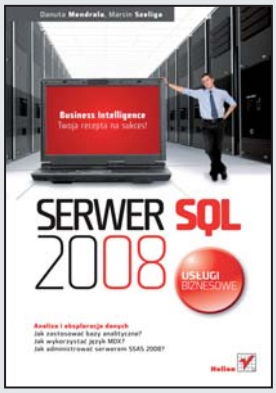
Serwer SQL 2008. Usługi biznesowe. Analiza i eksploracja danych



Autor: Danuta Mendrala, Marcin Szeliga
ISBN: 978-83-246-2111-8
Format: 158x235, stron: 360

Business Intelligence – Twoja recepta na sukces!

- Jak zastosować bazy analityczne?
- Jak wykorzystać język MDX?
- Jak administrować serwerem SSAS 2008?

Informacja jest bezcenna. Umiejętność jej właściwego wykorzystania i zarządzania nią również stanowi ogromną wartość. Autorzy Microsoft SQL Server, wydajnego serwera baz danych, zauważyli to już blisko dziesięć lat temu. To właśnie w Microsoft SQL Server 2000 zostały po raz pierwszy wprowadzone usługi analityczne. Dzięki możliwościom SQL Server 2008 również Ty możesz podejmować właściwe decyzje biznesowe i osiągnąć sukces!

Książka „Serwer SQL 2008. Usługi biznesowe. Analiza i eksploracja danych" jest długo oczekiwaną pozycją, której autorzy w sposób kompleksowy podejmują tematykę związaną z Business Intelligence. Dzięki niej zdobędziesz szczegółowe informacje na temat hurtowni danych, baz analitycznych oraz języka MDX. Ponadto dowiesz się, w jaki sposób administrować serwerem SSAS 2008 oraz jak wykorzystać arkusz kalkulacyjny Excel jako klienta baz analitycznych. Dodatkowo zapoznasz się z różnymi technikami eksploracji danych oraz sposobami tworzenia ich projektów. Jeżeli chcesz podejmować celne decyzje poprzez analizę danych, ta książka jest właśnie dla Ciebie!

- Hurtownie danych
- Projektowanie systemów Business Intelligence
- Modelowanie danych
- Bazy analityczne
- Serwery OLAP
- Serwer SSAS 2008
- Obiekty analitycznych baz danych
- Projektowanie i tworzenie baz analitycznych
- Wykorzystanie języka MDX
- Administrowanie serwerem SSAS 2008
- Wykorzystanie arkusza kalkulacyjnego Excel 2007 jako klienta baz analitycznych
- Zastosowania eksploracji danych
- Tworzenie projektów eksploracji danych
- Wykorzystanie dodatku Data Mining dla pakietu Office 2007

Zarządzaj informacjami tak, by przynosiły Ci korzyść

Helion SA
ul. Kościuszki 1c
44-100 Gliwice
tel. 032 230 98 63
e-mail: helion@helion.pl

Spis treści

Wstęp **7**

Część I Analiza danych **15**

Rozdział 1. Hurtownie danych **17**

Projektowanie systemów Business Intelligence 17
Uzgodnienie wymagań i założeń 17
Dostosowywanie projektu do zmieniających się wymagań użytkowników 19
Testy 20
Modelowanie danych 20
Hurtownie danych 21
Tabele faktów 21
Tabele wymiarów 25
Schematy 32
Obszar przejściowy 34
Indeksy 34
Partycjonowanie tabel faktów 35
Kompresja danych 37

Rozdział 2. Bazy analityczne **41**

Serwery OLAP 42
Kostki analityczne 42
Zunifikowany model wymiarowy 44
Serwer SSAS 2008 44
Funkcjonalność 44
Instalacja 45
Narzędzia 48
Interfejsy programistyczne 50
Podstawowe obiekty analitycznych baz danych 51
Praca z BIDS 51
Źródła danych 52
Widoki źródeł danych 53
Miary kostek analitycznych 54
Wymiary 56
Kostki analityczne, czyli miary plus wymiary 57

Rozdział 3. Tworzenie baz analitycznych 61
Założenia analitycznych baz danych 61
Projekt najprostszej analitycznej bazy danych 62
Źródła danych i ich widoki 65
Źródła danych 66
Widoki źródeł danych 68
Wymiary 72
Kreator wymiarów 72
Edytor wymiarów 74
Atrybuty 81
Hierarchie 85
Zapis zwrotny 96
Fizyczna struktura wymiarów 97
Kostki analityczne 99
Kreator kostek analitycznych 99
Edytor kostek analitycznych 100
Fizyczna struktura kostki 109
Zapis zwrotny 114
Projektowanie agregacji 116
Kreator agregacji 116
Przypisywanie zaprojektowanych agregacji 118
Samodzielne projektowanie agregacji 118
Kompilacja projektów 119
Techniki dystrybucji 120
Business Intelligence Development Studio 121
Analysis Services Deployment Wizard 122
Skrypty XMLA 124

Rozdział 4. Wprowadzenie do języka MDX 125
Konwencje 125
Krotki 126
Wyrażenia MDX 127
Operatory 127
Funkcje 128
Zapytania MDX 133
Kreator zapytań 133
Składnia instrukcji SELECT 135
Zbiory 137
Wykonywanie instrukcji MDX przez serwer SSAS 142

Rozdział 5. Dodatkowe funkcje kostek analitycznych 145
Wartości wyliczeniowe 145
Tworzenie wartości wyliczeniowych 146
Zalecenia 152
Zbiory nazwane 152
Nazywanie zbiorów krotek 152
Statyczne i dynamiczne zbiory nazwane 153
Zalecenia 155
Skrypty MDX 155
Modyfikowanie fragmentów kostki analitycznej 155
Zalecenia 159
Kluczowe wskaźniki efektywności 159
Elementy kluczowych wskaźników efektywności 159
Przeglądanie kluczowych wskaźników efektywności 163
Zalecenia 164

Akcje ... 164
Typy akcji ... 165
Zalecenia ... 168
Procedury składowane ... 168
Rejestrowanie zewnętrznych bibliotek ... 169
Wywoływanie procedur składowanych ... 170
Zalecenia ... 171

Rozdział 6. Administracja serwerem SSAS 2008 ... 173
Oszacowanie wymagań serwera SSAS 2008 ... 173
Pamięć i procesor ... 174
Dysk twardy ... 175
Aktualizowanie projektów baz analitycznych ... 177
Synchronizacja analitycznych baz danych ... 177
Przetwarzanie kostek i wymiarów ... 179
Tworzenie i odtwarzanie kopii zapasowych ... 185
Bezpieczeństwo ... 189
Role ... 189
Uprawnienia ... 190
Zalecenia ... 195
Monitorowanie pracy serwera ... 196
Dzienniki ... 197
Monitor wydajności ... 197
SQL Server Profiler ... 201
Widoki dynamiczne ... 203
Optymalizacja wydajności ... 204
Optymalizacja agregacji ... 204
Aktywny mechanizm buforowania ... 207
Skalowalność i dostępność ... 209
Równoważenie obciążania ... 209
Współdzielone bazy analityczne ... 211
Automatyzacja zadań administracyjnych ... 212
Skrypty XMLA ... 212
Zadania usługi SQL Server Agent ... 213
Pakiety SSIS ... 215

Rozdział 7. Excel 2007 jako klient baz analitycznych ... 217
Biznesowa analiza danych ... 217
Analiza danych przy użyciu programu Excel 2007 ... 218
Zewnętrzne źródła danych ... 219
Tabele przestawne ... 221
Wykresy przestawne ... 229
Formuły kostek analitycznych ... 232
Usługi programu Excel ... 237

Część II Eksploracja danych ... 239

Rozdział 8. Techniki eksploracji danych ... 241
Scenariusze biznesowe ... 241
Eksploracja danych jako część analizy biznesowej ... 242
Proces eksploracji danych ... 243
Zastosowania eksploracji danych ... 246
Klasyfikacja ... 246
Regresja ... 250
Segmentacja ... 252
Asocjacja ... 255

Analiza sekwencyjna ... 258
Prognozowanie ... 260
Serwer SQL 2008 ... 262
Integracja z usługami Business Intelligence ... 262

Rozdział 9. Tworzenie projektów eksploracji danych ... 265
Struktury eksploracji danych ... 265
Dane źródłowe ... 266
Dane treningowe (przypadki) ... 268
Przetwarzanie struktur eksploracji danych ... 274
Tworzenie struktur eksploracji danych w języku DMX ... 275
Modele eksploracji danych ... 279
Algorytmy ... 279
Dane treningowe ... 288
Tworzenie modeli eksploracji danych w języku DMX ... 291
Trening modeli ... 293
Ocena ... 296
Wykresy podniesienia i zysku ... 296
Macierz klasyfikacji ... 298
Walidacja krzyżowa ... 299
Odczytywanie wyników ... 302
Wizualizatory i odczytywanie dodatkowych danych ... 302
Zapytania predykcyjne ... 305

Rozdział 10. Dodatek Data Mining dla pakietu Office 2007 ... 309
Przygotowanie danych ... 310
Analiza danych ... 311
Oczyszczanie danych ... 312
Podział danych ... 314
Eksploracja danych tabelarycznych ... 315
Analiza kluczowych czynników ... 315
Kategoryzacja ... 318
Uzupełnianie na podstawie przykładu ... 321
Przewidywanie ... 322
Wykrywanie anomalii ... 324
Osiąganie celu ... 326
Analiza typu „Co będzie, jeśli?” ... 327
Ocena nowych przypadków ... 329
Analiza koszyka zakupów ... 331
Eksploracja zewnętrznych danych ... 333
Tworzenie, przeglądanie i zarządzanie modelami ... 334
Przewidywanie ... 335
Okresowość ... 338
Predykcje krzyżowe ... 339
Ocena ... 341

Skorowidz ... 343

Rozdział 8.

Techniki eksploracji danych

Eksplorację danych definiuje się jako *szeroką kategorię aplikacji i technologii do zbierania, przechowywania, analizowania i współużytkowania danych oraz zapewniania dostępu do nich w celu umożliwienia użytkownikom podejmowania lepszych decyzji biznesowych*. Wspomniane w tej definicji technologie korzystają z klasycznych metod statystyki i probabilistyki w celu zautomatyzowania analizy przechowywanych w bazach (zarówno relacyjnych, jak i analitycznych) dużych ilości informacji. Większość używanych w procesie eksploracji danych algorytmów jest dość nowa, ale ich skuteczność została już potwierdzona, zarówno teoretycznie (poprzez badania nad teorią baz danych), jak i praktycznie (instytucje finansowe oraz duże korporacje od wielu lat wykorzystują mniej lub bardziej zautomatyzowane techniki eksploracji danych).

Wyniki eksploracji danych, przede wszystkim predykcji (przewidywania zaistnienia pewnych zdarzeń lub zmian określonych wartości), zależą od jakości danych źródłowych. Jeżeli sytuacja ekonomiczna ulegnie radykalnej zmianie (jak to miało miejsce podczas rozpoczynającego się w czasie powstawania książki kryzysu gospodarczego), otrzymane na podstawie nieaktualnych danych wyniki będą niewiarygodne. Dlatego przeprowadzane w takich okresach prognozy (np. przyszłych cen ropy czy wzrostu produktu krajowego brutto) są dość przypadkowe. Po ustabilizowaniu się sytuacji i zebraniu wystarczającej ilości nowych, właściwie opisujących ją danych techniki eksploracji danych będą mogły ponownie dostarczać wartościowych i precyzyjnych wyników.

Scenariusze biznesowe

Współcześnie firmy dysponują dużymi zbiorami danych, ale mają coraz większy problem z ich praktycznym wykorzystaniem. **Zbyt dużo danych utrudnia ich analizę i skutkuje zmniejszeniem się ilości przydatnych informacji**. Techniki eksploracji danych rozwiązują ten problem, umożliwiając między innymi:

1. Przewidywanie utraty klientów — na podstawie historii zakupów oraz danych demograficznych możemy dokonać segmentacji klientów i określić, którzy z nich (i dlaczego) myślą o odejściu do konkurencji.
2. Wykrywanie nadużyć — bazując na historii użycia karty kredytowej, banki automatycznie oceniają ryzyko, że dana operacja nie jest autoryzowana przez ich posiadacza i w nietypowych przypadkach kontaktują się z klientem w celu ich potwierdzenia.
3. Wykrywanie oszustw i nieprawidłowości — zbierając dane o typowej aktywności użytkowników i porównując wyniki ich analizy z danymi z monitoringu, można wykryć nietypowe zachowania, w tym takie, które pozostałyby niewykryte przez tradycyjne systemy kontroli (na przykład fakt, że użytkownik, który do tej pory odczytywał z bazy tylko modyfikowane przez siebie dane dotyczące zamówień, odczytał z bazy komplet informacji o klientach firmy).
4. Budowanie skutecznych kampanii marketingowych — analizując dane demograficzne, można wybrać osoby, które będą prawdopodobnie zainteresowane otrzymywaniem informacji o promocjach i powinny być objęte akcją marketingową.
5. Ocenę ryzyka — dysponując danymi historycznymi i demograficznymi, można ocenić prawdopodobieństwo spłaty kredytu lub pożyczki przez daną osobę.
6. Przewidywanie sprzedaży — bazując na danych historycznych, można przewidzieć przyszłe wyniki sprzedaży danych produktów i odpowiednio wcześniej skorygować stan towarów w magazynach.
7. Zrozumienie potrzeb klientów — segmentacja klientów pozwala ocenić potrzeby każdej z grup i określić czynniki, którymi kieruje się dana grupa klientów, wybierając poszczególne produkty. Możliwe jest też znalezienie czynników, które mają największy wpływ na podejmowane przez nich decyzje.
8. Szukanie klientów przynoszących zyski — na podstawie danych osobowych można przewidzieć, która osoba (i z jakim prawdopodobieństwem) będzie dobrym klientem firmy.
9. Wyszukiwanie razem sprzedawanych towarów — każdy sprzedawca powinien wiedzieć, które towary często kupowane są razem, a które prawie nigdy nie trafiają do tego samego koszyka. Zdobycie tej wiedzy umożliwia analiza historii sprzedaży.

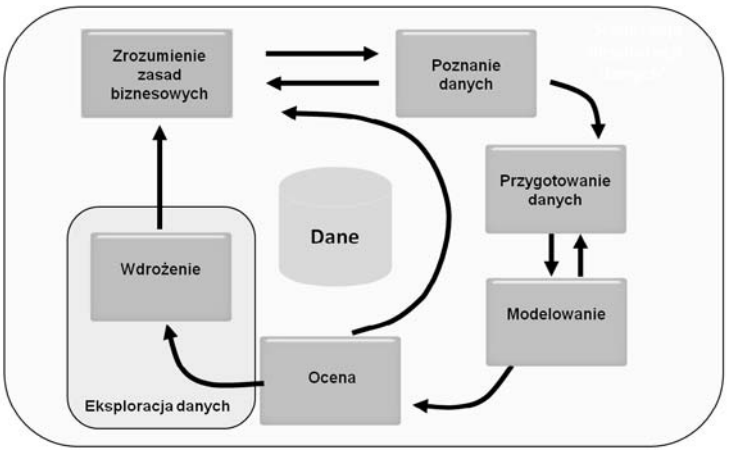
Eksploracja danych jako część analizy biznesowej

Eksploracja danych wymaga wykonania skomplikowanych obliczeń statystycznych, a następnie prawidłowego zinterpretowania otrzymanych w ten sposób wyników. Firma Microsoft, chcąc ułatwić przeprowadzanie analiz biznesowych z wykorzystaniem technik eksploracji danych:

1. Dodała do serwera SQL 2008[2] kompletny zestaw stosunkowo łatwych w użyciu narzędzi do eksploracji danych.
2. Umożliwiła eksplorację danych za pomocą wbudowanych algorytmów. Dzięki automatycznemu dostrajaniu i autoparametryzowaniu w wielu przypadkach nie wymagają one nawet zmiany domyślnych wartości parametrów.
3. Uzgodniła serwer SQL 2008 ze standardem PMML (ang. *Predictive Model Markup Language*), dzięki czemu tworzone w tym serwerze modele eksploracji danych są kompatybilne z rozwiązaniami firm trzecich, takich jak SAS, IBM czy Oracle.
4. Ułatwiła interpretację otrzymywanych wyników poprzez wbudowanie w konsolę SSMS i BIDS wizualizatorów.
5. Dołączyła do konsoli SSMS, BIDS oraz dodatku Data Mining dla pakietu Office 2007 zestaw narzędzi pozwalających ocenić dokładność i wiarygodność modeli eksploracji danych.
6. Umożliwiła użytkownikom, którzy nie mają doświadczenia w pracy z serwerem SQL, eksplorację danych za pomocą programu Excel 2007 i bezpłatnego dodatku Data Mining dla pakietu Office 2007[3].

Proces eksploracji danych

Proces eksploracji danych składa się z sześciu zdefiniowanych w ramach projektu CRISP-DM (ang. *Cross Industry Standard Process for Data Mining*) kroków (rysunek 8.1).



Rysunek 8.1. *Szczegółowe informacje na temat projektu CRISP-DM dostępne są pod adresem http://www.crisp-dm.org*

[2] Dotyczy to edycji Enterprise i częściowo edycji Standard.

[3] Excel 2007 jest tylko narzędziem klienckim serwera SQL 2008, a więc do eksploracji danych wymaga on połączenia z tym serwerem.

1. Pierwszym etapem procesu eksploracji danych powinno być dokładne zrozumienie zasad biznesowych. Pomocne w tym będzie otrzymanie od końcowego użytkownika (czyli właściciela firmy, kierownika działu marketingu czy pracownika kontrolingu) odpowiedzi na takie pytania jak:
   a) Jakie informacje mają być wykryte przez model eksploracji danych?
   b) Czy otrzymane dane będą używane do przewidywania przyszłości czy wyłącznie do analizowania ukrytych zależności pomiędzy historycznymi danymi?
   c) Jakie informacje i z jakim wyprzedzeniem model ma przewidywać?
   d) Jakie zależności między danymi model ma wykrywać?
2. Dopiero znając zasady biznesowe, możemy ocenić przydatność danych źródłowych. W tym celu należy posłużyć się reprezentatywną próbką danych, poznać ich charakterystykę i wykryć istniejące między nimi zależności[4]. Na tym etapie należy również ocenić jakość danych, czyli odpowiedzieć na pytanie: Na ile dokładnie dane źródłowe opisują proces biznesowy?
3. Trzeci etap, czyli przygotowanie danych źródłowych, polega na ich dostosowaniu do wymogów modelu eksploracji danych. W środowisku serwera SQL 2008 używa się do tego celu pakietów SSIS i widoków danych źródłowych. **Właściwe przygotowanie danych jest najbardziej pracochłonnym etapem procesu ich eksploracji i jednocześnie etapem, od którego w dużym stopniu zależy przydatność otrzymywanych wyników**. Proces ten obejmuje:
   a) Przekształcanie danych w celu ich oczyszczenia (na przykład wyeliminowania błędnie wpisanych nazw tego samego towaru) oraz sformatowania (na przykład zamienienia dat na liczby).
   b) Odizolowanie i oznaczenie nietypowych danych — jeżeli rozkład danych źródłowych będzie niereprezentatywny (na przykład będą one zawierały informacje o zbyt wielu nastoletnich klientach), otrzymane przez ich eksplorację wyniki też będą błędne.
   c) Wyeliminowanie lub uzupełnienie brakujących informacji — używane do eksploracji dane nie powinny zawierać wartości Null, a więc, przygotowując dane, należy albo usunąć niekompletne rekordy, albo zastąpić wartości Null określonymi wartościami.
   d) Dyskretyzację danych (podzielenie ciągłych wartości pomiędzy zakresy) — większość algorytmów eksploracji danych zwraca bardziej wartościowe wyniki dla danych dyskretnych niż ciągłych.
   e) Normalizację danych rozumianą jako zastąpienie różnych, reprezentujących tę samą sytuację biznesową wartości (na przykład fakt zakupu przez klienta kilku towarów z danej kategorii) jedną wartością (w tym przypadku fakt zakupu dowolnego roweru może reprezentować cyfra 1).

[4] Do oceny danych źródłowych z reguły używa się technik eksploracji danych.

f) Spłaszczenie danych — eksplorowane dane powinny znajdować się w jednej tabeli (lub widoku)[5], tymczasem w bazach relacyjnych przechowywane są one w wielu powiązanych ze sobą tabelach.

> W podobny sposób przygotowuje się dane umieszczane w bazach analitycznych, dlatego często eksploruje się dane pochodzące właśnie z kostek analitycznych, a nie bezpośrednio odczytane z tabel hurtowni danych.

4. Czwarty etap, modelowanie, polega na przetestowaniu wybranych algorytmów eksploracji danych. Jeżeli na tym etapie okaże się, że konieczne jest użycie nieplanowanych wcześniej algorytmów eksploracji danych, prawdopodobnie konieczne będzie również ponowne przygotowanie dla nich danych źródłowych.
5. Ocena modelu eksploracji danych polega na zweryfikowaniu otrzymanych za jego pomocą wyników. W pierwszej kolejności należy ocenić techniczną poprawność modelu, czyli prawdopodobieństwo, że otrzymane wyniki są dokładne i wiarygodne — w serwerze SQL 2008 taką ocenę przeprowadza się, dzieląc dane wejściowe na użyte do przetrenowania modelu dane treningowe i użyte do jego sprawdzenia dane testowe, a następnie porównując otrzymane wyniki. Następnie należy ocenić merytoryczną poprawność danych, czyli skonsultować otrzymane wyniki z użytkownikiem końcowym. **Pominięcie tego kroku grozi otrzymaniem trywialnych, niemających żadnej wartości biznesowej wyników**. Na przykład, jeżeli okaże się, że najczęściej kupowane razem z innymi towarami jest pieczywo (co jest wysoce prawdopodobne, a więc model eksploracji danych zwrócił prawidłowe wyniki), to powodem takiego zachowania klientów jest fakt, że pieczywo szybko czerstwieje, a więc jest ono kupowane przy okazji każdej wizyty w sklepie.
6. Ostatnim krokiem jest wdrożenie modelu eksploracji danych i udostępnienie jego wyników użytkownikom. Obejmuje on:
   a) Utworzenie finalnej definicji modelu eksploracji danych.
   b) Przetworzenie (trening) modelu, czyli wykorzystanie danych historycznych do predykcji i wykrycie istniejących pomiędzy tymi danymi zależności — ta operacja może być bardzo czasochłonna.
   c) Użycie modelu do oceny nowych danych na podstawie wykrytych podczas jego przetwarzania reguł — takie operacje są bardzo szybkie i z reguły mogą być wykonywane w czasie rzeczywistym.
   d) Aktualizację modelu — w zależności od potrzeb model eksploracji danych może być przetwarzany codziennie, co tydzień lub co miesiąc. Po każdym przetworzeniu modelu należy ponownie ocenić otrzymywane za jego pomocą wyniki.

[5] Serwer SQL 2008 umożliwia eksplorację danych zapisanych w zagnieżdżonych tabelach, czyli tabelach połączonych relacją typu „jeden do wielu".

Zastosowania eksploracji danych

Eksploracja danych ma bardzo szeroki zakres zastosowań, ale najczęściej używana jest do klasyfikacji, regresji, segmentacji, prognozowania, asocjacji i analizy sekwencyjnej. Techniki eksploracji danych znajdują też coraz częściej zastosowanie w analizie tekstów, na przykład są używane do rozpoznawania niechcianych wiadomości pocztowych (spamu). Eksplorację danych zaczęto również stosować w aplikacjach, np. do inteligentnego sprawdzania poprawności danych.

Klasyfikacja

Klasyfikacja jest formą analizy predykcyjnej polegającej na przewidywaniu jednej lub więcej podanych wartości. Wynikiem klasyfikacji może być prosta odpowiedź *Tak* lub *Nie*, bądź *Prawda* lub *Fałsz*, ale może to być również jedna z wielu wartości, np. nazwa towaru czy nazwisko klienta.

Klasyfikacja pozwala rozwiązywać takie problemy jak:

1. Problem kredytodawcy, który chce wiedzieć, czy udzielić danemu klientowi kredytu? Jeżeli tak, to na jakich warunkach? Jakie jest ryzyko niespłacenia kredytu w przypadku tego klienta?
2. Problem handlowca, który zastanawia się, czy utraci danego klienta? Jeżeli tak, co spowodowało, że wybrał on konkurencję?
3. Problem marketingowca zastanawiającego się, kim są klienci firmy? Czy istnieją jakieś zależności między danymi demograficznymi klientów a ich chęcią kupowania w tej a nie innej firmie? Na których klientach należy się bardziej skoncentrować, ponieważ mogą oni przynieść firmie największe zyski?
4. Problem brygadzisty, który chce dowiedzieć się, dlaczego pewne serie produkowanych towarów mają więcej usterek niż inne? Jakie zmiany w produkcji mogą spowodować, że towary będą lepszej jakości?

W serwerze SQL 2008 klasyfikację powinno się przeprowadzać za pomocą algorytmu drzew decyzyjnych, regresji logistycznej, naiwnego klasyfikatora Bayesa lub sieci neuronowych[6].

W przykładowej bazie danych *Adventure Works DW 2008 SE* znajduje się kilka gotowych modeli eksploracji danych, w tym model klasyfikacji potencjalnych klientów, czyli osób, do których warto wysłać ulotkę reklamową.

W pierwszej kolejności przyjrzymy się danym źródłowym modelu klasyfikującego najlepszych odbiorców kampanii reklamowej:

[6] Opis poszczególnych algorytmów eksploracji danych znajduje się w następnym rozdziale.

1. Uruchom konsolę SSMS i połącz się z serwerem SQL 2008.
2. W oknie eksploratora obiektów rozwiń listę widoków bazy danych *AdventureWorksDW2008*.
3. Kliknij prawym przyciskiem myszy widok *vTargetMail* i wybierz opcję *Select Top 1000 Rows* (rysunek 8.2).

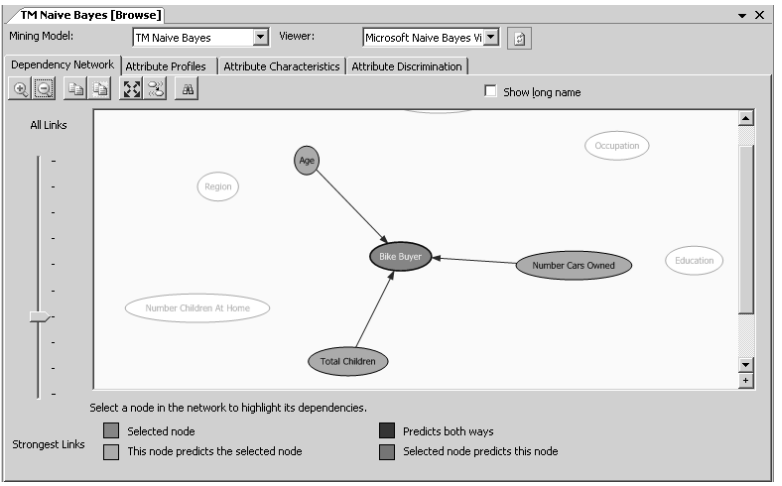
Rysunek 8.2. *Dane źródłowe, oprócz takich informacji jak liczba samochodów czy odległość pomiędzy miejscem zamieszkania a firmą, w której pracuje dana osoba, zawierają również informacje o tym, czy dana osoba kupiła u nas rower*

	g	NumberCarsOwned	AddressLine1	Phone	DateFirstPurchase	CommuteDistance	Region	Age	BikeBuyer
1		0	3761 N. 14th St	1 (11) 500 555-0162	2001-07-22	1-2 Miles	Pacific	42	1
2		1	2243 W St.	1 (11) 500 555-0110	2001-07-18	0-1 Miles	Pacific	43	1
3		1	5844 Linden Land	1 (11) 500 555-0184	2001-07-10	2-5 Miles	Pacific	43	1
4		1	1825 Village Pl.	1 (11) 500 555-0162	2001-07-01	5-10 Miles	Pacific	40	1
5		4	7553 Harness Circle	1 (11) 500 555-0131	2001-07-26	1-2 Miles	Pacific	40	1
6		1	7305 Humphrey Drive	1 (11) 500 555-0151	2001-07-02	5-10 Miles	Pacific	43	1
7		1	2612 Berry Dr	1 (11) 500 555-0184	2001-07-27	5-10 Miles	Pacific	42	1
8		2	942 Brook Street	1 (11) 500 555-0126	2001-07-12	0-1 Miles	Pacific	44	1
9		3	624 Peabody Road	1 (11) 500 555-0164	2001-07-28	10+ Miles	Pacific	44	1
10		1	3839 Northgate Road	1 (11) 500 555-0110	2001-07-30	5-10 Miles	Pacific	44	1
11		1	7800 Corrinne Court	1 (11) 500 555-0169	2001-07-17	5-10 Miles	Pacific	44	1
12		4	1224 Shoenic	1 (11) 500 555-0117	2001-07-02	10+ Miles	Pacific	45	1
13		2	4785 Scott Street	717-555-0164	2003-09-17	1-2 Miles	Nort...	40	0
14		3	7902 Hudson Ave.	817-555-0185	2003-10-15	0-1 Miles	Nort...	40	0
15		3	9011 Tank Drive	431-555-0156	2003-09-24	1-2 Miles	Nort...	40	0
16		1	244 Willow Pass Ro...	208-555-0142	2003-07-22	5-10 Miles	Nort...	29	1

Na podstawie powyższych danych trudno jest jednak ocenić, czy dana osoba byłaby zainteresowana kupnem kolejnego roweru. Musimy więc sklasyfikować posiadane dane pod kątem atrybutu *Bike Buyer*:

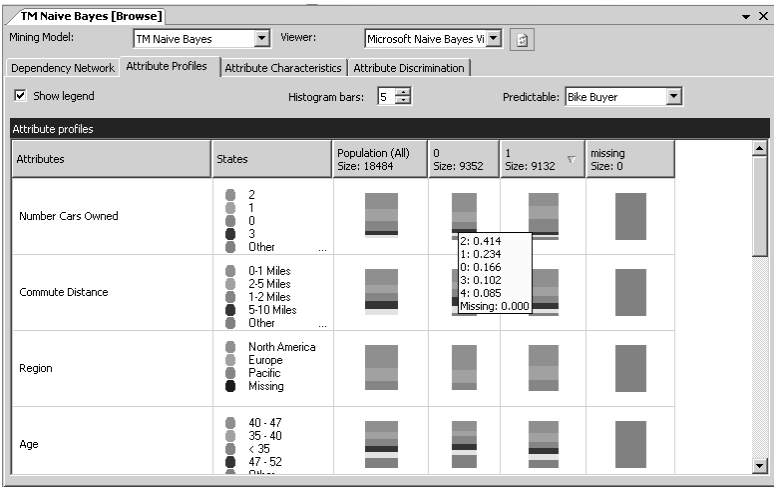
1. Za pomocą konsoli SSMS połącz się z serwerem SSAS.
2. W okienku eksploratora obiektów rozwiń sekcję *Databases/Adventure Works DW 2008 SE/Mining Structures/Targeted Mailing*.
3. Kliknij prawym przyciskiem myszy model *TM Naive Bayes* i z menu kontekstowego wybierz opcję *Browse*.
4. Każdy algorytm ma własny zestaw wizualizatorów. Wyniki uzyskane za pomocą naiwnego klasyfikatora Bayesa można oglądać w postaci wykresów zależności, rozkładu (profilu) wartości atrybutów, charakterystyki atrybutów oraz wykresu wpływu atrybutów.
5. Przyjrzyjmy się wykresowi zależności: w centralnej części zakładki *Dependency Network* umieszczony jest przewidywany atrybut — w tym przypadku jest nim *Bike Buyer* informujący o tym, czy dana osoba kupiła rower. Dookoła niego umieszczone są atrybuty, które miały wpływ na decyzję o zakupie roweru[7]. Kliknij atrybut *Bike Buyer*.
6. Po prawej stronie widoczny jest suwak — przesuń go na sam dół. W rezultacie na wykresie pozostanie zaznaczona tylko najsilniejsza zależność. W tym przypadku okazuje się, że największy wpływ na decyzję o zakupie roweru ma liczba posiadanych samochodów.
7. Przesuń suwak w górę — drugim pod względem wpływu na decyzję o zakupie roweru atrybutem okazał się wiek danej osoby, a trzecim — liczba dzieci (rysunek 8.3).

[7] Kierunek zależności symbolizuje strzałka oraz kolor tła atrybutu.



Rysunek 8.3. *Wykres wykrytych, istniejących w danych źródłowych zależności pomiędzy poszczególnymi atrybutami*

8. Przejdź do zakładki *Attribute profiles*. Zawiera ona histogramy rozkładu najważniejszych atrybutów w grupach osób, które kupiły i które nie kupiły roweru. W wierszach znajdują się histogramy rozkładu wartości kolejnych atrybutów, a w kolumnach informacje o rozkładzie wartości atrybutu w całej próbce danych w grupie osób, które nie kupiły roweru i w grupie osób, które kupiły rower (rysunek 8.4).

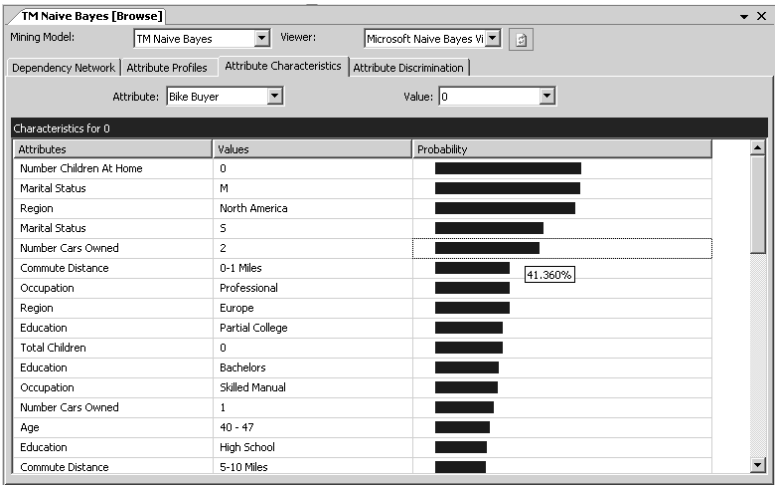


Rysunek 8.4. *Prawie ⅓ naszych klientów ma 2 samochody, a z wszystkich osób, które nie kupiły roweru, aż 41% to właśnie posiadacze 2 samochodów*

9. Przejdź do zakładki *Attribute Characteristics*. Prezentuje ona prawdopodobieństwo, z jakim wartość kolejnych atrybutów wpływa na decyzję o zakupie (*Characteristics for 1*) lub nie (*Characteristics for 0*) roweru. Po ustawieniu kursora nad wykresem prawdopodobieństwa, że osoby bezdzietne kupią rower, okazuje się, że wynosi ono prawie 63%.

10. Wybierz w polu *Value* wartość *0* — teraz wykres będzie pokazywał prawdopodobny wpływ poszczególnych atrybutów na brak decyzji o zakupie roweru. Sprawdź, ile procent posiadaczy 2 samochodów nie kupiło u nas roweru (rysunek 8.5).

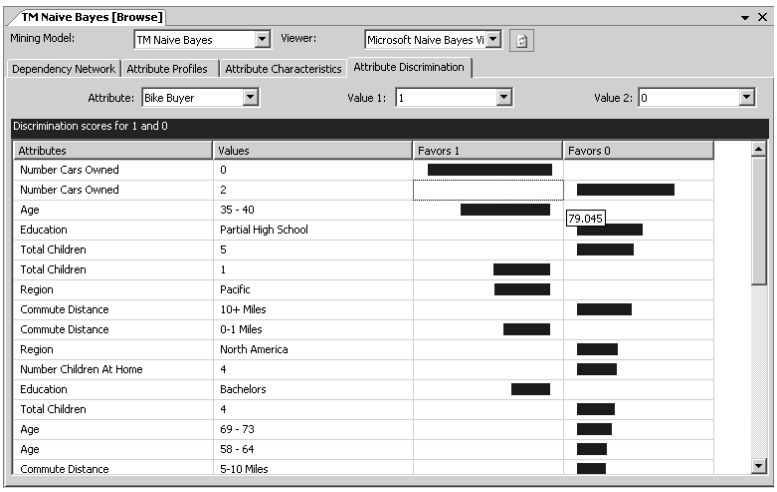
Rysunek 8.5. *Jako pierwsze pokazywane są atrybuty mające największy wpływ na brak decyzji o zakupie roweru*



11. Przejdź do zakładki *Attribute Discrimination*. Pokazuje ona wpływ wartości poszczególnych atrybutów na analizowany atrybut *Bike Buyer*. Żeby zobaczyć te dane, w polu *Value 1* wybierz *1*, a w polu *Value 2* wybierz *0*.

12. Przekonaj się, że brak samochodu oznacza, iż dana osoba prawie na pewno kupi u nas rower, natomiast posiadacze 2 samochodów prawie na pewno nie kupią roweru (rysunek 8.6).

Rysunek 8.6. *Umieszczając kursor nad wykresem, poznamy dokładną wielkość wpływu określonego stanu atrybutu (w tym przypadku posiadanie 2 samochodów) na decyzję o zakupie lub nie roweru*



Regresja

Regresja jest podobną do klasyfikacji formą analizy predykcyjnej, ale w jej przypadku przewidywane wartości nie muszą należeć do określonego zbioru. Wynikiem regresji może więc być np. wartość przyszłej sprzedaży lub przewidywany czas trwałości pewnego produktu.

Regresja pozwala rozwiązywać takie problemy jak:

1. Problem przedstawiciela handlowego, który chce wiedzieć, ile zysku przyniesie mu współpraca z danym klientem?
2. Problem klienta serwisu, którego interesuje, jak długo jego urządzenie będzie w naprawie?
3. Problem szefa działu PR, którego interesują czynniki wpływające na opinie klientów o firmie.

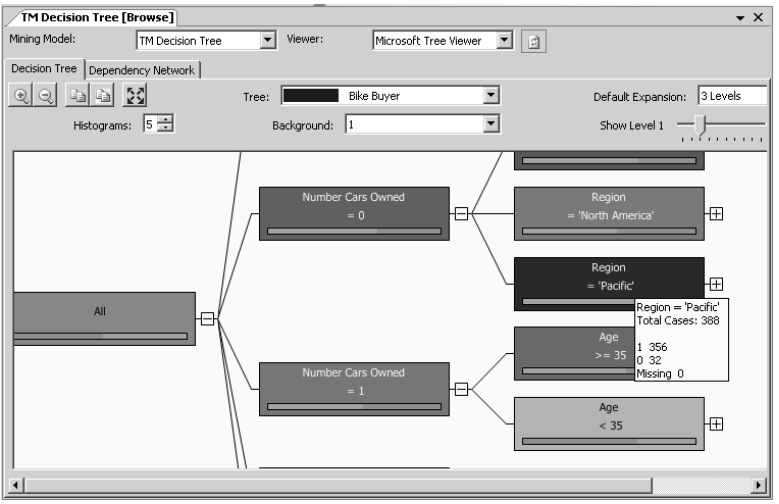
W serwerze SQL 2008 regresję powinno się przeprowadzać za pomocą algorytmu drzew decyzyjnych, regresji logistycznej, regresji liniowej lub sieci neuronowych.

Regresję przedstawiliśmy na przykładzie modelu utworzonego w tym samym celu, co poprzednio, a więc raz jeszcze będziemy zastanawiać się, do których użytkowników warto wysłać materiały reklamowe. Jednak tym razem przeanalizujemy go za pomocą algorytmu drzew decyzyjnych:

1. W konsoli SSMS połączonej z serwerem SSAS kliknij prawym przyciskiem myszy model eksploracji danych *TM Decision Tree* i wybierz opcję *Browse*.
2. Wyniki uzyskane za pomocą algorytmu drzew decyzyjnych można oglądać w postaci wykresów drzewa decyzji oraz zależności. W zakładce *Decision Tree* widoczny będzie pierwszy z tych wykresów.
3. Zwiń wszystkie węzły drzewa decyzji, tak aby widoczny był tylko jego pierwszy węzeł *All*. Domyślnie kolor tła węzła sygnalizuje liczbę opisywanych przez niego przypadków — im jest on ciemniejszy, tym więcej reprezentuje przypadków (rekordów tabeli źródłowej). Jeżeli przewidywane dane mają charakter dyskretny, w dolnej części węzła widoczny jest histogram rozkładu wartości pasujących do danego węzła przypadków.
4. Ustaw kursor myszki nad węzłem *All*. Po chwili wyświetlona zostanie informacja o rozkładzie przypadków. Zwróć uwagę, że prawie dokładnie połowa osób kupiła rower (9132 przypadki z 18 484) — **tak równomierny rozkład danych świadczy o ich prawidłowym przygotowaniu do eksploracji**.
5. Rozwiń znajdujące się na pasku narzędzi pole *Background* i wybierz *1* — od teraz kolor tła węzłów będzie sygnalizował liczbę osób, które kupiły rower.
6. Kliknij widoczny po prawej stronie węzła znak plus. Spowoduje to rozwinięcie pierwszego poziomu drzewa decyzji. Okazuje się, że największy wpływ na decyzję o zakupie roweru ma liczba posiadanych samochodów.

7. Najbardziej interesującą nas grupą klientów są osoby nieposiadające samochodu (ten węzeł jest najciemniejszy, a histogram wskazuje, że w ponad 60% przypadków takie osoby kupiły rower). Najmniej obiecującą grupą klientów są osoby posiadające 4 samochody (tylko w 466 na 1261 przypadków kupiły one u nas rower). Rozwiń węzeł *Number Cars Owned = 0*.

8. W grupie osób nieposiadających samochodu największy wpływ na decyzję o zakupie roweru ma rejon, w którym dana osoba mieszka. Sprawdź jeszcze, jaki czynnik ma największy wpływ na decyzję o zakupie roweru w grupie osób posiadających jedno auto — okazuje się, że w tej grupie był to wiek (rysunek 8.7).



Rysunek 8.7. *Algorytm drzew decyzyjnych jest jednym z najczęściej używanych algorytmów eksploracji danych, między innymi dlatego, że otrzymane za jego pomocą wyniki są wyjątkowo intuicyjne i łatwe do zinterpretowania*

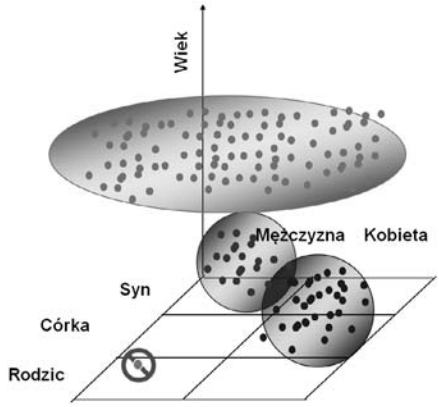
9. W poszukiwaniu charakterystyki idealnego odbiorcy naszej kampanii promocyjnej rozwiń węzeł *Region = 'Pacyfic'*, a następnie węzeł *Total Children not = 4*. Na widocznym z prawej strony pasku legendy sprawdź, że nieposiadający auta mieszkaniec rejonu Pacyfiku, o ile tylko nie ma on czwórki dzieci, kupi rower z prawdopodobieństwem przekraczającym 94%.

10. Przejdź do zakładki *Dependency Network*. Wyświetlony zostanie znany nam już wykres zależności pomiędzy atrybutami. Również ten model potwierdził, że największy wpływ na decyzję o zakupie roweru ma liczba posiadanych samochodów, ale drugi w kolejności okazał się roczny dochód, a większy wpływ niż wiek klienta miał jego rejon zamieszkania. Oba modele (klasyfikacji i regresji) zostały użyte do eksploracji tych samych danych[8], ale zwróciły nieco inne wyniki. Jest to zupełnie normalna sytuacja — **wyniki eksploracji danych są mniej lub bardziej prawdopodobne, ale nigdy nie są w 100% pewne**. Dlatego tak ważną rolę pełni ich testowanie oraz weryfikowanie.

[8] Ponieważ naiwny klasyfikator Bayesa nie pozwala eksplorować danych ciągłych, w rzeczywistości atrybut *Yearly Income* został w nim zignorowany.

Segmentacja

Segmentacja polega na grupowaniu rekordów i w przeciwieństwie do obu wcześniej przedstawionych technik nie jest formą analizy predykcyjnej. Zamiast tego pozwala ona łączyć w klastry przypadki mające podobną charakterystykę. W najprostszym przypadku dane o osobach zawierające informacje o ich wieku, płci i formie pokrewieństwa mogą być pogrupowane w pokazany na rysunku 8.8 sposób.



Rysunek 8.8. *Ubocznym, ale bardzo przydatnym efektem segmentacji jest wykrywanie anomalii, czyli przypadków niepasujących do żadnego klastra. W tym przypadku jest to bardzo młody mężczyzna będący rodzicem*

Segmentacja pozwala rozwiązywać takie problemy jak:

1. Problem pracownika działu kontrolingu, którego zadaniem jest wykrywanie podejrzanych transakcji finansowych.
2. Problem osoby odpowiedzialnej za bezpieczeństwo systemu komputerowego, która w dziennikach zdarzeń musi znaleźć podejrzane (nietypowe), mogące świadczyć o ataku, wpisy.
3. Problem marketingowca, który musi zaklasyfikować klientów do najlepiej opisujących ich kategorii.
4. Problem asystentki, która chce optymalnie uporządkować dokumenty.

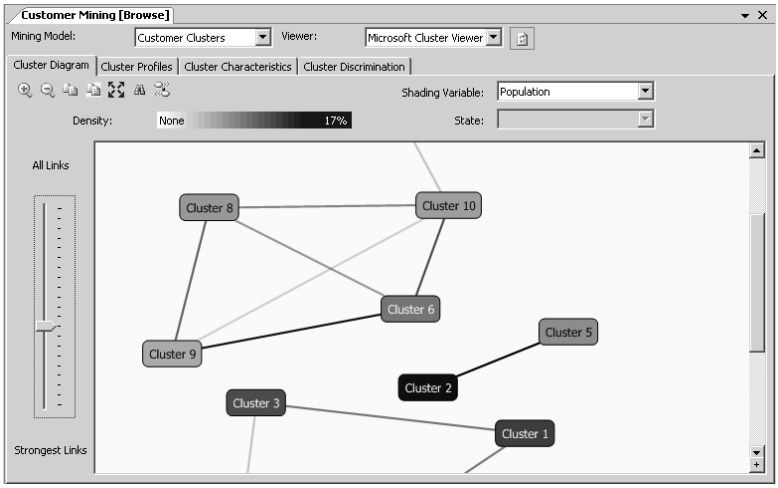
W serwerze SQL 2008 segmentację powinno się przeprowadzać za pomocą algorytmu klastrowania lub klastrowania sekwencyjnego.

Segmentację przedstawiliśmy na przykładzie algorytmu klastrowania użytego do eksploracji danych demograficznych o klientach:

1. W oknie eksploratora obiektów konsoli SSMS kliknij prawym przyciskiem myszy strukturę eksploracji danych *Customer Mining* i wybierz opcję *Browse*.
2. Wyniki uzyskane za pomocą algorytmu klastrowania można oglądać w postaci wykresów diagramu klastrów, profilów klastrów, charakterystyki klastrów i wpływu atrybutów na przynależność przypadku do klastra. Zakładka *Cluster Diagram* będzie zawierała pierwszy z tych wykresów.

3. Domyślnie odcień tła klastra reprezentuje jego liczebność — im ciemniejszy klaster, tym więcej zawiera on przypadków. Związki pomiędzy poszczególnymi klastrami można wyświetlać lub ukrywać, przesuwając znajdujący się z lewej strony wykresu suwak.

4. Ustaw ten suwak na tyle nisko, żeby na wykresie widoczne były tylko najsilniejsze związki pomiędzy klastrami (rysunek 8.9).



Rysunek 8.9. *Serwer SQL numeruje otrzymane w wyniku segmentacji danych klastry, dlatego po poznaniu charakterystyki każdego z nich należy zmienić ich nazwy na bardziej opisowe. Możemy to zrobić, klikając klaster prawym przyciskiem myszy i wybierając opcję Rename Cluster*

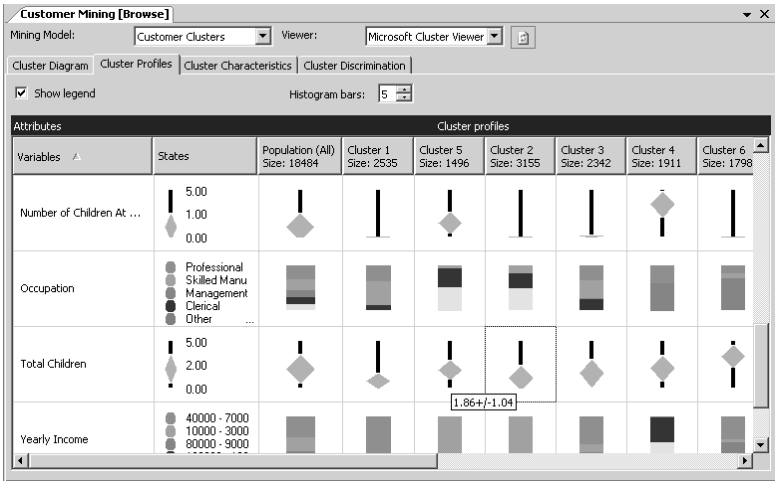
5. Odcień tła klastrów może reprezentować rozkład wartości dowolnego atrybutu. Żeby przekonać się, które klastry zawierają osoby o najniższych zarobkach, w polu *Shading Variable* wybierz atrybut *Yearly Income*, a w polu *State* wybierz najniższy przedział *10000 – 30000*. Okazuje się, że w klastrach 2. i 5. 100% osób zarabia więcej niż 10 tysięcy, ale mniej niż 30 tysięcy dolarów.

6. Przejdź do zakładki *Cluster Profiles*. Zawiera ona histogramy rozkładu dyskretnych wartości atrybutów w klastrach oraz wykresy średnich wartości ciągłych w klastrach. Porównaj rozkład rocznych zarobków, wykonywanych zawodów oraz liczbę dzieci osób z klastrów 2. i 5. (rysunek 8.10).

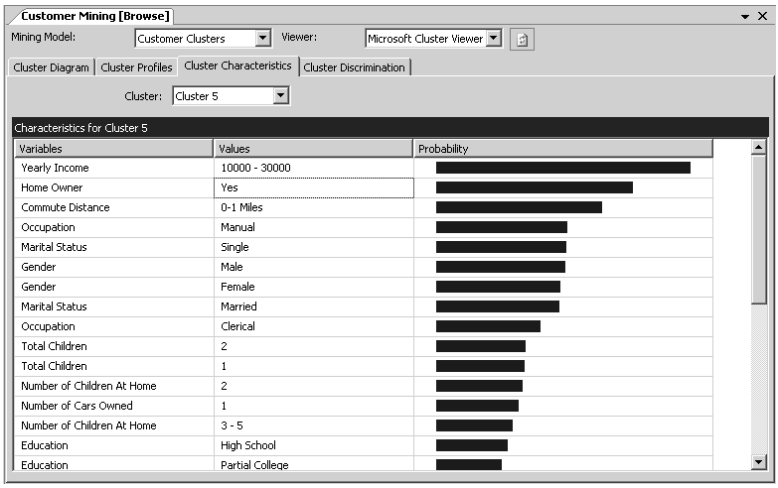
7. Przejdź do zakładki *Cluster Characteristics*. Zawiera ona informacje na temat prawdopodobieństwa, z jakim określony stan poszczególnych atrybutów wpływa na przynależność osoby do danego klastra. Wybierz klaster 5. — oczywiście najsilniejszy wpływ na przynależność do tego klastra ma niski dochód, jednak istotne są również: posiadanie domu (ponad 77% prawdopodobieństwa), zamieszkanie blisko miejsca pracy i bycie niewykwalifikowanym robotnikiem lub urzędnikiem (rysunek 8.11).

8. Żeby potwierdzić charakterystykę klastrów 2. i 5., przejdź do zakładki *Cluster Discrimination*. W pierwszej kolejności w polu *Cluster 1* wybierz *Cluster 5*. Wartość w polu *Cluster 2* automatycznie zmieni się na *Complement of Cluster 5* (dopełnienie klastra 5.). Tak jak przypuszczaliśmy, najsilniejszy wpływ na przynależność do tego klastra ma niski dochód, natomiast dochód powyżej 40 000$ jest silnym kontrargumentem. Stanowisko niewykwalifikowanego robotnika lub urzędnika ma duży wpływ na zaklasyfikowanie do klastra 5., natomiast praca na stanowisku specjalisty jest równie silnym argumentem przeciw.

Rysunek 8.10. *Roczny dochód w klastrach 2. i 5. jest taki sam, podobny jest też rozkład zawodów wykonywanych przez zaklasyfikowane do tych klastrów osoby (w obu przypadkach są to głównie niewykwalifikowani robotnicy lub urzędnicy), ale osoby z klastra 5. mają więcej dzieci, które w większości przypadków wciąż z nimi mieszkają, tymczasem osoby z klastra 2. mieszkają bez dzieci*



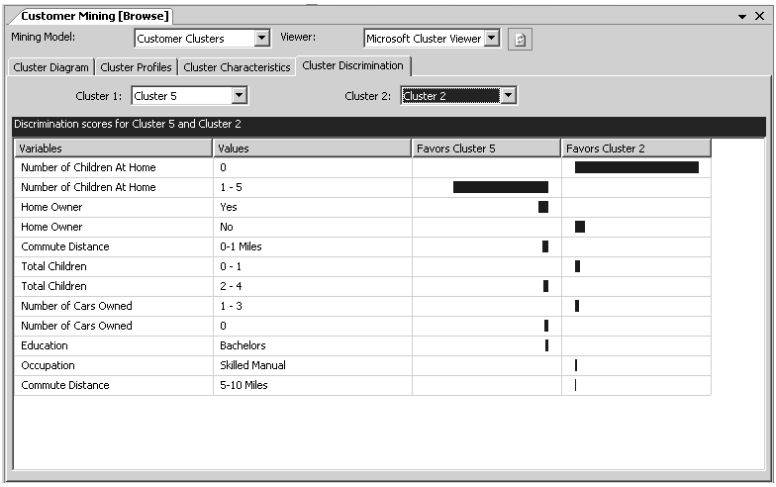
Rysunek 8.11. *Analizując wpływ wartości atrybutów na wynik segmentacji, należy zwracać uwagę na wszystkie możliwe wartości danego atrybutu. W tym przypadku bycie mężczyzną ma prawie taki sam wpływ na przynależność do klastra 5., co bycie kobietą, a więc ten atrybut nie ma w rzeczywistości żadnego znaczenia*



9. Pozostało nam jeszcze porównanie ze sobą tylko czynników wpływających na przynależność do klastra 5., a nie do silnie z nim powiązanego klastra 2. W tym celu w polu *Cluster 2* wybierz *Cluster 2* (rysunek 8.12).

10. Wróć do zakładki *Cluster Diagram* i zmień nazwę klastra 5. na `UbogieRodzinyRobotnicze`.

Uzyskane w wyniku segmentacji klastry mogą być używane jako wymiary kostek analitycznych w zapytaniach zwracających nazwę klastra, do którego należy dany rekord, lub w „inteligentnych" (tj. samouczących się) aplikacjach.



Rysunek 8.12. *Jedyną istotną różnicą pomiędzy wybranymi klastrami jest liczba mieszkających z rodzicami dzieci*

Asocjacja

Asocjacja jest techniką wykrywania istniejących pomiędzy poszczególnymi przypadkami zależności i najczęściej jest stosowana do analizy koszyka zakupów, czyli wyszukiwania najczęściej razem kupowanych towarów.

Asocjacja pozwala rozwiązywać takie problemy jak:

1. Problem sprzedawcy chcącego wiedzieć, które towary są kupowane w ramach jednej transakcji, w celu zarekomendowania ich klientom i zwiększenia w ten sposób szansy na sprzedaż krzyżową (ang. *Cross-selling*). Wyniki tego typu analizy są powszechnie używane na witrynach sklepów internetowych, w których po wybraniu produktu wyświetla się lista kupowanych razem z tym produktem towarów[9].
2. Problem kierownika sklepu, który musi właściwie zaplanować rozmieszczenie towarów — jeżeli na sąsiednie półki trafią często razem kupowane towary, prawdopodobnie klient kupi oba towary, ale jeżeli obok siebie będą leżały produkty, których klienci razem nie kupują (np. w Stanach Zjednoczonych osoby, które kupują francuskie pieczywo, nie chcą widzieć w pobliżu amerykańskich pączków), mogą oni zrezygnować z zakupu i wyrobić sobie negatywną opinię o sklepie.

W serwerze SQL 2008 asocjację powinno się przeprowadzać za pomocą algorytmu reguł asocjacyjnych lub drzewa decyzyjnego.

Praktyczny przykład użycia asocjacji do zasugerowania klientom sklepu internetowego zakupu dodatkowych towarów dostępny jest na witrynie SQL Server Data Mining:

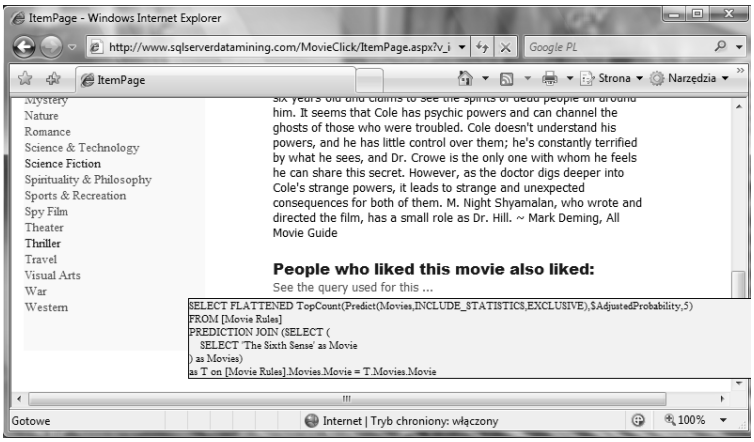
1. Połącz się z adresem *http://www.sqlserverdatamining.com*.

[9] Podsuwane w ten sposób klientom sugestie nie powinny być zbyt oczywiste — z reguły zaproponowanie klientom niepowiązanych ze sobą na pierwszy rzut oka towarów przynosi lepsze rezultaty.

2. Kliknij odnośnik *Live!Samples*. Jednym z dostępnych przykładów będzie *Movie!Click* — wybierz go i uruchom.

3. Wyświetlona zostanie strona fikcyjnego sklepu zajmującego się sprzedażą filmów. Wybierz dowolny film.

4. Oprócz informacji na temat wybranego filmu wyświetlona zostanie informacja o filmach, które są często kupowane razem z tym filmem (ang. *People who liked this movie also liked:*). Powyżej listy tytułów znajdować się będzie odnośnik *See the query used for this...*. Po ustawieniu nad nim kursora myszki wyświetlone zostanie zapytanie MDX, które zwróci tę listę tytułów (rysunek 8.13).



Rysunek 8.13. *Zapytanie zwracające listę filmów najczęściej kupowanych razem z filmem Szósty zmysł. Ponieważ w odczytywanym modelu zastosowano zagnieżdżone tabele, wynik został spłaszczony za pomocą słowa kluczowego FLATTENED*

Natomiast w przykładowej bazie danych AdventureWorksDW2008 znajdują się dwa widoki zwracające dane o transakcjach zakupu i dokonujących je klientach:

1. Za pomocą konsoli SSMS połącz się z serwerem SQL 2008.

2. W oknie edytora kodu wykonaj poniższe zapytanie:

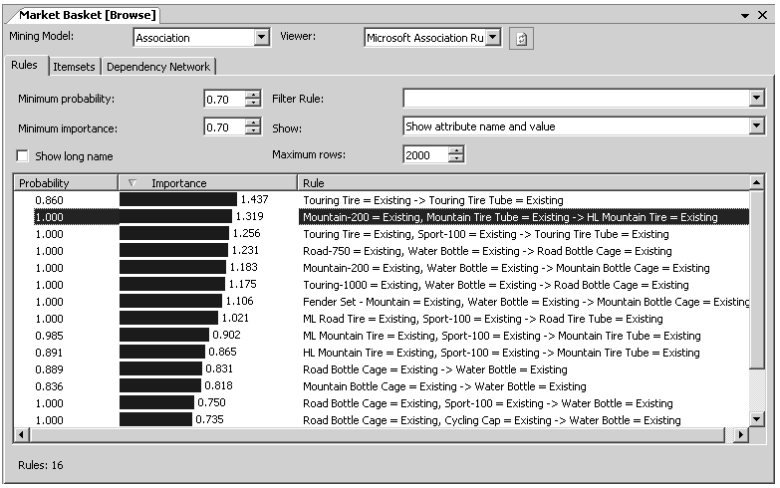
```
SELECT O.CustomerKey, O.IncomeGroup, O.Region, L.OrderNumber,L.Model
FROM AdventureWorksDW2008.dbo.vAssocSeqOrders AS O
JOIN AdventureWorksDW2008.dbo.vAssocSeqLineItems AS L
ON O.OrderNumber = L.OrderNumber
-----------------------------------------
15053    Low          Europe   SO66532          Classic Vest
15061    Low          Europe   SO63299          Mountain-500
15061    Low          Europe   SO63299          Sport-100
15086    Moderate     Europe   SO61491          Mountain-200
15086    Moderate     Europe   SO61491          HL Mountain Tire
15086    Moderate     Europe   SO64841          Cycling Cap…
```

Chociaż zwraca ono wszystkie potrzebne do analizy koszyka zakupów dane, to odczytanie z niego interesujących nas informacji wymagałoby skomplikowanego pogrupowania. Lepiej będzie w tym celu posłużyć się modelem eksploracji danych:

1. Za pomocą konsoli SSMS połącz się z serwerem SSAS.

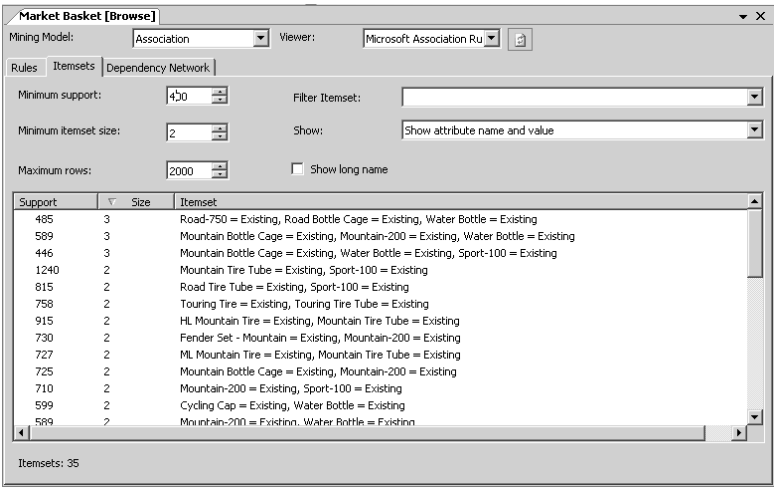
2. Kliknij prawym przyciskiem myszy strukturę eksploracji danych *Market Basket* i wybierz opcję *Browse*.

3. Wyniki uzyskane za pomocą algorytmu reguł asocjacyjnych można oglądać w postaci listy znalezionych reguł, listy istniejących koszyków zakupów oraz wykresu zależności. Zakładka *Rules* będzie zawierała listę reguł, którymi kierują się klienci.
4. Wpisz w polu *Minimum Probability* wartość `0.70` — w ten sposób wyświetlone zostaną tylko reguły obowiązujące z co najmniej 70% prawdopodobieństwem.
5. W polu *Minimum Importance* wybierz wartość `0.70` — ograniczy to liczbę wyświetlanych reguł do tych, które mają największy wpływ na budowanie koszyków zakupów (rysunek 8.14).



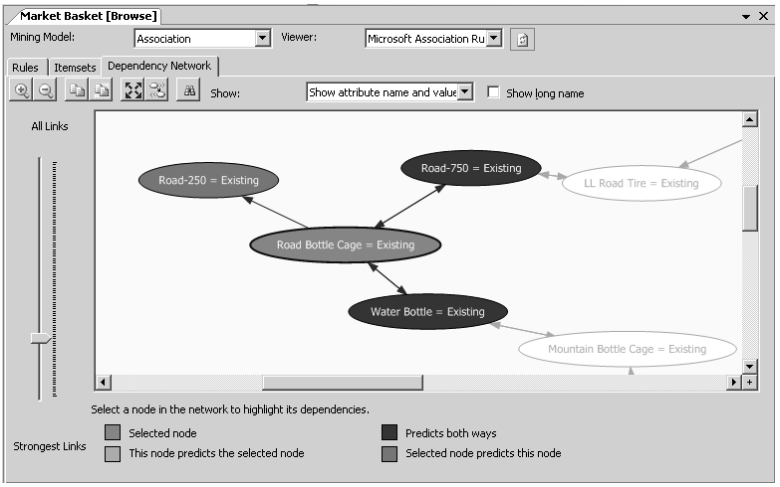
Rysunek 8.14. *Wśród reguł obowiązujących ze 100% prawdopodobieństwem najsilniejsza jest pewna reguła — według niej osoba, która kupiła rower Mountain-200 oraz dętkę Mountain Tire Tube, kupi też oponę HL Mountain Tire*

6. Przejdź do zakładki *Itemsets* i wyeliminuj mniej popularne (rzadziej występujące) koszyki zakupów, zwiększając wartość pola *Minimum Support* do *400*.
7. Wyświetl jedynie koszyki zakupów zawierające co najmniej dwa towary — w tym celu w polu *Minimum itemset size* wybierz wartość *2* (rysunek 8.15).



Rysunek 8.15. *Najczęściej występujący, trzyelementowy koszyk zakupów zawiera rower Road-750, uchwyt na bidon Road Bottle Cage oraz bidon Water Bottle*

8. Przejdź do zakładki *Dependency Network* i przesuń w dół znajdujący się po lewej stronie suwak, tak żeby zaznaczone zostały tylko najsilniejsze powiązania pomiędzy towarami.
9. Zaznacz towar *Road Bottle Cage*, kliknij przycisk *Improve Layout* i powiększ otrzymany wykres (rysunek 8.16).



Rysunek 8.16. *Osoby kupujące uchwyt na bidon często kupują też bidon i rower Road-750. Ponieważ pomiędzy tymi towarami istnieje dwukierunkowa zależność, osoby kupujące bidon lub rower Road-750 prawdopodobnie kupią też jego uchwyt. Ponadto osoby kupujące uchwyt na bidon wybierają też rower Road-250, ale ta zależność jest jednokierunkowa, a więc kupno roweru Road-250 nie ma silnego wpływu na decyzję o zakupie uchwytu*

Analiza sekwencyjna

Wszystkie przedstawione do tej pory techniki eksploracji danych nie uwzględniały kolejności, w jakiej występowały badane przez nie zdarzenia. Analiza sekwencyjna wykrywa właśnie często powtarzające się sekwencje zdarzeń.

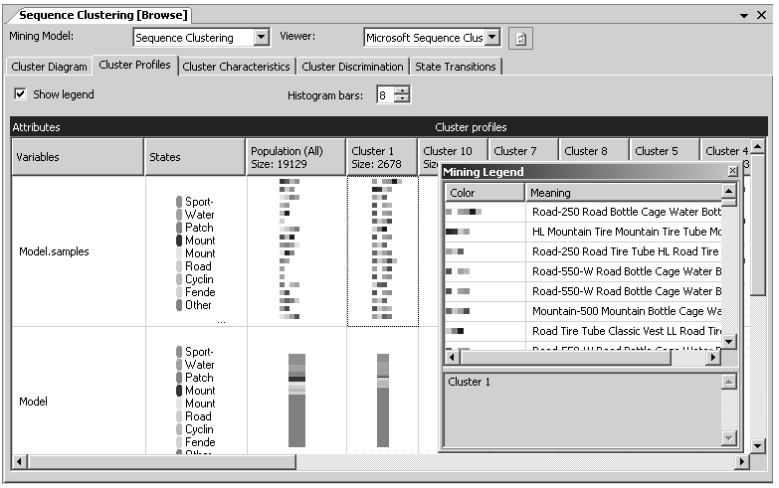
Analiza sekwencyjna pozwala rozwiązywać takie problemy jak:

1. Problem projektanta witryny WWW, który chce wiedzieć, w jakiej kolejności odwiedzające jego witrynę osoby wyświetlają poszczególne podstrony. Dysponując taką informacją, może on dostosować witrynę do potrzeb użytkowników i zmniejszyć liczbę osób rezygnujących z poszukiwania interesujących je podstron na niewłaściwie zaprojektowanej witrynie.
2. Problem naukowca chcącego przewidzieć dalszy rozwój przeprowadzanych badań (np. wyniki przyszłych eksperymentów) lub chcącego zweryfikować hipotezę statystyczną.

W serwerze SQL 2008 analizę sekwencyjną powinno się przeprowadzać za pomocą algorytmu klastrowania sekwencyjnego.

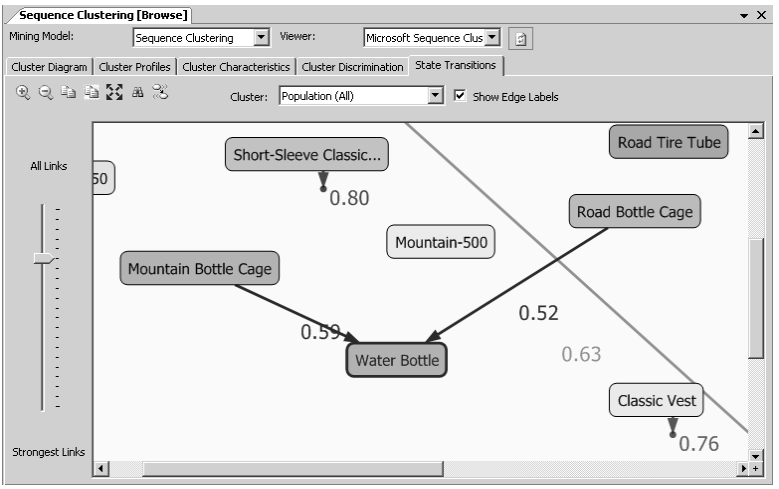
W przykładowej bazie danych *AdventureWorksDW2008* analiza sekwencyjna została zastosowana do segmentacji koszyków zakupów na podstawie kolejności dodawanych do nich towarów. Do przeprowadzenia tej analizy użyto danych pochodzących z tych samych co w poprzednim modelu dwóch widoków, tym razem uzupełnionych o numery poszczególnych pozycji zamówień:

1. Wyświetl w konsoli SSMS strukturę eksploracji danych *Sequence Clustering*.
2. Wyniki uzyskane za pomocą algorytmu klastrowania sekwencyjnego można oglądać w postaci wykresów diagramu klastrów, profilów klastrów, charakterystyki klastrów, wpływu atrybutów na przynależność przypadku do klastra oraz wykresu zmian stanu. Zakładka *Cluster Diagram* będzie zawierała pierwszy z tych wykresów. Ponieważ ten typ wykresu został już opisany przy okazji segmentacji, przejdź do zakładki *Cluster Profiles*.
3. Oprócz charakterystyki poszczególnych klastrów (w tym przypadku utworzonej na podstawie różnych modeli robienia zakupów) w wierszu *Model.samples* widoczne będą sekwencje, w jakich poszczególne towary były dodawane do zamówień (rysunek 8.17).



Rysunek 8.17. *Wykres charakterystyki znalezionych klastrów z wyświetlonymi w okienku legendy sekwencjami dodawania towarów w ramach 1. klastra*

4. Wykresy charakterystyki klastrów oraz wpływu wartości atrybutów na przynależność przypadku do danego klastra przypominają wykresy przedstawione podczas opisywania techniki segmentacji i nie zostały ponownie opisane. Nowym typem wykresu jest tylko wykres zmian stanu. Żeby go wyświetlić, przejdź do zakładki *State Transitions*.
5. W głównym okienku widoczne będą wszystkie możliwe stany, w tym przypadku reprezentowane przez nazwy poszczególnych towarów. Przesuń znajdujący się z lewej strony suwak w dół, tak aby zaznaczone zostały tylko najbardziej prawdopodobne zmiany stanów.
6. W polu *Cluster* można wybrać analizowany klaster, domyślnie pokazywane są zmiany stanów wszystkich przypadków. Zaznacz stan *Water Bottle* (rysunek 8.18).



Rysunek 8.18. *Kupno dowolnego uchwytu na bidon z ponad 50% prawdopodobieństwem prowadzi do zakupu bidonu, po zakupieniu którego klienci kończą wizytę w sklepie*

Prognozowanie

Prognozowanie, czyli szacowanie przyszłości, w dużym stopniu opiera się na przedstawionej regresji, ale w serwerze SQL 2008 wizualizatory klasycznych algorytmów regresji znacznie różnią się od wizualizatora algorytmu szeregów czasowych.

Prognozowanie pozwala rozwiązywać takie problemy jak:

1. Problem prezesa, który chciałby wiedzieć, jakie zyski osiągnie firma w przyszłym roku.
2. Problem maklera, który chciałby przewidzieć zmiany cen akcji.
3. Problem magazyniera, który musi tak zaplanować kolejne dostawy, żeby nie mieć problemu ze składowaniem towaru.
4. Problem agencji reklamowej, która chce przedstawić klientowi prognozowane dane na temat liczby odwiedzin stron z reklamą oraz planowane wpływy reklamy na wyniki sprzedaży.

W serwerze SQL 2008 prognozowanie powinno się przeprowadzać za pomocą algorytmu szeregów czasowych.

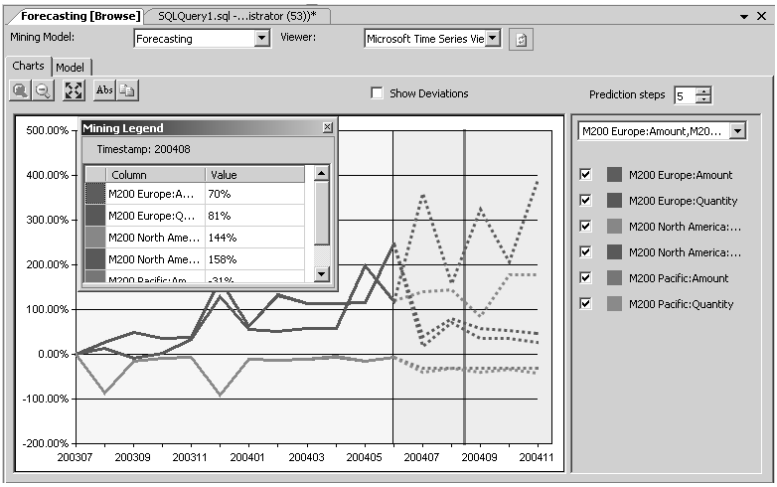
W przykładowej bazie danych AdventureWorksDW2008 znajduje się widok zawierający historyczne dane o sprzedaży w poszczególnych rejonach. Żeby je odczytać, wykonaj w oknie edytora SQL poniższe zapytanie:

```
SELECT ModelRegion, TimeIndex, Quantity,Amount
FROM AdventureWorksDW2008.dbo.vTimeSeries
----------------------------------------
M200 Europe      200310        50     115449.50
M200 Europe      200311        51     117644.49
```

```
M200 Europe      200312      99      228304.01
M200 Europe      200401      60      138349.40
M200 Europe      200402      86      198344.14
…
```

Natomiast w bazie analitycznej *Adventure Works DW 2008 SE* znajduje się model eksploracji danych *Forecasting*, który na podstawie tych samych danych przewiduje przyszłe wyniki sprzedaży w poszczególnych rejonach:

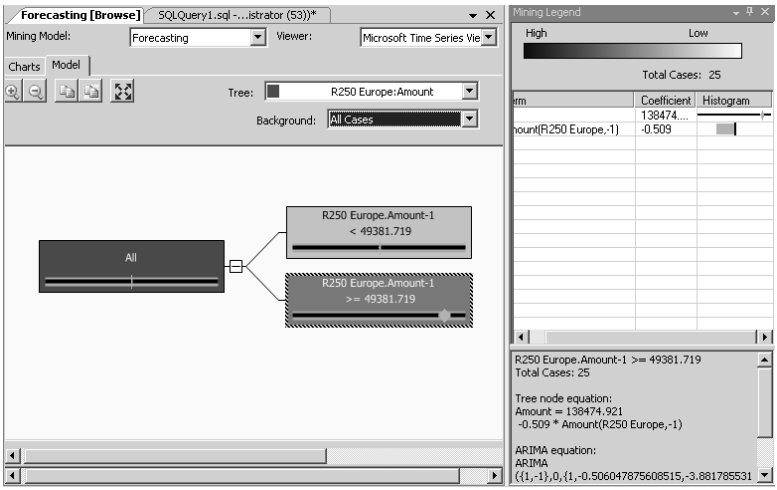
1. Połącz się za pomocą konsoli SSMS z serwerem SSAS i wyświetl model *Forecasting*.
2. Wyniki uzyskane za pomocą algorytmu szeregów czasowych można oglądać w postaci wykresów zmian wartości w czasie oraz wykresu modelu. Zakładka *Charts* będzie zawierała pierwszy z tych wykresów.
3. W tym przypadku jednostką czasu jest rok, a więc, wybierając w polu *Prediction steps* wartość *5*, oszacujemy wartość sprzedaży na pięć lat wprzód. Im więcej okresów przewidujemy, tym mniej precyzyjne otrzymamy wyniki — żeby się o tym przekonać, zaznacz pole wyboru *Show Deviations* i zwiększ na chwilę liczbę prognoz do 15.
4. Z lewej strony wykresu znajduje się też pole listy pozwalające wybrać widoczne na wykresie wartości oraz pola wyboru pozwalające ukryć niektóre z wybranych z tej listy wartości (rysunek 8.19).



Rysunek 8.19. *Prognozowane wartości zaznaczone są linią przerywaną, natomiast legenda pokazuje prognozy na okres wskazany pionową linią*

5. Działanie algorytmu szeregów czasowych polega na skonstruowaniu drzewa regresji dla każdego prognozowanego parametru. Drzewa te możemy zobaczyć w zakładce *Model* (rysunek 8.20)[10].

[10] Dodatkowe informacje na temat używanych przez serwer SQL 2008 do prognozowania algorytmów ARTXP i ARIMA znajdują się w rozdziale 9.



Rysunek 8.20. *Do prognozowania używane są scałkowane drzewa autoregresyjne z predykcją krzyżową i średnią ruchomą*

Serwer SQL 2008

Tylko edycje Enterprise, Standard oraz Developer serwera SQL 2008 zawierają zintegrowany z usługą SSAS moduł eksploracji danych, niestety w edycji Standard nie znajdziemy m.in. niektórych algorytmów.

Programowy dostęp do modeli eksploracji danych w serwerze SQL 2008 zapewniają:

1. Model AMO (ang. *Analysis Management Objects*) — jest to ten sam interfejs API, który umożliwia tworzenie i modyfikowanie obiektów baz analitycznych, zarządzanie tymi bazami oraz ich zabezpieczanie.
2. Model ADOMD.NET (*ActiveX Data Objects Multi-dimensional .NET*) — interfejs API pozwalający łączyć się z serwerem SSAS oraz odczytywać i modyfikować przechowywane w bazach analitycznych dane.
3. Model Server ADOMD.NET — interfejs rozszerzający wsparcie języka MDX o procedury składowane stworzone w językach platformy .NET, takich jak C# oraz VB.NET.

Ponadto serwer SQL 2008 umożliwia rejestrowanie dodatkowych algorytmów eksploracji danych oraz dodatkowych wizualizatorów zwracanych przez te algorytmy wyników.

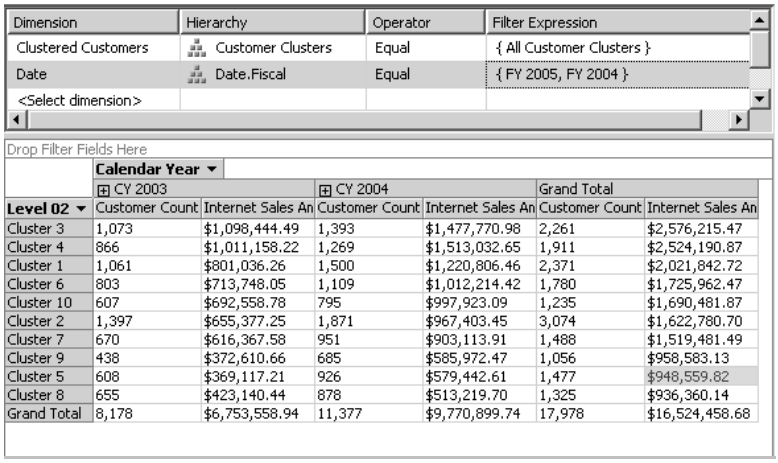
Integracja z usługami Business Intelligence

Serwer SQL 2008 jest kompletną platformą Business Intelligence, w której usługi eksploracji danych są ściśle powiązane z pozostałymi usługami biznesowymi.

Integracja z serwerem SSAS

Integracja z serwerem SSAS oznacza, że modele eksploracji danych mogą być tworzone na podstawie danych odczytanych z kostek analitycznych oraz że wyniki eksploracji danych (z reguły uzyskane za pomocą algorytmów drzew decyzyjnych, klastrowania lub reguł asocjacyjnych) mogą być używane w roli wymiarów. **W obu przypadkach modele eksploracji danych i kostki wielowymiarowe muszą znajdować się w tej samej bazie analitycznej**.

Użycie modelu eksploracji danych w roli wymiaru pozwala analizować fakty biznesowe w nowych kontekstach, na przykład używać wykrytych grup klientów do analizowania danych o sprzedaży. W przykładowej bazie danych znajduje się kostka *Mined Customers*, której jeden z wymiarów (*Clustered Customers*) jest stworzony na podstawie opisanego w punkcie Segmentacja modelu eksploracji danych (rysunek 8.21).



Dimension	Hierarchy	Operator	Filter Expression
Clustered Customers	Customer Clusters	Equal	{ All Customer Clusters }
Date	Date.Fiscal	Equal	{ FY 2005, FY 2004 }
<Select dimension>			

Drop Filter Fields Here

	Calendar Year					
	CY 2003		CY 2004		Grand Total	
Level 02	Customer Count	Internet Sales An	Customer Count	Internet Sales An	Customer Count	Internet Sales An
Cluster 3	1,073	$1,098,444.49	1,393	$1,477,770.98	2,261	$2,576,215.47
Cluster 4	866	$1,011,158.22	1,269	$1,513,032.65	1,911	$2,524,190.87
Cluster 1	1,061	$801,036.26	1,500	$1,220,806.46	2,371	$2,021,842.72
Cluster 6	803	$713,748.05	1,109	$1,012,214.42	1,780	$1,725,962.47
Cluster 10	607	$692,558.78	795	$997,923.09	1,235	$1,690,481.87
Cluster 2	1,397	$655,377.25	1,871	$967,403.45	3,074	$1,622,780.70
Cluster 7	670	$616,367.58	951	$903,113.91	1,488	$1,519,481.49
Cluster 9	438	$372,610.66	685	$585,972.47	1,056	$958,583.13
Cluster 5	608	$369,117.21	926	$579,442.61	1,477	$948,559.82
Cluster 8	655	$423,140.44	878	$513,219.70	1,325	$936,360.14
Grand Total	8,178	$6,753,558.94	11,377	$9,770,899.74	17,978	$16,524,458.68

Rysunek 8.21. *Klienci zaklasyfikowani do grupy ubogich rodzin robotniczych (klaster 5.), pomimo że są dość liczni, stanowią prawie że najmniej dochodową grupę naszych klientów*

Integracja z serwerem SSIS

Techniki eksploracji danych doskonale nadają się do przekształcania w ramach rozbudowanych procesów ETL. Serwer SSIS zawiera trzy komponenty ułatwiające użycie tych technik:

1. Zadanie *Data Mining Query Task* pozwala wykonywać zapytania predykcyjne (instrukcje języka DMX), na przykład sprawdzić, z jakim prawdopodobieństwem dana osoba zostanie naszym klientem.
2. Dostępna tylko w edycji Enterprise transformacja *Data Mining Query* pozwala wykonywać w trakcie przekształcania danych dowolną liczbę zapytań predykcyjnych.
3. Dostępne tylko w edycji Enterprise zadanie *Data Mining Model Training Destination* pozwala przetwarzać modele eksploracji danych.

Zastosowanie technik eksploracji danych w pakietach SSIS umożliwia przeprowadzenie „inteligentnego" oczyszczania danych, zaawansowaną klasyfikację przypadków oraz wypełnienie brakujących danych już na etapie ich importowania do hurtowni.

Integracja z serwerem SSRS

Serwer SSRS pozwala tworzyć raporty zawierające wyniki eksploracji danych (dane zwrócone przez zapytania MDX, których budowanie ułatwia graficzny kreator). Wyniki zapytań predykcyjnych mogą być też używane jako parametry raportów, możliwe jest więc wyświetlenie danych osób, które z ponad 85% prawdopodobieństwem zostaną naszymi klientami.

Praca z raportami tego typu nie różni się od pracy z raportami prezentującymi dane odczytane z baz relacyjnych lub wielowymiarowych kostek analitycznych, w szczególności możliwe jest dowolne grupowanie odczytanych z modeli eksploracji danych oraz dystrybuowanie raportów poprzez subskrypcje serwera SSRS.