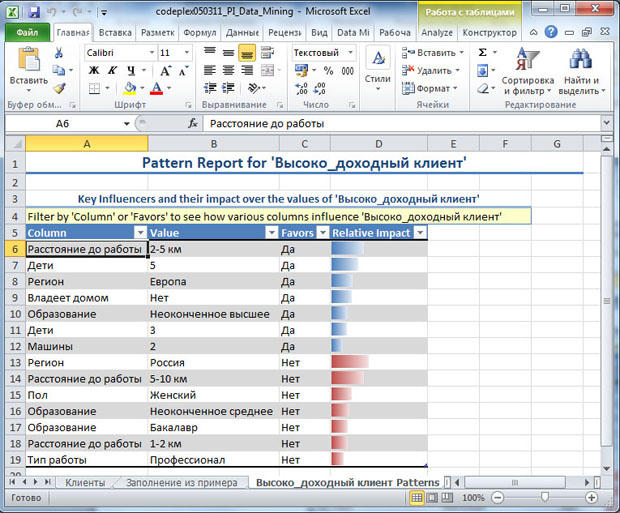
Відповідно, даний інструмент непридатний для завдання передбачення значень параметра, який може приймати безперервні числові значення.

Ще одна особливість - аналіз проводиться за стовпцями (тобто передбачається значення стовпця). Якщо ряд, який необхідно заповнити, зберігається у вигляді рядка, перед початком аналізу треба виконати транспонування (скопіювати в буфер, вибрати в контекстному меню "Спеціальна вставка" і відзначити прапорець "Транспонувати").

Запустимо інструмент FillFromExample. У першому вікні буде запропоновано вибрати стовпець, що містить зразки даних. У нашому випадку він автоматично визначено вірно - "Високоприбутковий клієнт". Як і в попередніх випадках, за посиланням "Choosecolumnstobeusedforanalysis", можна вибрати стовпці, що враховуються при аналізі. Евристичний механізм визначив, що поле ID враховувати не треба. На практиці, рекомендовані настройки варто міняти тільки в разі, якщо точно відомо про взаємну незалежності параметрів. Після запуску, інструмент формує звіт про виявлені шаблонах (рис. 2.3), і додає стовпець з передбаченими значеннями до вихідної таблиці.

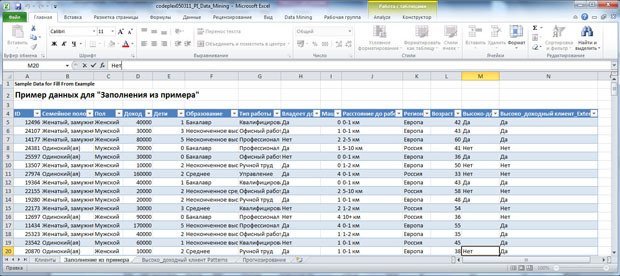
[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/6/files/19_02.jpg)

**Рис. 2.2.** Налаштування інструменту FillFromExample

[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/6/files/19_03.jpg)

**Рис. 2.3.** Звіт про виявлені шаблонах

У звіті описуються виявлені залежності між значенням цільового шпальти (в нашому випадку "так" або "ні") і значеннями інших стовпців. На рис. 2.3 видно, що ваговий коефіцієнт для "Так", відповідний значенням "2-5 км" параметра "Відстань до роботи", дорівнює 34. Це значення має найбільшу питому вагу при виборі варіанту "Так". Це можна інтерпретувати, як "відстань 2-5 км до роботи" багато в чому визначає вибір на користь покупки велосипеда.

[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/6/files/19_04.jpg)

**Рис. 2.4.** Отримані оцінки заносяться в вихідну таблицю

Для кожного рядка розраховується підсумкова оцінка для кожного варіанта (в прикладі для "Так" і "Ні") і робиться вибір на користь значення з найбільшою сумарною питомою вагою. Воно заноситься в стовпець з суфіксом "\_Extended" (на малюнку "Високо\_доходний кліент\_Extended"). Для записів, на яких модель навчалася, значення цього стовпця збігається зі зразком.

Припустимо, ми отримали додаткові дані про якісь клієнтів. Можна змінити зразок (рис. 2.4, останній рядок) і знову запустити інструмент. Нові значення будуть отримані з урахуванням уточнень в наборі навчальних даних. Подібні ітерації дозволяють послідовно уточнювати вироблену оцінку значень.

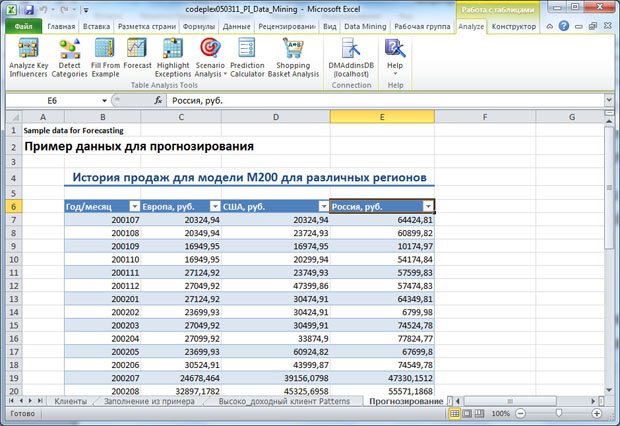
Завдання. Проведіть аналіз і опишіть отримані результати.

Змініть навчальний набір даних в такий спосіб. Знайдіть рядок зі значенням "відстань до роботи 2-5 км", (наприклад, рядок з ідентифікатором 19562, 97-й рядок у таблиці) і для параметра "Високо-прибутковий клієнт" поставте значення "Ні" .Повторіте аналіз. Як змінився звіт про шаблони? Поясніть ці зміни.

Для того, щоб повністю видалити результати роботи інструменту, досить видалити лист із звітом і доданий стовпець в таблиці з вихідними даними.

**Прогноз**

Інструмент Forecast дозволяє побудувати прогноз значень числового ряду. Ряд повинен бути представлений стовпцем в таблиці (якщо досліджувані значення організовані у вигляді рядка, потрібно, як і в разі інструменту "FillFromExample", виконати транспонування).

У використовуваному нами файлі Excel на аркуші прогнозування є набір даних за сумами продажів велосипедів марки М200 по місяцях в трьох різних регіонах. Таким чином, для дослідження ми маємо три числові послідовності, можливо пов'язані між собою (рис. 2.5). В процесі роботи інструмент будує модель з використанням алгоритму часових рядів (MicrosoftTimeSeries). Для його роботи необхідно, щоб в досліджуваних шпальтах були тільки числові значення (пропуски допустимі). Передбачати можна числові (безперервні) або "грошові" (тип currency) значення. Інструмент не розрахований на пророкування дат.[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/6/files/19_05.jpg)

**Рис. 2.5.** Зразок даних для прогнозування - продажу по місяцях в різних регіонах

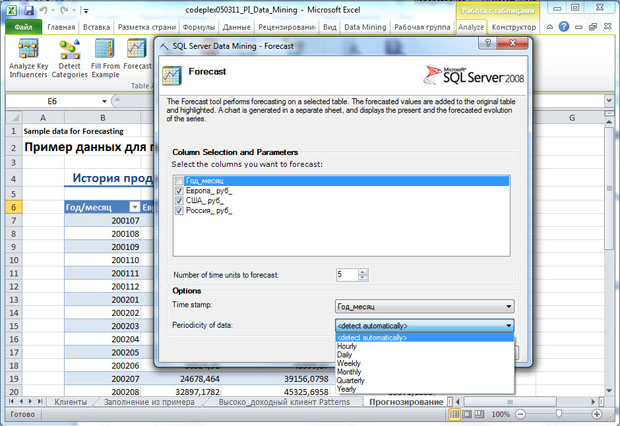
Як зазначається в [1], інструмент шукає в аналізованої послідовності шаблони наступних типів:

тренд - тенденцію зміни значень. Тренд може бути висхідним (зростання значень ряду) або низхідним (зменшення значень);

періодичність (сезонність) - подія повторюється через певні інтервали;

взаємна кореляція - залежність значень одного ряду від інших (наприклад, вартість акцій нафтових компаній від цін на нафту). Алгоритми, які виявляють взаємну кореляцію, входять в поставку MS SQLServer 2008 версії Enterprise або Developer, а у версії Standard недоступні.

Налаштування параметрів полягає у виборі аналізованих стовпців, кількості передбачаються значень ряду, вказівки тимчасової позначки і типу періодичності.

[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/6/files/19_06.jpg)

**Рис. 2.6.** Налаштування параметрів інструмента Forecast

У нашому випадку в якості тимчасової позначки логічно вибрати поле "Рік / місяць" (інструмент змінив його назву на "Год\_месяц" для сумісності з вимогами SQLServer) і погодитися з виключенням його зі списку передбачаються. Треба відзначити, що значення в стовпці, що використовується в якості тимчасової мітки, повинні бути унікальні.

Що стосується періодичності, то пропоновані для вибору варіанти визначаються наступним чином [1]:

Hourly (погодинна) - шукається періодичність 12;

Daily (денна) -іщется періодичність 5 і 7 (робочі дні і тиждень повністю);

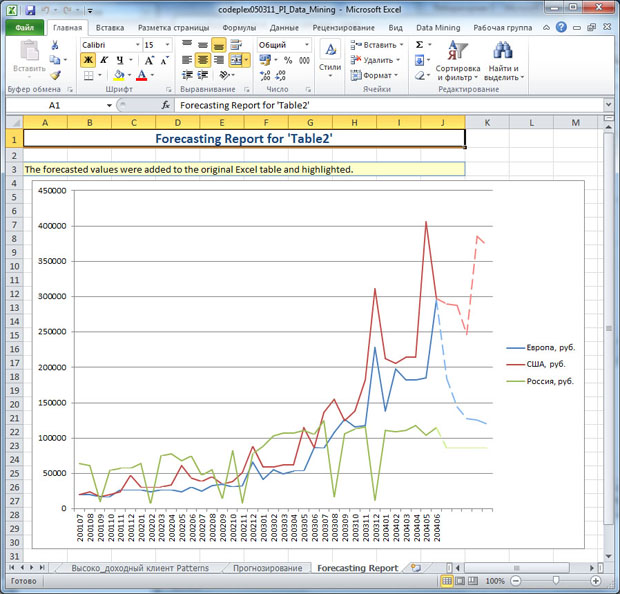
Weekly (тижнева) - 4 і 13 (число тижнів у місяці та кварталі);

Monthly (місячна) - 12 (число місяців в році);

Yearly - інструмент буде автоматично виявляти періодичності.

Якщо періодичність невідома, то рекомендується залишити "detectautomatically", щоб інструмент перевірив дані на наявність періодичності різних типів.

Інструмент створює звіт з графіком (рис. 2.7), на якому безперервної линів позначений "історичний тренд", побудований за наявними значеннями. Пунктирною лінією показано пророкує продовження тренда. Зверніть увагу, що тимчасові мітки для прогнозованих у допустимих проставлені.

[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/6/files/19_07.jpg)

**Рис. 2.7.** Звіт інструменту "Прогноз"

Крім того, в вихідну таблицю додаються результати прогнозу (стільки значень, скільки було вказано при запуску - рис. 2.6). На рис. 2.8 вони виділені світло-жовтим фоном. Щоб продовжити ряд тимчасових міток, можна виділити кілька останніх значень стовпця "Рік / місяць" і незаповнену область в рядках з прогнозом, вибрати на панелі управління в стрічці "Головна" кнопку "Заповнити" (рис. 2.8 підкреслена червоним), зі списку вибрати варіант "Прогресія" і вказати автоматичне визначення кроку. Відсутні значення будуть додані. Тепер ина графіку будуть автоматично проставлені відсутні тимчасові мітки.

[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/6/files/19_08.jpg)

**Рис. 2.8.** Передбачені значення і заповнення стовпчика тимчасових міток

Щоб прибрати результати роботи інструменту, треба видалити аркуш звіту і рядки вихідної таблиці з передбаченими значеннями.

**Самостійна робота 3: Використання інструментів "HighlightExceptions" і "ScenarioAnalysis"**

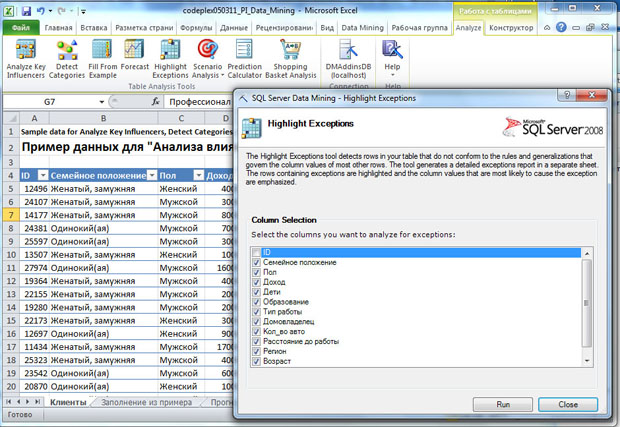
**Виділення винятків**

Як випливає з назви, інструмент дозволяє виявити дані, що виділяються серед наявного набору. Це може бути корисно в ряді випадків. По-перше, це можуть бути помилкові дані (наприклад, результати помилки оператора при введенні якихось значень). По-друге, виключення можуть становити окремий інтерес (як, наприклад, в разі виявлення шахрайських дій з банківськими картами і т.п.). Крім того, аналіз винятків може розглядатися як попередня частина інтелектуального аналізу даних за допомогою інших методів. Зокрема, це дозволяє виключити потрапляння нетипових прикладів в навчальну вибірку.

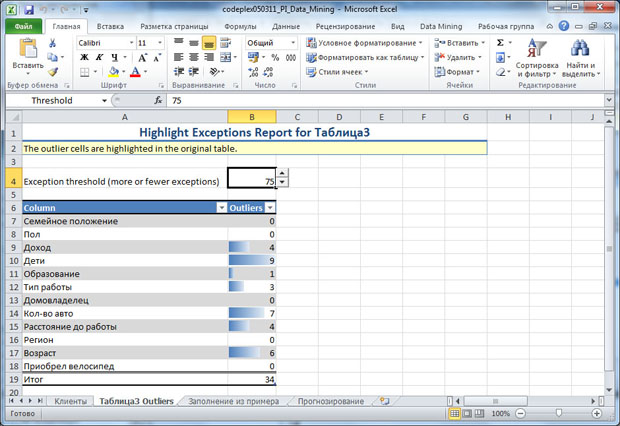
В ході роботи інструмент HighlightExceptions створює тимчасову модель інтелектуального аналізу з використанням алгоритму MicrosoftClustering. Для кожної аналізованої рядки оцінюється ступінь приналежності виявлених кластерам.Значенія, що знаходяться далеко від всіх кластерів, позначаються як виключення.

При запуску інструменту можна відзначити стовпці, що не враховуються при аналізі. У рекомендаціях щодо використання [1,3] вказується, що бажано виключити з аналізу стовпці з унікальними значеннями (імена, ідентифікатори), а також містять багато порожніх значень або довільний текст. На рис. 3.1 видно, що при аналізі набору даних "Клієнти" інструмент пропонує виключити з розгляду поле ID.

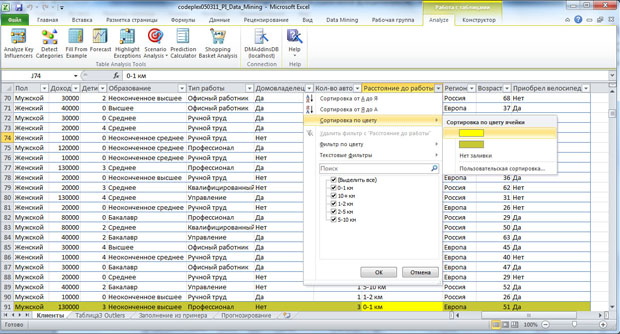
За підсумками роботи (а працює цей інструмент трохи довше розглянутих нами раніше) формується звіт (рис. 3.2) і в початковому наборі даних виключення виділяються кольором (рис. 3.3).

[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_01.jpg)

**Рис. 3.1.** Запуск инструмента HighlightExceptions

[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_02.jpg)

**Рис. 3.2.** Отчет по проведенному анализу данных

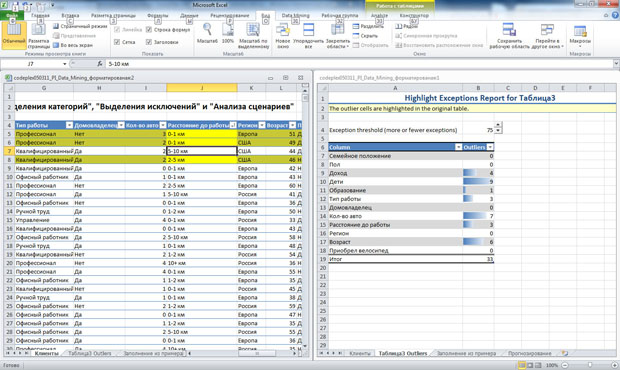
[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_03.jpg)

**Рис. 3.3.** Винятки виділяються кольором, що дозволяє провести сортування

На рис. 3.2 видно, що інструмент дозволяє вказати поріг відхилення від норми (Exception threshold), вимірюваний у відсотках (оцінка ймовірності того, що виділене значення відноситься до винятків). Зменшення порога призведе до того, що більше записів буде розглядатися як виключення, збільшення - навпаки. При значенні за замовчуванням в 75% нашому наборі даних виявлено 34 виключення. Звіт показує, в яких стовпчиках скільки винятків було виявлено.

Перейдемо на лист Excel з даними. Розглянуті як викиди значення виділяються в таблиці кольором: вся рядок-коричневим, конкретне значення - жовтим. Щоб згрупувати потрібні рядки можна скористатися функціями Excel, що дозволяють провести сортування за кольором.

Також можна скористатися інструментами вкладки "Вид", щоб створити нове вікно і розташувати поруч з вікном зі звітом і даними (рис. 3.4). Нехай у відібраному наборі записів ми виявили ошібку.Скажем відстань до роботи у якогось клієнта з США, що володіє двома машинами, не «0-1 км", а "5-10 км" (саме тому йому потрібно в родині 2 машини). Якщо ми змінимо значення, буде проведений автоматичний перерахунок. У разі, представленому на рис. 3.4, нове значення вже не розглядається як викид.

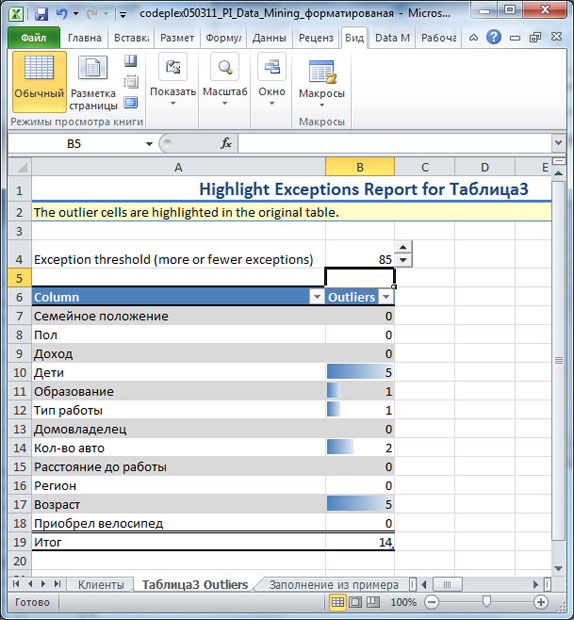
[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_04.jpg)

**Рис. 3.4.** Виправлення помилкового значення

Зверніть увагу, що не тільки змінилася розфарбування рядки таблиці, а й відбулися зміни в звіті, що показує тепер наявність 33 винятків. Автоматичний перерахунок працює тільки в тому випадку, якщо сесія роботи з аналітичними службами SQLServer залишається відкритою. Якщо таблиця Excel була закрита і знову відкрита, то автоматичного перерахунку не буде (потрібно знову провести аналіз).

Також в описах зазначається, що інструмент реагує тільки на зміни даних в діапазоні осередків, використовувався при навчанні. Якщо після початку роботи інструменту в кінець таблиці додати нові рядки, вони оцінюватися не будуть.

Як вже зазначалося вище, якщо потрібно розглядати тільки найбільш сильні викиди, можна збільшити значення порога відхилення і інструмент змінить оцінки відповідно до заданого значенням (рис. 3.5).

Повторний запуск інструменту видалить результати попереднього аналізу. З огляду на, що проводяться інструментом зміни досить складні (розфарбування рядків таблиці і т.д.), якщо потрібно видалити результати роботи, рекомендується запустити повторний аналіз, погодитися з видаленням результатів і потім у вікні, аналогічному представленому на рис. 3.1, натиснути кнопку Close (відмовитися від аналізу даних).[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_05.jpg)

**Рис. 3.5.** Увеличение порога отклонения уменьшает число исключений

**Аналіз сценаріїв**

Інструмент Scenario Analysis дозволяє моделювати вплив, який чиниться зміною одного з параметрів (значень одного стовпчика) на інший, пов'язаний з першим. В основі роботи інструменту лежить використання алгоритму Microsoft Logistic Regression. Для формування часової моделі потрібно навчальна вибірка, що містить не менше 50 записів [3].

Інструмент Scenario Analysis включає дві складові частини - "Аналіз сценарію пошуку рішень" (GoalSeek) і "Аналіз можливих варіантів" ( "What-If").

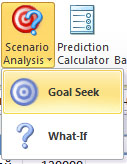
"Аналіз сценарію пошуку рішень" (GoalSeek)

Використання інструменту GoalSeek дозволяє оцінити, чи зможемо ми досягти бажаного значення в цільовому стовпці, змінюючи значення обраного параметра. Інструмент дозволяє провести аналіз як для запису, так і для всієї таблиці.

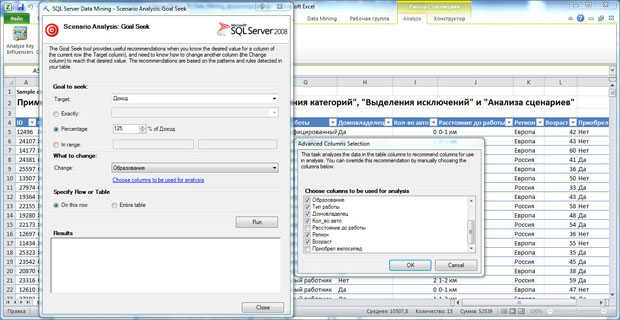
Використовуючи цей інструмент треба бути готовим, що не для всіх варіантів запиту може бути отримана відповідь. Це може бути пов'язано з тим, що у вихідних даних немає цікавлять нас поєднань. Також можуть бути проблеми через типів даних.

Крім того, не можна забувати, що запит потрібно формувати з урахуванням знань про предметну область. Наприклад, можна запросити систему, якщо людина хоче збільшити річний дохід на 20 відсотків, чи треба йому купувати велосипед. І навіть отримати якусь відповідь. Але зрозуміло, що в такій постановці саме питання є безглуздим.

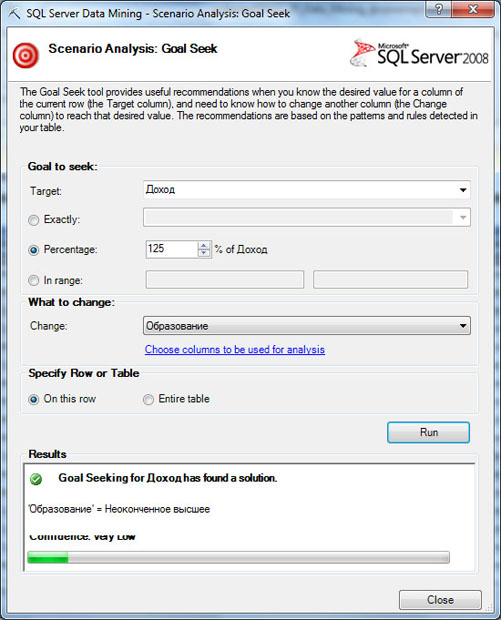
Нехай ми хочемо дізнатися, як буде впливати освіту на рівень достатку людини. Спочатку проведемо аналіз для запису. Наприклад, нас цікавить клієнт з ідентифікатором 12496 (перший запис в наборі даних) .Откройте набір даних "Клієнти" і на вкладці Analysis виберіть ScenarioAnalysis-> GoalSeek (рис. 3.6).



**Рис. 3.6.** Выбор инструмента GoalSeek

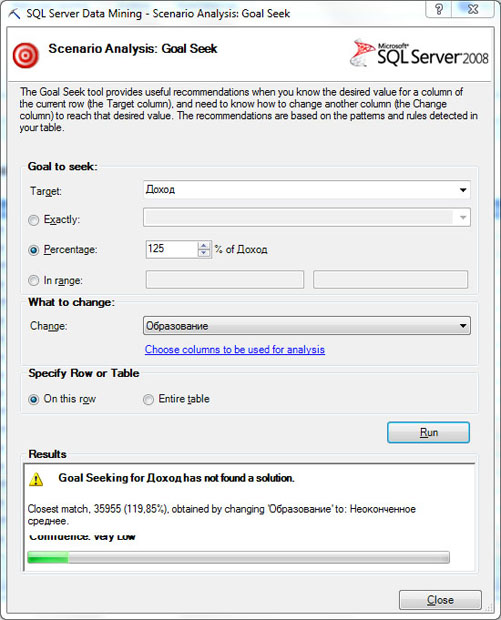
У вікні параметрів (рис. 3.7) вкажемо цільової стовпець - "Дохід", а також бажане значення - 125% від поточного. У цьому випадку інструмент вважає успішним результат, що не менше заданого (в нашому прикладі і більше). Якщо задається значення менше 100%, то успішним вважається результат, що не більше заданого. Також можна вказати точне значення і діапазон (вибравши "Inrange"). Для значень, які не є числовими, варіанти "Percentage" і "Inrange" будуть неактивні. Для досягнення шуканого значення будемо міняти стовпець "Освіта".[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_07.jpg)

**Рис. 3.7.** Настройка параметров для GoalSeek

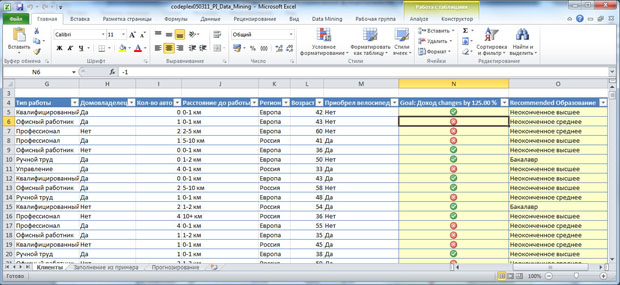
Перейшовши за посиланням "Choose columns to be used for analysis", відзначимо, що при аналізі в розгляд не беремо стовпці "ID", "Діти", "Відстань до роботи", "Придбав велосипед". Після закриття вікна "Advanced Columns Selection" варто ще раз перевірити настройки в секції "Goaltoseek" - іноді при переході між вікнами перемикач між "Exactly", "Percentage" і "Inrange" скидається в значення за замовчуванням ("Exactly")[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_08.jpg)

**Рис. 3.8.** Результат анализа для одной строки - решение найдено

Результат аналізу, виконаного після натискання кнопки Run, представлений на рис. 3.8. Для вибраного рядка знайдений шаблон, що рекомендує для параметра "Освіта" значення "Незакінчена вища". При цьому рівень достовірності - Confidence (іноді верхня частина напису затирається, як на малюнку), оцінюється як дуже низький ("Very low").

Якщо перейти на наступний рядок і знову натиснути Run, отримаємо результат для нових даних (рис. 3.9). В цьому випадку, відповідного рішення не було знайдено, і був запропонований найбільш близький варіант.[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_09.jpg)

**Рис. 3.9.** Результат анализа для одной строки - решение не найдено

[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_10.jpg)

**Рис. 3.10.** Аналіз для всієї таблиці

А якщо в секції "Specify Rowor Table" встановити перемикач в "Entire table", то сценарії будуть пораховані для всіх рядків (рис. 3.10). Результати будуть вказані в двох стовпчиках, доданих в вихідну таблицю. Для тих рядків, які відзначені хрестиком в червоному колі, відповідного бажаного сценарієм шаблону знайдено не було.

Завдання. Проведіть аналіз для окремого рядка і таблиці, аналогічний описаному вище. Прокоментуйте результати.

Примітка. Запуск процедури аналізу для ряду інших комбінацій стовпців (наприклад - цільової стовпець "покупка велосипеда" = "так", незалежна змінна - "відстань до роботи") призводить до помилки "Query (1, 50) Синтаксичний аналізатор Невірний синтаксис" value ". ", мабуть пов'язаної з некоректною обробкою деяких типів даних.

"Аналіз можливих варіантів" ( "What-If")

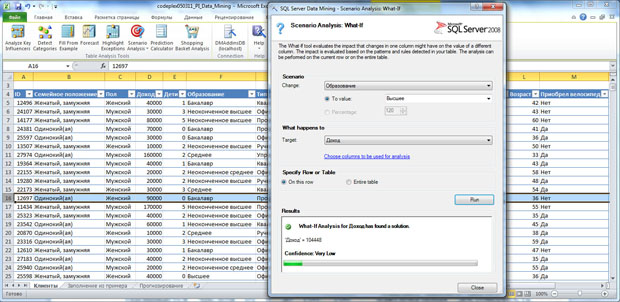
Інструмент What-If дозволяє вирішити зворотну по відношенню до GoalSeek завдання: оцінити значення цільової змінної при певному зміні заданого параметра.

Наприклад, можна оцінити, як змінився б рівень доходу людини, якби підвищився його рівень освіти. Перейдемо на запис з ідентифікатором 12697 і запустимо інструмент: Scenario Analysis-> What-If. Зазначимо параметри сценарію: освіта змінюється на "Вища" і цільової стовпець "Дохід". Отриманий для рядка результат показує, що при зміні рівня освіти дохід може дещо зрости (початкове значення 90000, середнє значення для нового шаблону 104448). Але ступінь впевненості в прогнозі не надто висока.

Аналогічно до попереднього інструменту, подібний аналіз сценарію можна зробити і для всієї таблиці цілком. В цьому випадку до вихідної таблиці додаються два стовпці - один показує нове значення цільового параметра, другий - оцінку достовірності (рис. 3.12). Достовірність оцінюється числом від 0 до 100: 100 - максимальна достовірність (абсолютна впевненість в прогнозі), 0 - мінімальна.

Завдання. Проведіть аналіз даних, аналогічний описаному вище.

Для того щоб видалити результати роботи з таблицею інструментів What-If і Scenario Analysis, досить видалити додані стовпці. При роботі з окремими рядками, ніяких додаткових дій не потрібно.

[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_11.jpg)

**Рис. 3.11.** Работа инструмента What-If для отдельной строки

[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_12.jpg)

**Рис. 3.12.** Прогноз What-If для всей таблицы

**Самостійна робота 4: Використання інструментів "Prediction Calculator" і "ShoppingbasketAnalysis"**

**Розрахунок прогнозу**

Інструмент Prediction Calculator допомагає згенерувати і налаштувати "калькулятор", який дозволяє оцінити шанси на отримання очікуваного значення цільового параметра без підключення до аналітичних служб SQL Server. Зокрема, така можливість може бути дуже корисна для віддалених користувачів.

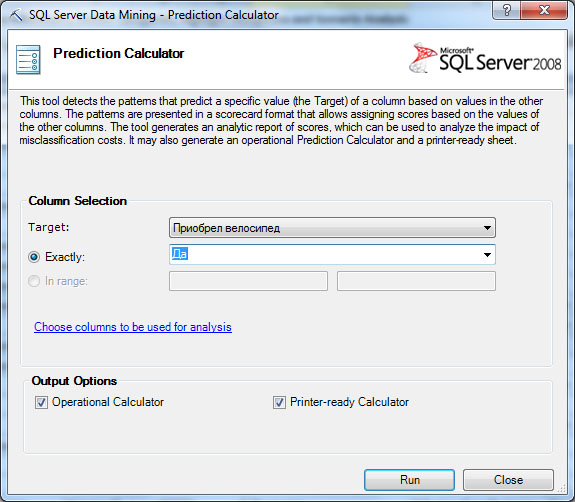
Перейдемо на набір даних "Клієнти" і на вкладці Analyze виберемо Prediction Calculator. У вікні налаштувань треба вказати цільової стовпець і шукане значення (рис. 4.1). Якщо значення цільового стовпчика розглядаються як числові з безперервного діапазону, то можна вказати, як точне значення, так і бажаний інтервал. В іншому випадку - тільки точне значення.

Рис. 4.1. Настройки инструмента Prediction Calculator

Нехай мета аналізу - визначити, чи купить клієнт велосипед. В якості цільового стовпчика вказуємо "Придбав велосипед" і значення "Так". Далі можна вказати стовпці для аналізу. Як і раніше, рекомендується враховувати при розгляді стовпці з унікальними значеннями і стовпці, один з яких дублює інший (наприклад, точне значення заробітної плати і діапазон заробітної плати).

Інструмент завжди формує звіт Prediction Calculator Report, крім того за умовчанням формуються два необов'язкових звіту - Prediction Calculator ( "калькулятор" прогнозу у вигляді таблиці Excel) і Printable Calculator (таблиця калькулятора для друку і ручної обробки).

Щоб краще розібратися з результатами роботи інструменту, перейдемо спочатку на лист зі звітом Prediction Calculator.В верхній частині звіту розташований сам калькулятор (рис. 4.2), в нижній - таблиця балів, що відповідають різним значенням параметрів (рис. 4.3).

Працюючи з калькулятором, можна описати аналізований приклад, вказуючи значення для кожного параметра. Значення в стовпець Value можна вводити або вибирати з випадаючих списків (що краще, тому що менше шансів ввести некоректне значення або діапазон). Для описуваного прикладу розраховується сума балів, яка порівнюється з рекомендованим граничним значенням. Якщо значення вище "порога", то прогноз отримує значення "істина" (на малюнку сума балів 572, порогове значення 565). Друга частина звіту пояснює отриманий результат, показуючи, скільки балів за яке значення ставиться.

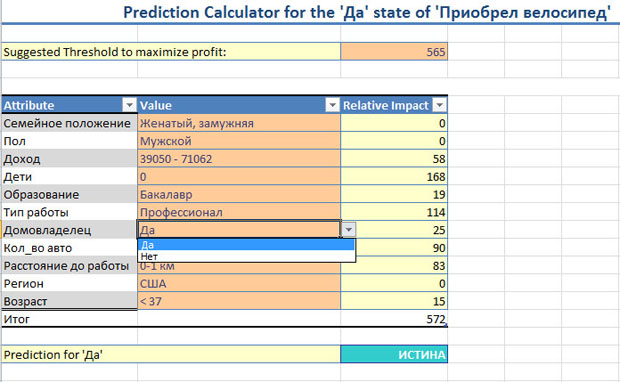
[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/8/files/21_02.jpg)

Рис. 4.2. "Калькулятор"

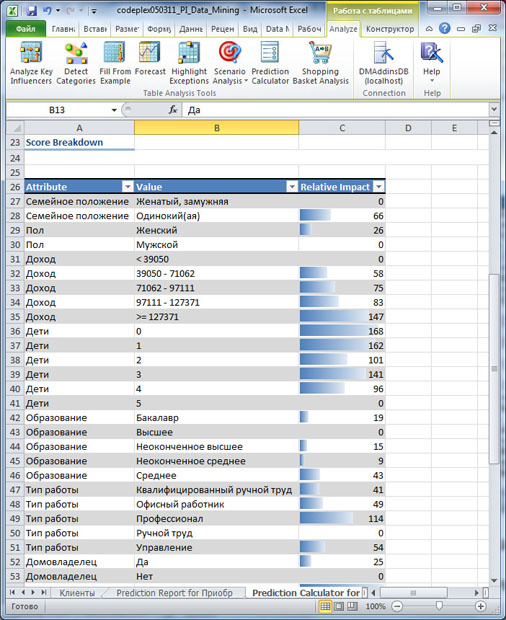
[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/8/files/21_03.jpg)

Рис. 4.3. Таблица баллов для параметров

Представлений на рис. 4.4 звіт "Printable Calculator" дозволяє вивести на друк готову форму для ручного підрахунку балів і отримання оцінки без використання комп'ютера. Це може бути зручно, наприклад, для торгових представників або інших співробітників, що працюють поза офісом і не мають доступу до комп'ютера. Все що потрібно для розрахунку прогнозу - відзначити варіанти, підсумувати бали і порівняти з граничним значенням.

Тепер перейдемо до більш цікавого питання - як же було визначено граничне значення. Звіт Prediction Calculator Report дозволяє з цим розібратися (рис. 4.5). За підсумками аналізу формується прогноз, який може бути віднесений до однієї з чотирьох категорій [1]:

істинний позитивний прогноз (TruePositive) - вірний прогноз. Наприклад, клієнт, для якого прогноз показав істину, насправді зацікавлений в покупці велосипеда. Магазин отримав прибуток;

істинний негативний прогноз (TrueNegative) - вірний негативний прогноз. Клієнт, для якого прогноз показав незацікавленість в покупці, насправді не збирається купувати велосипед. Магазин не отримав прибутку, але і не поніс витрат (на розсилку рекламних пропозицій та ін.);ложный позитивный прогноз (FalsePositive; ошибка 1 рода) - неверный прогноз, показывающий,что клиент хочет сделать покупку, хотя на самом деле это не так (может привести магазин к затратам на сопровождение клиента);

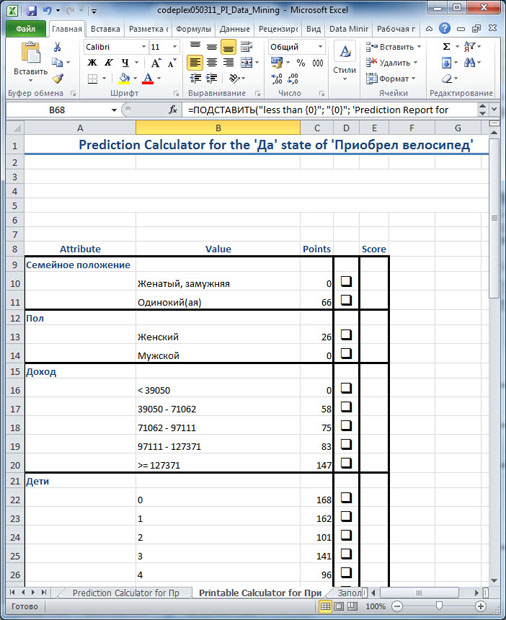
[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/8/files/21_04.jpg)

Рис. 4.4. Отчет "PrintableCalculator"

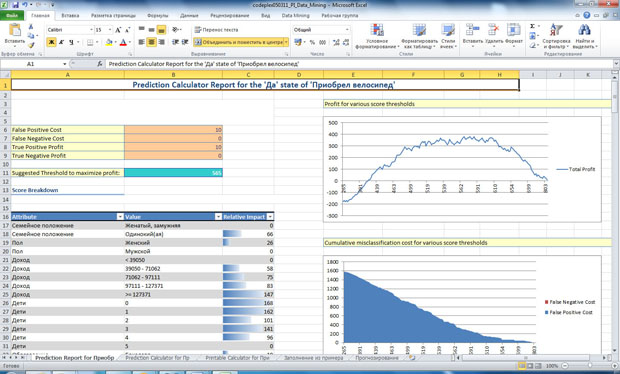
Звіт Prediction Calculator Report дозволяє вказати прибуток від справжніх прогнозів і збиток від помилкових. На основі цих даних визначається порогове значення, що забезпечує максимум прибутку. За замовчуванням, для істинного позитивного прогнозу вказується прибуток 10 (доларів або інших одиниць), для помилкового позитивного - такий же збиток (рис. 4.5, таблиця в лівій верхній частині екрану). В цьому випадку максимум прибутку (графік на рис. 4.5 справа вгорі) якраз і буде відповідати граничному значенню для прогнозу в 565 балів.[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/8/files/21_05.jpg)

Рис. 4.5. Звіт Prediction Calculator Report

У нижній частині звіту Prediction Calculator зліва розташовується таблиця з відносними вагами значень даних параметрів (її ми вже зустрічали в таблиці Prediction Calculator) і графіком втрат від помилкових прогнозів.

Нехай продаж велосипеда приносить магазину не 10, а 50 доларів. В цьому випадку, прибуток від одного продажу буде перекривати витрати на супровід до 5 відмовилися від покупки клієнтів. Відповідно зміниться і співвідношення прибутку / витрати. На рис. 4.6 показано, що в цьому випадку, для максимізації прибутку рекомендується встановити граничне значення для прогнозу в 443 бали. Нове значення буде автоматично підставлений і в таблицю Prediction Calculator.

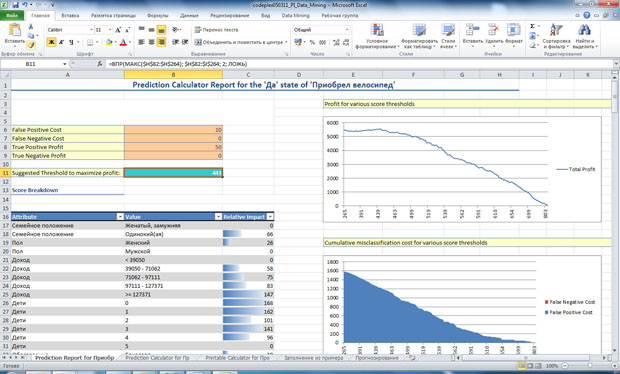
Завдання. Проведіть аналіз для двох різних наборів значень прибутку від справжніх прогнозів і збитків від помилкових. Прокоментуйте результати.[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/8/files/21_06.jpg)

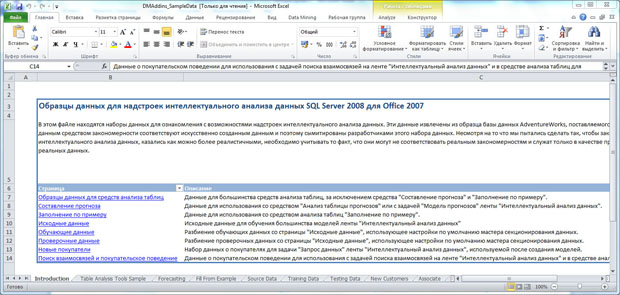
Рис. 4.6. Отчет Prediction Calculator Report: при вводе новой оценки прибыли от правильного прогноза меняется рекомендуемое пороговое значение

**Аналіз купівельної корзини**

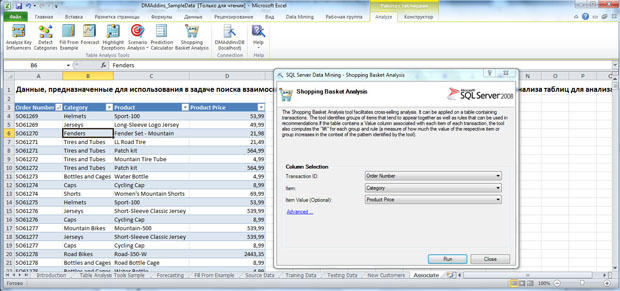
У наборі Table Analysis Tools нам залишилося розглянути інструмент Shopping Basket Analysis. Він дозволяє, наприклад, на основі даних про покупки виділити товари, що найчастіше зустрічаються в одному замовленні, і сформувати рекомендації щодо спільних продажів.

У процесі аналізу використовується алгоритм MicrosoftAssociationRules.

Для вивчення цього інструменту, замість використаного раніше локалізованого набору даних, звернемося до прикладу з поставки надбудов інтелектуального аналізу (в попередньому файлі потрібного набору даних просто немає). Через меню "Пуск" знайдіть "Надбудови інтелектуального аналізу даних" -> "Зразки даних Excel". У цій книзі Excel c першого аркуша (рис. 4.7) перейдіть за посиланням "Пошук взаємозв'язків і купівельну поведінку". Відповідний набір даних (рис. 4.8) містить інформацію про замовлення (номер замовлення - Order Number), включених в них товари (їх категорії - Category і власне товар - Product) і цінах

[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/8/files/21_07.jpg)

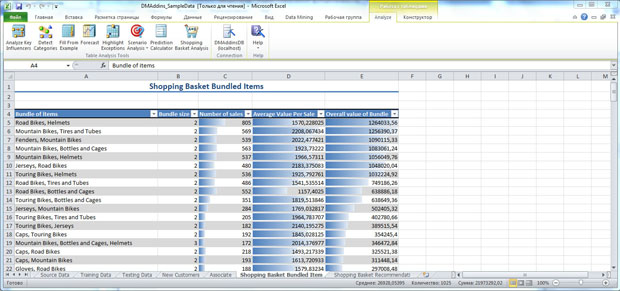
**Рис. 4.7.** Образцы данных

[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/8/files/21_08.jpg)

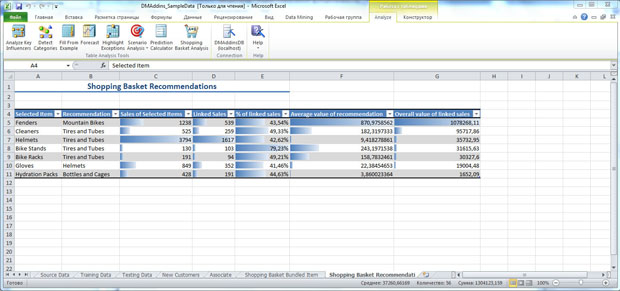
**Рис. 4.8.** Запуск инструмента Shopping Basket Analysis

Проаналізуємо, які категорії найчастіше потрапляють в одне замовлення. Запустимо інструмент Shopping Basket Analysis. У його настройках треба вказати ідентифікатор транзакції (TransactionID), в нашому випадку, це Order Number і предмет аналізу (ми будемо проводити аналіз для категорій - Category). Необов'язковим параметром, кількісно характеризує предмет аналізу (Item Value), в нашому випадку буде ціна. Якщо Item Value не вказано, то аналізуватися буде тільки частота виявлених сполучень.

Результати роботи Shopping Basket Analysis відображаються в двох звітах - Bundled Items і Recommendations.

[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/8/files/21_09.jpg)

**Рис. 4.9.** Отчет Shopping Basket Bundled Items

[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/8/files/21_10.jpg)

**Рис. 4.10.** Отчет Shopping Basket Recommendations

Перший із звітів містить інформацію про найбільш часто зустрічаються в "одному чеку" поєднаннях категорій товарів. Так, наприклад, в першому рядку звіту на рис. 4.9 ми бачимо, що найчастіше зустрічається поєднання категорій "дорожні велосипеди" і "шоломи" (RoadBikes, Helmets). В даному наборі воно виявлено в 805 замовленнях. Далі вказується середня ціна набору і сумарна вартість всіх подібних наборів. Можна сказати, що цей звіт описує купівельні шаблони клієнтів.

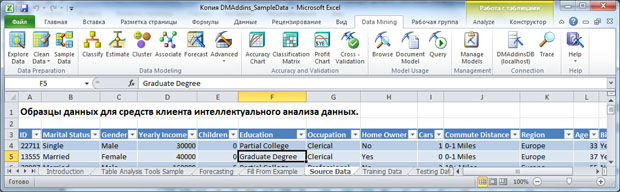
Другий звіт Shopping Basket Recommendations містить рекомендації про товари, які можуть бути запропоновані разом. Наприклад, третя сходинка звіту вказує, що людям купили шолом, варто також запропонувати придбати шини. Цей висновок базується на тому, що серед 3794 покупок включають шоломи, 1617 включали і шини. Частка таких пов'язаних продажів дорівнює 42,62%. Далі наводиться середній дохід від пов'язаних продажів (загальна вартість, поділена на число транзакцій, які містять "рекомендує" продукт, в нашому випадку - шолом) і загальна сума пов'язаних продажів. Грунтуючись на подібному звіті, власник магазину може вирішити, як розмістити товари, які пов'язані пропозиції можна сформувати і т.д.

Для видалення результатів роботи інструменту досить видалити сформовані звіти.

**Самостійна робота 5: Використання інструментів Data Mining Client для Excel 2007 для підготовки даних**

Розглянуті в попередніх лабораторних роботах "Засоби аналізу таблиць для Excel" (TableAnalysisTools) дозволяють швидко провести "стандартний" аналіз наявних даних. У той же час, цей набір інструментів не надає особливих можливостей з підготовки даних до аналізу, оцінці результатів і т.д. З Excel це можна зробити, використовуючи клієнт інтелектуального аналізу даних (DataMiningClient), який також входить в набір надбудов інтелектуального аналізу. В ході "Надбудови інтелектуального аналізу даних для MicrosoftOffice", зазначалося, що бажано зробити повну установку надбудов, в яку входить і DataMiningClient.

Відкриємо вже використовувався нами набір даних, що входить в поставку надбудов (меню "Пуск", знайдіть Надбудови інтелектуального аналізу даних-> Зразки даних Excel) .Щоб можна було спокійно вносити зміни, краще зберегти його під новим іменем.Перейдіте на лист "Вихідні дані" (SourceData) і клацніть на закладці DataMining. Стрічка з пропонованими інструментами представлена ​​на рис. 13.1.

[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/13/files/22_01.jpg)

**Рис. 13.1.** Інструменти Data Mining Client

Перша група інструментів (Data Preparation - Підготовка даних), дозволяє провести перше знайомство з набором даних і підготувати його для подальшого аналізу.

Наприклад, в попередніх роботах ми неодноразово стикалися з тим, що ряд алгоритмів (MicrosoftNaiveBayes і ін.) Вимагають попередньої дискретизації безперервних значень числових параметрів. Але в ряді випадків користувачеві бажано подивитися можливі діапазони, уточнити їх число і т.д. Окремий інтерес може представляти і розподіл рядків за значенням обраного параметра.

**Explore Data**

Інструмент Explore Data дозволяє проаналізувати значення стовпця (або діапазону комірок) і відобразити їх на діаграмі. Розглянемо його роботу на прикладі значення річного доходу клієнта (Income). Додатковий інтерес становить те, що це значення може розглядатися і як безперервне, і як дискретне. Отже, запускаємо інструмент (рис. 13.2).

В процесі роботи буде потрібно вказати, для якої таблиці (або діапазону комірок) і стовпця буде проводитися аналіз (рис. 13.2-1 і рис. 13.2-2). Після чого зазначені значення будуть проаналізовані і результат представлений у вигляді гістограми.

Як вже зазначалося вище, значення річного доходу можна розглядати і як безперервне, і як дискретне (за рахунок того, що в нашому наборі даних присутні тільки значення, кратні 10 тисячам). Для безперервного значення буде запропонований варіант розбиття на діапазони (рис. 13.2-3). Число діапазонів можна поміняти і діаграма з розподілом значень будуть побудована заново. Натиснувши кнопку "Add New Column" можна додати в вихідну таблицю новий стовпець з інтервалами річного доходу. Наприклад, якщо для рядка значення Yearly Income = 30000, то значення нового параметра Yearly Income 2 при використанні представленого на малюнку розбиття буде " '30000 - 50000" (саме так, з апострофом на початку, щоб розглядалося як строкове). В ході інтелектуального аналізу, отриманий стовпець може використовуватися замість справжнього (включення обох стовпців одночасно небажано).

Кнопками із зображеннями графіка і гістограми (на рис. 13.2-3, рис. 13.2-4 вони підкреслені), можна вказати тип аналізованого значення - безперервне або дискретне. Якщо значення річного доходу розглядаємо як дискретне, то для нього буде побудована діаграма, що показує розподіл числа рядків за значенням річного доходу (рис. 13.2-4). При цьому сортування проводиться по спадаючій числа рядків з даних значенням, через що перший стовпець гістограми відповідає значенню "60000", другий - "40000" і т.д. Сформовану гистограмму можна скопіювати в буфер (кнопка правіше кнопки "Add New Column", рис. 13.2-3, рис. 13.2-4) і використовувати для подальшої роботи.

**Clean Data**

Інструмент Clean Data (рис. 13.3) дозволяє підготувати дані для аналізу, відкинувши нетипові або помилкові дані (викиди), а також провівши заміну окремих значень. Як зазначається в документації, під викидом мається на увазі значення даних, що є проблематичним по одній з наступних причин:

значення знаходиться за межами очікуваного діапазону;

дані були введені неправильно;

значення відсутній;

дані представляють собою пробіл або порожній рядок;

значення може значно відхилитися від розподілу, якому підпорядковуються дані в моделі.

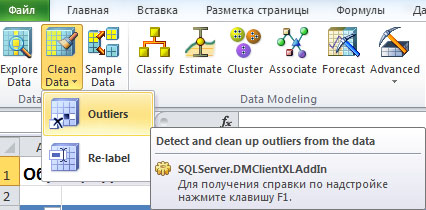


Рис. 13.3. інструмент CleanData

Використання даного інструменту проілюструємо на прикладі все тієї ж таблиці з даними про клієнтів (лист Source Data). Звернемося до колонку з віком. Нехай нам потрібно очистити набір даних від інформації про нехарактерних за віком покупців. Запускаємо інструмент Clean Data-> Outliers, у вікні аналогічному представленому на рис. 13.2-1 вибираємо таблицю для аналізу, потім у вікні Select Column (рис. 13.2-2) - стовпець Age.

В даному наборі даних є рядки зі значеннями стовпчика Age від 25 до 96 років. Якщо цей параметр вважаємо безперервним, то він буде представлений графіком, де по осі X вказується вік, по осі Y-число клієнтів з таким віком. У наборі даних частка клієнтів похилого віку дуже мала. На рис. 13.4-1 показано, що встановивши порогове значення в 75 років, ми відкидаємо заштрихований "хвіст", що включає нехарактерні значення (покупці велосипедів у віці від 76 до 96 років, яких переважна меншість).

Багато в чому аналогічно виглядає робота з параметром, які приймають дискретні значення. Для нього будується гістограма, а для визначення порогу потрібно вказати мінімальне число прикладів, "підтримують" значення. Наприклад, на рис. 13.4-2, встановлено граничне значення в 15. На жаль, при великому числі стовпців гістограми, значення параметра на ній не відображаються. Тому не зрозуміти, що саме потрапляє в "хвіст" розподілу.

Отже, ми виділили нехарактерні дані. Тепер потрібно визначити, що з ними робити. Пропоновані майстром рішення дещо відрізняються для випадків безперервного і дискретного параметра. Відповідний рядок можна видалити (Delete rows containing outliners) або замінити значення параметра на порожнє (Change value to null). Крім того, для безперервних даних (рис. 13.2-3) можна замінити нехарактерне значення середнім або граничним (зверху чи знизу, в залежності від того, який діапазон відкидається). Для дискретного параметра (рис. 13.2-4) можна вказати значення (з числа вже наявних в наборі), на яке будуть замінюватися "викиди".

Останнє вікно майстра (воно на малюнку не представлено) пропонує вибрати, куди заносити зміни - в вихідні дані (Change data inplace), в їх копію на новому аркуші Excel (Copy sheet data with changes to a new work sheet) або в новий стовпець в вихідної таблиці (Add as a new column to the current work sheet) .Останній опція для випадку видалення рядків недоступна.

CleanData.Re-label

У деяких випадках у вихідних даних можуть бути значення, які ускладнюють автоматизований аналіз. Наприклад, є параметр "місто" і серед його значень - Санкт-Петербург, С-Петербург, СПб. Для того, щоб в процесі інтелектуального аналізу цих значень враховувалися коректно, треба їх замінити на одне. Для цього можна використовувати інструмент Re-label. Його ж можна застосувати, якщо потрібно знизити рівень деталізації значень параметра. Треба відзначити, що інструмент працює тільки з дискретними значеннями (ну або розглядаються як дискретні).

Для прикладу, в таблиці з інформацією про клієнтів нам треба зменшити число значень параметра CommuteDistance (відстань щоденних поїздок). Вихідні значення "0-1 Miles", "1-2 Miles", "2-5 Miles", "5-10 Miles", "10+ Miles". Нехай все, що менше 2 миль, буде "близько", решта - "далеко". Додамо в таблицю два порожні рядки і вкажемо для однієї CommuteDistance "близько", для іншого - "далеко". Робиться це тому, що значення, на які замінюємо, теж повинні бути присутніми в стовпці.

Запустимо інструмент: CleanData-> Re-label. Перші два екрани, як і раніше, дозволяють вказати таблицю і стовпець. Далі вказуємо порядок заміни (рис. 13.5-1) і вибираємо створення нового стовпчика (рис. 13.5-2), щоб не втратити вихідні дані. Заміна буде проведена, після чого не забудемо видалити додані порожні рядки з "близько" - "далеко".

## **SampleData**

Останній інструмент в групі Data Preparation називається Sample Data (Зразки даних). Він дозволяє вирішити задачу формування навчає і тестового множин даних, а також виконувати "балансування" даних.

У тих випадках, коли використовується метод інтелектуального аналізу вимагає попереднього навчання моделі (наприклад, для вирішення задачі класифікації) необхідно сформувати кілька наборів даних - для навчання моделі, перевірки її роботи, власне аналізу. Інструмент Sample Data дозволяє підготувати потрібні набори.

Нехай необхідно випадковим чином розділити наявний набір даних на навчальну і тестову вибірку. Для цього треба запустити інструмент Sample Data, вказати звідки беремо дані для обробки (рис. 13.6-1) і тип формується вибірки. Спочатку зробимо випадкову вибірку, тобто тип - Random Sampling (рис. 13.6-2). Далі вказується відсоток записів з вихідного набору (або точне число записів) розміщені в вибірку (рис. 13.6-3) і місце для збереження отриманих результатів. На рис. 13.6-4 видно, що можна окремо зберегти сформовану вибірку і дані, в неї не потрапили. У підсумку можемо отримати навчальний і тестовий набори. Хотілося б звернути увагу на можливість використання зовнішнього джерела даних при формуванні вибірки (рис. 13.6-1). Це дозволяє використовувати дані зберігаються на MS SQLServer для формування наборів значень. Але як зазначається в описі інструменту, при використанні зовнішнього джерела даних у вікні, представленому на рис. 13.2, буде доступний тільки параметр випадкової вибірки.

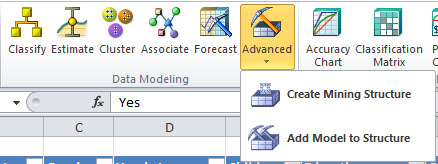
При використанні коштів інтелектуального аналізу для виявлення рідкісних подій, в навчальному наборі рекомендується збільшити частоту появи потрібного події в порівнянні з вихідними даними. Формування подібної вибірки часто називають балансуванням даних, і інструмент SampleData дозволяє її виконати.

За допомогою інструменту Explore Data проаналізуємо розподіл клієнтів в наборі даних по регіонах. На рис. 13.7-1 видно, що приблизно п'ята частина клієнтів у нас з регіону Pacific (будемо вважати це Азіатсько-Тихоокеанським регіоном). Сформуємо набір даних, де таких клієнтів буде 50%.

Запустимо інструмент Sample Data, вкажемо в якості джерела даних використовувану таблицю Excel і виберемо варіант формування надлишкової вибірки з балансуванням даних (Oversample to balance data distributions, рис. 13.7-2). Далі вкажемо стовпець, для якого виконується балансування, і частоту появи потрібного значення і розмір вибірки (рис. 13.7-3) .Буде створена нова таблиця з зазначеним користувачем назвою. Знову застосуємо Explore Data і переконаємося в тому, що вибірка сформована відповідно до зазначених вище вимог (рис. 13.7-4).

**Самостійна робота 7: Використання інструментів Data Mining Client для Excel 2007 для створення моделі інтелектуального аналізу даних**

Розглянуті в лабораторних роботах "Надбудови інтелектуального аналізу даних для MicrosoftOffice" - "Використання інструментів" Prediction Calculator "і" ShoppingbasketAnalysis "" "Засоби аналізу таблиць для Excel" (TableAnalysisTools) для кінцевого користувача багато в чому представляються "чорним ящиком", який виконує аналіз, але що не дає інформації про те, як отриманий результат. Якщо таке рішення не влаштовує, можна перейти з вкладки Analyze на вкладку DataMining і скористатися інструментами DataMiningClient для Excel (рис. 14.1).

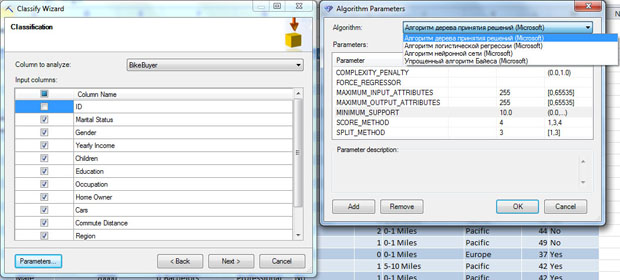


**Рис. 14.1.** Група інструментів DataModeling

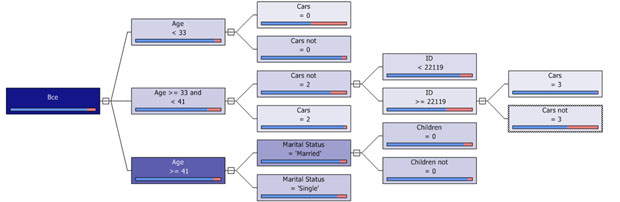
В "Використання інструментів Data Mining Client для Excel 2007 для підготовки даних" ми розглянули інструменти, що дозволяють підготувати дані для аналізу. Наступна група показані на рис. 14.1 інструменти DataModeling, що дозволяють створити моделі інтелектуального аналізу даних.

Використовуємо інструмент Classify. У поставляється з надбудовами наборі даних (меню "Пуск" -> "Надбудови інтелектуального аналізу даних" -> "Зразки даних Excel") виберемо таблицю TrainingData, що містить випадкову вибірку 70% даних з таблиці SourceData. Запустимо майстер Classify, в першому вікні якого буде коментар щодо застосування інструменту, а друге вікно дозволить вказати джерело даних для аналізу (таблиця TrainingData). Далі буде потрібно описати мета аналізу.

Нехай нас цікавить, чи зробить даний клієнт покупку. У цільовому стовпці вказуємо параметр BikeBuyer (рис. 14.2, вікно зліва), скидаємо в переліку вхідних стовпців позначки навпроти ID (порядковий номер клієнта в базі ніяк не впливає на його рішення про покупку). Якщо ID залишити серед аналізованих параметрів, то підсумкова модель може його врахувати. Зокрема, на рис. 14.3 показано дерево рішень, що враховує значення поля ID в процесі класифікації, що однозначно неправильно.

[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/14/files/23_02.jpg)

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

[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/14/files/23_03.jpg)

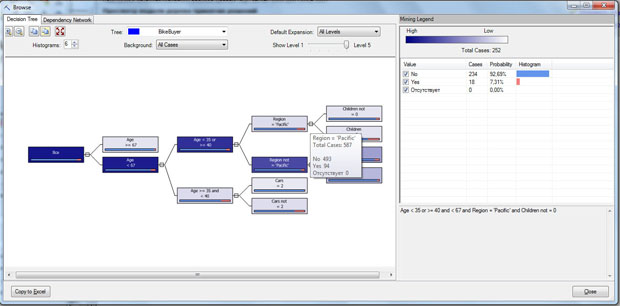
**Рис. 14.3.** Невдалий варіант дерева рішень

Якщо потрібно більш точне налаштування, можна відкрити вікно Parameters і явно вказати використовуваний алгоритм і його параметри (рис. 14.2, вікно праворуч). Далі майстер запропонує розділити наявні дані на набір для навчання моделі і для її тестування (рис. 14.4-1). За замовчуванням на набір для тестування виділяється 30% рядків вихідного набору.

Останній етап роботи майстра - вказівка ​​імені створюваної структури і моделі (рис. 14.4-2). У нашому прикладі структура буде назватися TrainingDataStructure, а модель Classify BikeBuyer\_1. Ці назви нам знадобляться згодом для роботи з моделлю.

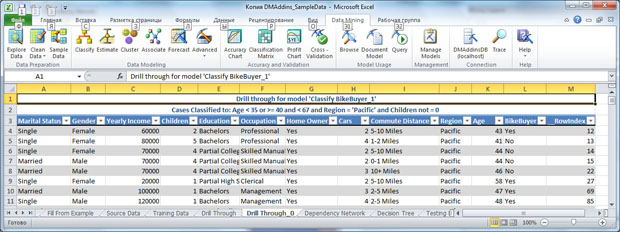
Якщо виконує аналіз користувач не має прав адміністратора в базі Аналітичних Служб (цей параметр ми робили в "Надбудови інтелектуального аналізу даних для MicrosoftOffice"), то створити постійною модель інтелектуального аналізу на сервері він не зможе. В цьому випадку можна використовувати тимчасову модель, для чого поставити галочку навпроти пункту Use temporary model.Временная модель буде автоматично видалена з сервера по завершенню сеансу роботи користувача.

Зазначена за замовчуванням настройка Brose model вказує на те, що після створення моделі буде відкрито вікно перегляду. Для моделі, створеної з використанням алгоритму DecisionTrees, відображається побудоване дерево рішень і діаграма залежностей. Представлене на рис. 14.5 дерево рішень дозволяє оцінити побудовану модель. Розташовані у верхній частині екрану "повзунок" і випадає дозволяють встановити число відображуваних рівнів дерева (на малюнку показані всі п'ять). Якщо навести курсор миші на точку розгалуження, можна побачити підказку із зазначенням того, скільки і яких випадків в навчальному наборі їй відповідає. Для виділеного вузла в правій частині екрану відображається його опис та гістограма з розподілом значень. Кнопкою Copy to Excel можна перенести результат з вікна перегляду на новий лист Excel (для дерева рішень в Excel буде перенесено його растрове зображення)

.[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/14/files/23_05.jpg)

**Рис. 14.5.** . Побудоване дерево рішень

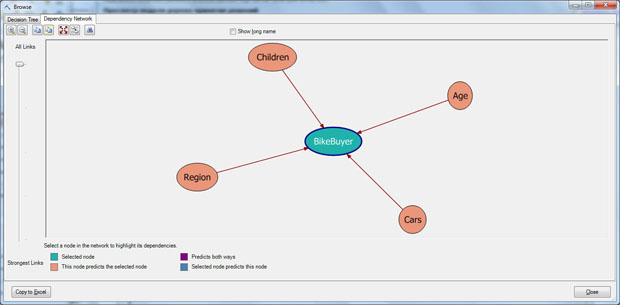
Клацнувши по вузлу дерева правою клавішею миші і вибравши в контекстному меню DrillThroughModelColumns (можна приблизно перекласти як "деталізація використовувалися моделлю даних") ми отримаємо нову таблицю Excel, що містить набір рядків з навчальної вибірки, які відповідають даному вузлу (рис. 14.6).

[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/14/files/23_06.jpg)

**Рис. 14.6.** Результат виконання DrillThroughModelColumns

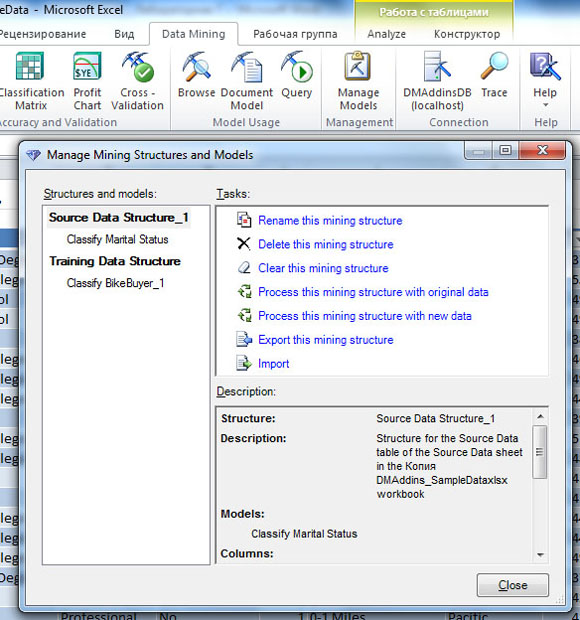
На рис. 14.7 представлена діаграма залежностей, що показує виявлені взаємозв'язку між параметрамі.Ее також можна скопіювати в Excel.Виделяя на діаграмі вузол, можна побачити все що впливають на нього.

Закриємо вікно перегляду моделі. Якщо потрібно буде знову переглянути її параметри, скористайтеся інструментом Browse, який знаходиться в групіmodelUsage.

[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/14/files/23_07.jpg)

**Рис. 14.7.** діаграма залежностей

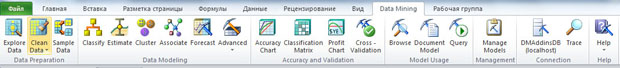
Для того щоб управляти наявними на сервері структурами і моделями інтелектуального аналізу, можна скористатися відповідним майстром, запускаються після натискання кнопки ManageModels на вкладці DataMining (рис. 14.8). Він дозволяє переглянути наявні структури і моделі, перейменувати їх, видалити непотрібні, виконати інші дії на сервері прямо з DataMiningClient.

[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/14/files/23_08.jpg)

**Рис. 14.8.** Окно управления моделями

**Самостійна робота 8: Аналіз точності прогнозу і використання моделі інтелектуального аналізу**

У попередній лабораторній роботі ми створили модель для класифікації клієнтів магазину з метою визначити, чи зробить даний клієнт покупку чи ні. Наступне завдання - оцінити точність побудованої моделі інтелектуального аналізу. Для цього можна використовувати інструменти з групи Accuracy and Validation (в російській варіанті - Точність і Правильність).

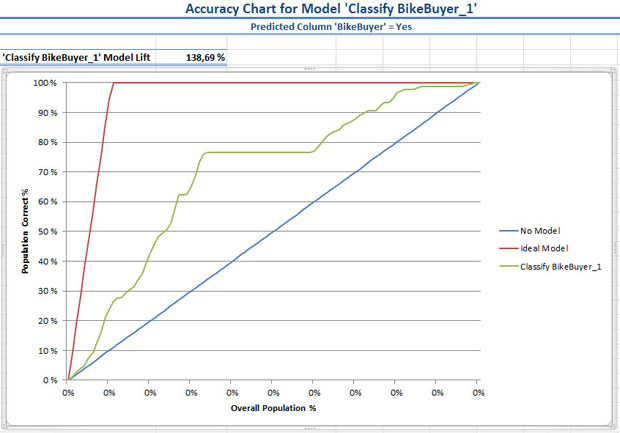
[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/15/files/24_01.jpg)

**Рис. 15.1.** Інструменти DataMiningClient

Діаграма точності (AccuracyChart) дозволяє, застосувавши модель на тестовій вибірці даних, оцінити результати її роботи. В ході виконання "Використання інструментів Data Mining Client для Excel 2007 для створення моделі інтелектуального аналізу даних» була створена структура TrainingDataStructure і модель класифікації Classify BikeBuyer\_1. При створенні моделі ми резервували 30% даних для цілей тестування (рис. 15.4-1 в "Використання інструментів Data Mining Client для Excel 2007 для створення моделі інтелектуального аналізу даних »).

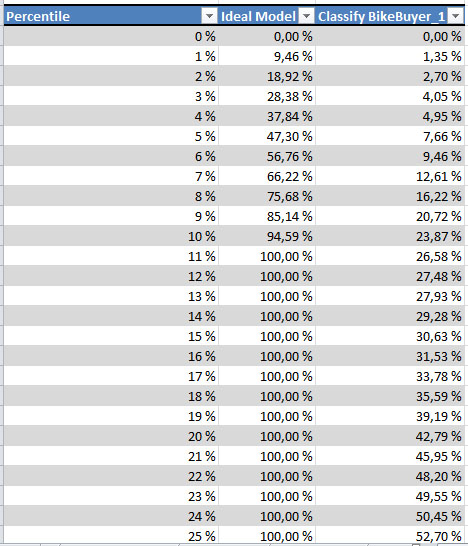
Запустимо інструмент AccuracyChart.Первое вікно майстра містить короткий опис інструменту, в наступному - треба вказати структуру або модель (рис. 15.2-1). Якщо для однієї структури визначені кілька моделей, по діаграмі можна буде провести їх порівняльний аналіз. Наступне вікно (рис. 15.2-2) служить для вибору пророкує параметра і його значення. У нашому випадку параметр - BikeBuyer, а оцінювати будемо точність передбачення значення "Yes". Далі потрібно вказати джерело даних для тестування. Це можуть бути зарезервовані при створенні моделі дані, дані з таблиці або діапазону комірок Excel, або зовнішнього джерела (рис. 15.2-3). Зараз виберемо дані з моделі. У разі зазначення таблиці Excel (що будемо робити в вправах), треба описати відповідність стовпців в моделі і використовуваної для тестування таблиці (рис. 15.2-4). Після цього будуть сформовані і поміщені на новий лист Excel діаграма точності (рис. 15.3) і таблиця зі значеннями, представленими на діаграмі (рис. 15.4).

На діаграмі червона лінія відповідає ідеальної моделі, світло-зелений - нашої моделі, нижня (синя) лінія - лінія випадкового вибору, завжди йде під кутом 45 градусів.

[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/15/files/24_03.jpg)

**Рис. 15.3.** Діаграма точності (AccuracyChart)

Дані на діаграмі і в таблиці можна інтерпретувати в такий спосіб. Нехай нам необхідно вибрати всіх клієнтів, які зроблять покупки. Формована ідеальною моделлю вибірка обсягом в 11% від числа вихідних записів буде включати всі 100% потрібних записів (в тестовому безлічі їх мабуть трохи менше 11%). Випадкова вибірка обсягом в 11% містить 11% потрібних записів, а вибірка такого ж обсягу, яка формується нашою моделлю - 26,58%. До вибірки в 25% від загального обсягу даних, наша модель помістить 52,7% "правильних" клієнтів і т.д. Якість прогнозу падає (горизонтальну ділянку зеленого графіка) після виявлення 76% цікавлять випадків. При візуальному аналізі - чим ближче графік оцінюваної моделі до ідеального, тим точніший прогноз вона видає.

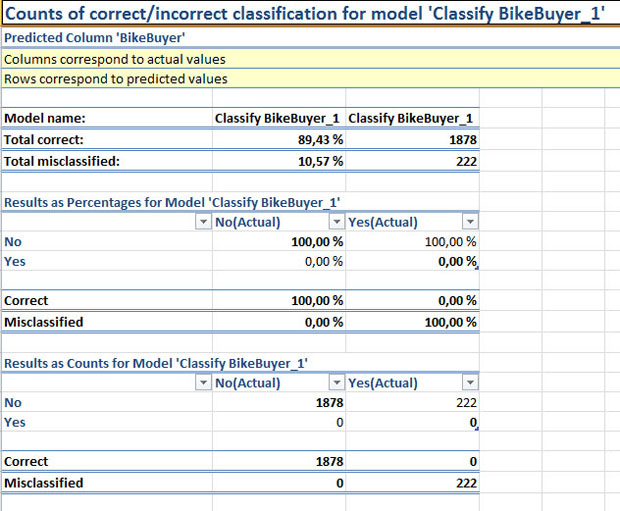


**Рисунок. 15.4.** Фрагмент таблиці з оцінками точності прогнозу

Завдання 1. Побудувати діаграму точності аналогічну тій, що представлена ​​вище (використовуваний файл - "Зразки даних Excel"). Додатково побудуйте діаграму для BikeBuyer = "No". Поясніть відмінність у зовнішньому вигляді графіків.

Завдання 2. У попередньому завданні для цілей тестування використовувалися дані з моделі. Модель формувалася на даних з таблиці TrainingData. У таблиці TestingData знаходяться 30% даних з вихідного набору SourceData. Перевірте точність моделі на наборі TestingData.

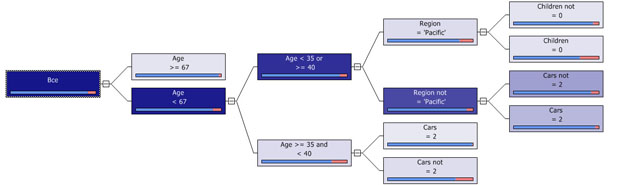
Аналізуючи графік на рис. 15.3 можна припустити, що у нас з моделлю все добре. Але звернемося до ще одного інструменту аналізу точності - ClassificationMatrix (Матриця класифікації). З його допомогою ми отримуємо таблицю з результатами точних і помилкових передбачень (рис. 15.5). З неї видно, що створена нами модель при тестуванні на зарезервованих даних зробила 89,43% правильних прогнозів, що можна розцінити як успіх (тому й діаграма точності на рис. 15.3 виглядає добре) .Але при цьому в 100% випадків правильно передбачала значення " No "і помилково" Yes ". Інакше кажучи, у всіх випадках 100% ставиться "No". І використовувати таку модель для передбачення безглуздо.



**Рисунок. 15.5.** матриця класифікації

Завдання 3. Побудувати матрицю класифікації, проаналізуйте отриманий результат.

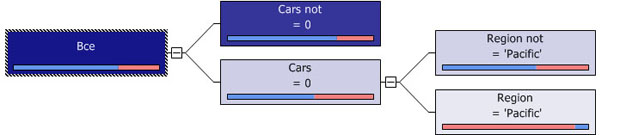
Проблема, з якою ми зіткнулися, могла бути виявлена і раніше, якщо уважно подивитися на побудоване дерево рішень (рис. 15.6). Але тоді не вдалося б продемонструвати можливості DataMining за оцінкою точності моделі.

[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/15/files/24_06.jpg)

**Рисунок 15.6.** дерево рішень

З рис. 15.6 видно, що всі кінцеві вузли дерева дають рішення BikeBuyer = "No" (йому відповідає синя смужка на діаграмі характеризує розподіл відповідей в навчальній вибірці). Відповіді "Yes" відповідає більш коротка червона смужка, що говорить про те, що підтримують такий результат прикладів було менше. По всій видимості, це пов'язано з тим, що таких прикладів і взагалі меншість в даному наборі (близько 10%).

Спробуємо використовувати навчальний набір більшого обсягу і з більш часто зустрічається значенням BikeBuyer = "Yes". Відкриємо таблицю SourceData, де даних більше. Але відсоток цікавлять нас записів залишається таким же (це можна визначити за допомогою інструменту ExploreData) .Тому скористаємося інструментом SampleData ( "Використання інструментів Data Mining Client для Excel 2007 для підготовки даних"), щоб сформувати "надлишкову" вибірку з 2000 року рядків, де в 30% випадків BikeBuyer = "Yes". У отриманого набору залишимо автоматично призначений назву SampledData. За допомогою інструменту Classify побудуємо модель аналогічна тій, як це було зроблено в "Використання інструментів Data Mining Client для Excel 2007 для створення моделі інтелектуального аналізу даних" (алгоритм - DecisionTrees, цільової параметр BikeBuyer, стовпець ID при аналізі не враховуємо, інші настройки за замовчуванням ). Отримане дерево рішень представлено на рис. 15.7. Воно простіше попереднього, але в залежності від значень параметрів може давати як прогноз "Yes", так і "No". "Yes" буде в тому випадку, якщо у клієнта 0 машин і він з регіону "Pacific".

[](http://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/15/files/24_07.jpg)

**Рисунок 15.7.** Нове дерево рішень

На основі нового набору даних також створимо модель для класифікації, засновану на алгоритмі NeuralNetworks (нейронних мереж). Якщо побудувати для них матриці класифікації (рис. 15.8-1, рис. 15.8-2) буде видно, що модель на основі нейронних мереж дає більш точний прогноз. Розглянутий приклад показує, що в деяких випадках точність прогнозу можна підвищити за рахунок спеціальної підготовки навчальної вибірки і вибору найбільш підходящого алгоритму. Хоча, з огляду на відносно високий відсоток помилок, ні та, ні інша модель, напевно, не може бути визнана вдалою.

**Висновок**

В ході виконання роботи ми навчилися використанню технології інтелектуального аналізу даних (Data Mining) і її реалізації в Microsoft SQL Server 2008 і пов'язаних програмних продуктах. Такожм, розгляданемо всі алгоритми DM, підтримувані Microsoft SQL Server 2008, надбудови інтелектуального аналізу даних для Microsoft Office, основи мови DMX.