STUDIA INFORMATICA 2010
Volume 31 Number 2B (90)

Paweł DRZYMAŁA, Łukasz SOBCZAK, Henryk WELFLE, Sławomir WIAK
Politechnika Łódzka, Instytut Mechatroniki i Systemów Informatycznych

SYSTEM GROMADZENIA I ANALIZY DANYCH KORPORACYJNYCH Z WYKORZYSTANIEM NARZĘDZI BUSINESS INTELLIGENCE I BAZ DANYCH

Streszczenie. W artykule zaprezentowano współczesne techniki przetwarzania danych pozyskiwanych przez korporacje i instytucje finansowe. Przedstawiono proces przepływu danych od momentu ich pozyskania do momentu otrzymania końcowych wyników obliczeń i uzyskania wiedzy. Opracowano i zaprezentowano aplikację, która pozwala użytkownikowi na szybkie i skuteczne raportowanie informacji z systemu oraz wspomaga proces podejmowania decyzji we wszystkich aspektach działalności organizacji przez jej pracowników począwszy od doradcy, skończywszy na kadrze zarządzającej. W procesie projektowania systemu uwzględniono różnorakie aspekty działalności korporacji finansowych.

Słowa kluczowe: Bazy Danych w Zarządzaniu, Business Intelligence Hurtownie Danych, OLAP

CORPORATE DATA ACQUISITION AND ANALYSING SYSTEM BASED ON BUSINESS INTELLIGENCE TOOLS AND DATABASE

Summary. The paper presents a modern data acquisition techniques and techniques of the data processing obtained by corporations and financial institutions. It also presents the data flow process from the beginning (moment of obtaining the data) to the final results of the calculations and knowledge. It presents an developed application that allows users to quickly and effectively reporting information from the system and assists in decision-making in all aspects of the organization by its employees as a staff adviser to the ending of management.

Keywords: Database in Management, Business Intelligence, Data Warehouse, OLAP.

1. Wstęp

Współczesne korporacje i instytucje finansowe potrzebują narzędzi zapewniających wspomaganie sprzedaży. Wysokie koszty przeprowadzenia kampanii marketingowej oraz czas jej przygotowania wymagają posiadania narzędzi ułatwiających to zadanie. Nadrzędnym celem takich działań jest maksymalizacja zysków przy jak najmniejszych kosztach. Istnieje wiele sposobów dotarcia do klienta. Rozwój technologii informacyjnych ułatwił to zadanie.

Klienci instytucji finansowych również otrzymali nowe możliwości. Należy tu wymienić np. Internet Banking, który umożliwia zarządzanie portfelem klienta w sposób zdalny. Pozwala to zaoszczędzić czas potrzebny na wykonanie zleconych dyspozycji oraz zwalnia klienta z wymogu fizycznej obecności w oddziale instytucji.

Instytucje finansowe chcąc generować zyski zmuszone są oferować klientom nowe produkty finansowe. Każdy produkt ma swoją specyfikę i nie może być adresowany do wszystkich klientów. Jakkolwiek większość klientów można nakłonić do oszczędzania, tak kredyt hipoteczny może być zaoferowany tylko i wyłącznie klientom spełniającym określone warunki. Taka sytuacja wymaga od korporacji narzędzi zapewniających wspomaganie procesu sprzedaży.

Przykładowo oddział zajmujący się produktami oszczędnościowymi przygotował nowy rachunek a vista gwarantujący klientowi wysokie oprocentowanie przy zachowaniu niektórych własności rachunku bieżącego. Produkt ten będzie zaoferowany wszystkim klientom. Rolą doradców finansowych jest przekonywanie do zakupu tego produktu. Jednakże pomimo tego, że produkt ten może nabyć każdy klient, nie każdy go zakupi. Doradca zobowiązany jest poinformować klienta o nowym produkcie i przekonywać go o jego zaletach. Istnieje wiele możliwości dotarcia z taką informacja. Są to głównie mass media, poczta elektroniczna, SMS, telefony. Jednakże zwiększa to znacząco środki finansowe wydane na kampanię marketingową i może doprowadzić do sytuacji, kiedy pozyskane depozyty i zyski z nich osiągnięte w całości zostaną przeznaczone na kampanię marketingową.

Innym podejściem jest wykorzystanie technik drążenia danych oraz wiedzy eksperckiej analityków. Można wyróżnić pewne cechy wspólne klientów, którzy już posiadają tego typu produkt lub produkty podobne. Na podstawie zachowań klienta można wnioskować o jego skłonnościach do oszczędzania. Monitorowanie salda rachunku bieżącego i transakcji przeprowadzanych przez klienta pozwala stwierdzić, czy klient nie posiada depozytów u konkurencji, czy też ma w wysokim stopniu niezagospodarowane środki, które zalegają na nieoprocentowanym rachunku. Środki takie można alokować przykładowo na lokatę terminową z korzyścią dla korporacji i klienta. Spełnienie jednego z tych warunków gwarantuje, że istnieje pewne prawdopodobieństwo, że klient dany produkt zakupi.

Przedstawiony w pracy projekt spełnia warunki drugiego podejścia. Aplikacja wykonana na potrzeby pracy jest narzędziem wspomagającym doradcę finansowego w określeniu grupy docelowej danego produktu. Pozwala na zarządzanie klientami, raportowanie bieżącej i historycznej

skuteczności sprzedaży oraz na określenie grupy docelowej dla danych produktów, tak aby oferta była skierowana do właściwego klienta spełniającego dane wymagania finansowe.

2. Założenia systemu

Projekt został zbudowany w oparciu o bazę danych SQL Server, która stanowi główne źródło danych dla wszystkich jej modułów. Narzędzia Bussiness Intelligence w MS SQL Server oferują kompleksowe rozwiązania w procesach drążenia danych, ich analizy i raportowaniu. Dane mają być prezentowane za pomocą aplikacji napisanej w języku Visual Basic .Net, pozwalającej użytkownikowi na szybkie i skuteczne pozyskiwanie informacji z systemu. Dodatkowo aplikacja wspomaga podejmowanie decyzji we wszystkich aspektach działalności organizacji przez wszystkich jej pracowników począwszy od doradcy skończywszy na kadrze zarządzającej.

Artykuł prezentuje całościowo proces przepływu danych od momentu pozyskania ich w systemie transakcyjnym, do momentu otrzymania końcowych wyników obliczeń i uzyskania wiedzy ze zgromadzonych danych. System wspomaga proces sprzedaży określonej grupy produktów oferowanych przez daną instytucję finansową, a także proces doradzania zarówno klientowi, jaki i osobie oferującej dany produkt.

Od rozwiązań tego typu wymaga się, aby były kompleksowe i uwzględniały każdy aspekt działalności danej organizacji. Rozwiązanie oparte na budowie systemu ze specjalizowanych modułów ma zagwarantować łatwe wprowadzanie zmian i modyfikacji wynikających z charakterystyki prowadzonej działalności.

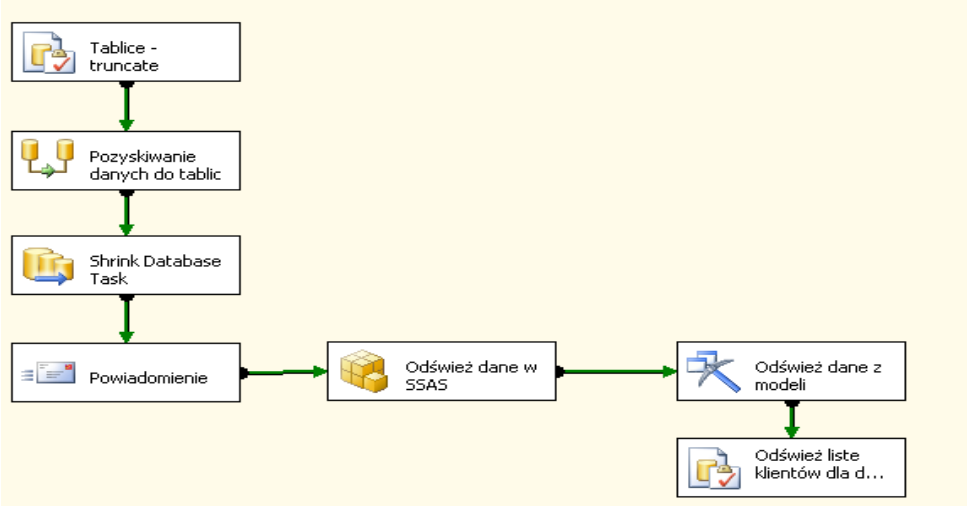
Z uwagi na dużą ilość pozyskiwanych i gromadzonych danych, system powinien oferować wyniki w określonym czasie. Szybko zmieniające się dane i warunki wymagają, aby procesy pozyskiwania wiedzy działały nie dłużej niż kilka godzin, przy zachowaniu ich skuteczności i jakości. Ponieważ dane są pozyskiwane z systemu zewnętrznego, konieczne jest zastosowanie narzędzi ETL, mających na celu szybkie przemieszczanie danych pomiędzy różnymi punktami oraz zapewnienie pełnej integralności.

3. Pozyskiwanie, przetwarzanie i przechowywanie danych źródłowych dla systemu

Każda instytucja finansowa posiada swój własny system do zarządzania klientem oraz jego produktami. Zazwyczaj systemy transakcyjne dostarczają danych do zasilania hurtowni i aplikacji. Z uwagi na charakter takiego systemu oraz jego przeznaczenie, nie jest możliwe przeprowadzanie na nim specyficznych dla hurtowni operacji. Od systemu transakcyjnego wymaga się precyzyjnej i w miarę natychmiastowej odpowiedzi. Zapytania analityczne i związany z tym

duży stopień przetwarzania danych ograniczają znacznie wydajność działania takich systemów. Kolejnym powodem do tworzenia systemów przeznaczonych do zastosowań analitycznych jest bardzo często fakt, że dane pochodzą z wielu źródeł. Zastosowanie podejścia opartego na hurtowni jest często jedyną alternatywą „wyłuskania" informacji.

Dane z systemu transakcyjnego są przetwarzane, wstępnie agregowane i umieszczane w aplikacji.



Rys. 1. Schemat procesów w usłudze SSIS (SQL Server Integration Services)
Fig. 1. Diagram of processes in SSIS service (SQL Server Integration Services)

Całość opisanego procesu można zautomatyzować tak, aby zadania wykonywane były zgodnie z określonym harmonogramem (np. w momencie najmniejszego wykorzystania aplikacji). W projekcie wykorzystano do tego celu narzędzie SQL Server Agent.

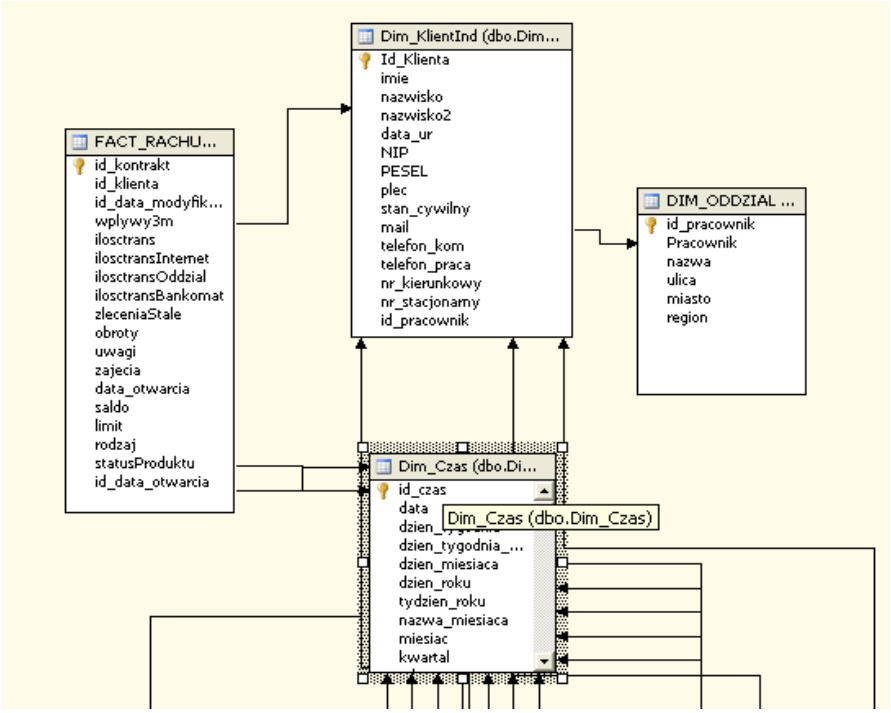
4. Moduł analityczny aplikacji i drążenie danych

4.1. Tworzenie wielowymiarowych raportów z użyciem kostek OLAP

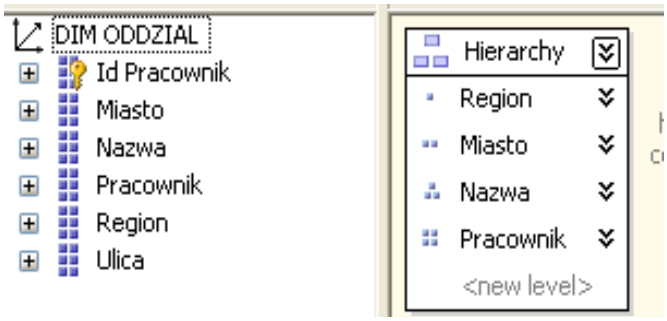
Narzędzia OLAP mają za zadanie prowadzenie dynamicznej analizy danych oraz wspieranie procesów decyzyjnych zachodzących w korporacji. Narzędzia te są stosowane głównie przez kadrę zarządzającą i analityków. Użycie kostek pozwala użytkownikowi na otrzymanie raportu w czytelnej formie, najczęściej przy użyciu tabeli przestawnej. Dodatkowym atutem jest możliwość różnicowania szczegółowości raportu.

Kostki zbudowane są z wymiarów i miar. Wymiary są zmiennymi pozwalającymi na odnalezienie w bazie danej informacji. Miary to wartości przypisane punktom określonym przez wymiary. Inaczej, wymiary są indeksami pozwalającymi na odczytanie wartości miary znajdującej się na przecięciu wartości wszystkich ustalonych wymiarów. Oprócz wymiarów definiuje się również hierarchię wymiaru, która to jest zbiorem wymiarów. Dzięki tworzeniu wymiarów hie-

rarchicznych istnieje możliwość opracowywania raportów dla różnych poziomów szczegółowości.



Rys. 2. Fragment schematu źródła danych ukazujący rachunki
Fig. 2. Data source diagram (fragment) showing bills



Rys. 3. Hierarchia wymiaru pracownik
Fig. 3. Dimension hierarchy of worker

Przygotowanie danych ogranicza się do stworzenia dwóch rodzajów tabel:

- tabele wymiarów – zawierają informacje o wymiarach,
- tabele faktów – zawierają informacje o miarach.

Po opracowaniu struktury hurtowni przystępuje się do budowy kostki. Budowę tę znacząco wspomaga kreator.

4.2. Proces drążenia danych

Data Mining jest „procesem odkrywania znaczących nowych powiązań, wzorców i trendów przez przeszukiwanie dużych ilości danych zgromadzonych w skarbnicach danych, przy wykorzystaniu metod rozpoznawania wzorców, jak również metod statystycznych i matematycznych" [3,4,5].

Narzędzia drążenia danych pozwalają na uzyskanie informacji, które nie są dostrzegalne „na pierwszy rzut oka". W projekcie użyto narzędzi do drążenia danych dla modułu marketingowego aplikacji. Wytypowanie klientów do danej akcji marketingowej mającej na celu ofertę i sprzedaż danego produktu może się odbyć w dwojaki sposób:

- przez podejście eksperckie,
- przez podejście z użyciem narzędzi drążenia danych.

Podejście eksperckie opiera się na wiedzy i doświadczeniu analityka. Analityk za pomocą zwykłego zapytania SQL może wyselekcjonować grupę klientów, która jego zdaniem, może być zainteresowana daną ofertą. Przykładowo, ofertę lokaty można skierować do klientów, którzy:

- posiadają wolne środki na rachunku lub posiadają już inne lokaty,
- ich wpływy na rachunek są znaczące,
- są w wieku powyżej 40 lat (takie osoby mają większą skłonność do oszczędzania).

Takie podejście nie gwarantuje, że wszystkie warunki zostały dostrzeżone i wzięte pod uwagę. Nie ma pewności, że nie istnieją inne cechy, które decydują o tym, że danych produkt zostanie zakupiony lub też nie.

Podejście z użyciem narzędzi drążenia danych pozwala na odkrywanie reguł, których analityk nie byłby wstanie dostrzec. Narzędzia te, wg [3], gwarantują wykonanie poniższych zadań: opis, szacowanie (estymacja), przewidywanie (predykcja), klasyfikacja, grupowanie, odkrywanie reguł.

SSAS (SQL Server Analysis Services) dostarcza kompletny zestaw narzędzi do eksploracji danych, które nie wymagają zaangażowania ekspertów z dziedzin statystycznych. Oferuje następujące modele [2,5,6]:

Microsoft Decision Trees Algorithm – algorytm drzew decyzyjnych jest algorytmem, który najlepsze rezultaty daje przy modelowaniu przewidującym. Jest to głównie algorytm klasyfikacji. Można go zastosować do wszystkich typów danych dyskretnych, jak i ciągłych. Zasada działania opiera się na oszacowaniu, w jakim stopniu zmienne wejściowe wpływają na wartość przewidywaną. Wynikiem działania tego algorytmu jest ukazanie za pomocą struktury drzewa, jak zmienne wejściowe wpływają na zmienną wynikową.

Naive Bayes – algorytm jest możliwy do zastosowania tylko w przypadku dyskretnych wartości. Oblicza on, z jakim prawdopodobieństwem może wystąpić dany stan zmiennych wejściowych i zmiennej wyjściowej. Celem pracy tego algorytmu jest doprowadzenie do sytuacji, w której wszystkie zmienne wejściowe osiągną niezależnie przewidywalne wartości. Z uwagi na

szybkość działania tego algorytmu, ma on szczególne zastosowanie w przypadku wstępnego przeglądania zmiennych. Równie często stosuje się go jako algorytm klasyfikujący i prognozujący.

Microsoft Clustering Algorithm – czyli algorytm grupujący. Jest to algorytm służący do grupowania danych o podobnych cechach. Na podstawie utworzonych w ten sposób klastrów można dokonać analizy struktury danych oraz odkrywać uogólnienia. Ważną cechą tego algorytmu jest zredukowanie dużej ilości danych pierwotnych do podstawowych grup, które są najistotniejsze w badanym problemie.

Microsoft Association Algorithm – reguły asocjacyjne. Jest to rodzaj algorytmu „a priori". Stanowi on metodę pozwalającą na znalezienie korelacji pomiędzy zmiennymi. Pozwala na określenie, które dane są ze sobą powiązane. Zastosowanie tego algorytmu ułatwia także odkrycie powiązań pomiędzy atrybutami, które dotąd nie były znane. Wynikiem działania algorytmu jest zbiór reguł „Jeżeli poprzednik to następnik" wraz z określeniem prawdopodobieństwa wystąpienia danej reguły.

Microsoft Time Series Algorithm – szeregi czasowe. Algorytm ten pozwala na odnajdywanie trendów w danych. Umożliwia prognozowanie zmiennej bądź wielu zmiennych wyjściowych na podstawie wartości ciągłych zmiennej wejściowej. Prognozowanie trendów jest możliwe tylko na podstawie zmiennych użytych przy budowie modelu.

Microsoft Neural Network Algorithm – sieci neuronowe. Algorytm ten jest działem sztucznej inteligencji. Analizuje wszelkie możliwe relacje pomiędzy danymi. Jest on jednym z najdokładniejszych, a zarazem najwolniejszych algorytmów. Obszary, w których jest on stosowany, są bardzo rozległe.

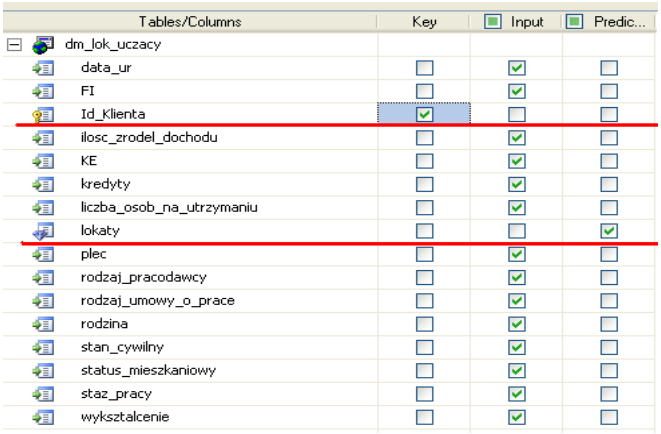
4.3. Modelowanie za pomocą drzew decyzyjnych

Proces tworzenia modelu predykcyjnego zostanie zaprezentowany na przykładzie modelu dotyczącego lokat [1,3]. Zadaniem tego modelu jest przewidywanie, czy dany klient będzie skłonny zakupić lokatę, czy też lepiej takiej oferty mu nie składać. Konieczne jest zdefiniowanie zmiennych, które będą brały udział w procesie modelowania. Najważniejszym punktem jest określenie zmiennej wyjściowej, która musi być dyskretna. W przypadku lokat zmienną określającą, czy klient posiada rachunek terminowy, jest kolumna o nazwie lokaty w tabeli status_ProduktyFinansowe. Zmienna ta jest typu bit i określa, czy dany klient posiada aktualnie lokatę, czy też nie. Na rys. 4 zaprezentowano zmienne biorące udział przy tworzeniu modelu.

Kryteria doboru tych zmiennych są znane tylko analitykowi. Nie ma jednoznacznej definicji, które zmienne wziąć pod uwagę, a których nie brać. Dobór zmiennych pozostaje w gestii twórcy modelu i opiera się na jego doświadczeniu. Jedynym warunkiem, na jaki należy zwrócić uwagę, to aby zmienne wejściowe i wyjściowa nie były ze sobą skorelowane. Innymi słowy, należy uważać, aby w pytaniu nie zawrzeć odpowiedzi.

Rys. 4. Zmienne biorące udział w procesie tworzenia modelu „DM_Lokaty”
Fig. 4. The variables involved in the process of creating a model “DM_Lokaty”

Po wyborze zmiennych i ich przygotowaniu konieczne jest stworzenie trzech zbiorów danych: zbioru uczącego, zbioru testowego, danych produkcyjnych.



Rys. 5. Definiowanie przeznaczenia zmiennych w modelu
Fig. 5. Defining the destination of the variables in the model

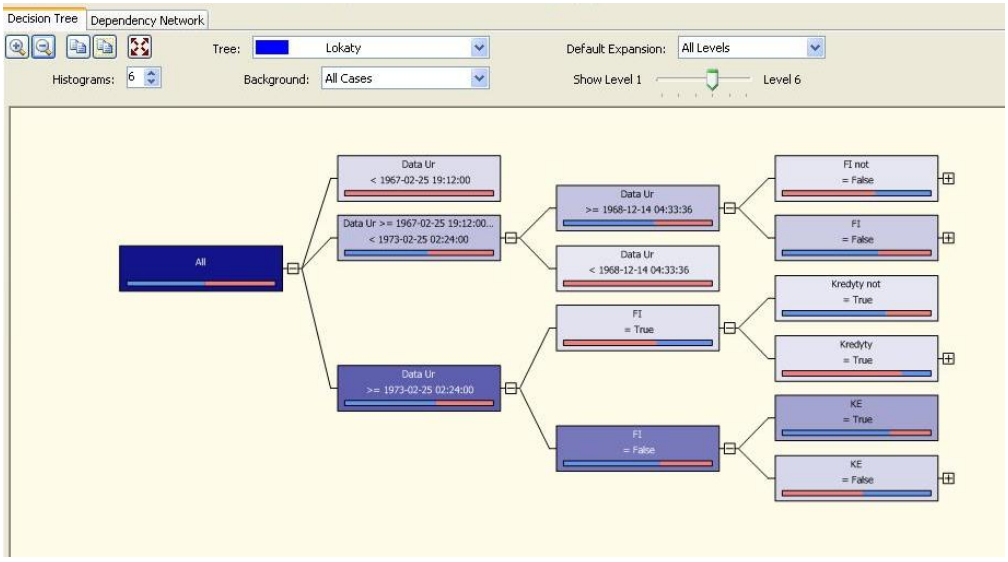
Zbiór uczący, to zbiór danych, na podstawie którego dokona się proces uczenia modelu. Musi się on składać z równolicznych próbek zmiennej przewidywanej. Innymi słowy, zbór ten w naszym przypadku musi się składać w połowie z osób, które posiadają lokatę, jak i tych, którzy tej lokaty nie mają. Istotna jest również liczba rekordów. Im większa próbka rekordów, tym proces uczenia modelu będzie trwał dłużej, przy jednoczesnym zwiększeniu dokładności modelu.

Po dokonaniu wyboru zmiennych wejściowych i przygotowaniu zmiennej wyjściowej można przystąpić do procesu modelowania. Na rys. 5 zaznaczono podział zmiennych. Kolumna Id_Klienta jest kluczem, natomiast kolumna lokaty zawiera zmienną przewidywaną.

Rys. 5 ukazuje strukturę kolumn i ich znaczenie. Zmienne oznaczone jako input to zmienne wejściowe. PredictOnly oznacza kolumnę zawierającą zmienną przewidywaną, natomiast Key jest zmienną zawierającą klucz.

Zakładka Mining Model Viewer pozwala na graficzne zilustrowanie sposobu, w jaki model dokonuje predykcji. Algorytm drzew decyzyjnych można zobrazować za pomocą czytelnego schematu przypominającego swą strukturą drzewo. Taki sposób prezentacji budowy modelu

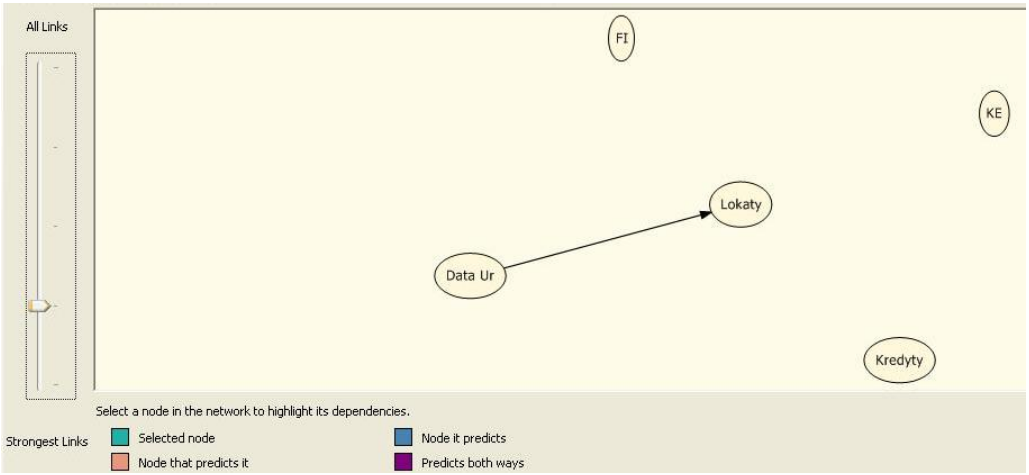
pozwala przedstawić reguły, na postawie których model dokonuje przewidywania. Rys. 6 prezentuje strukturę drzewa dla modelu DM Lokaty.



Rys. 6. Struktura drzewa decyzyjnego w modelu „DM_Lokaty”
Fig. 6. The structure of the decision tree model „DM_Lokaty”

Wybierając za pomocą listy rozwijanej Background wartość True, podświetla się warunki, które określają sukces. Blok o nazwie All jest korzeniem, czyli głównym węzłem decyzyjnym. Jest on połączony z innymi węzłami za pomocą gałęzi. Natomiast ostatnie węzły są nazwane liśćmi. Za pomocą paska umieszczonego pod nazwą węzła zilustrowano prawdopodobieństwo, z jakim może się pojawić dana wartość zmiennej.

Zakładka Dependecncy Network ukazuje sieć zależności pomiędzy określonymi węzłami. Jak widać na rys.7, najważniejszym kryterium jest wiek klienta. Kolejnym ważnym kryterium jest posiadanie funduszy inwestycyjnych, co zapewne ma związek z dywersyfikacją portfela oszczędnościowego klienta i świadczy o długoterminowym oszczędzaniu bądź awersji do ryzyka. Inne ważne kryterium jest to posiadanie przez klienta konta emerytalnego.

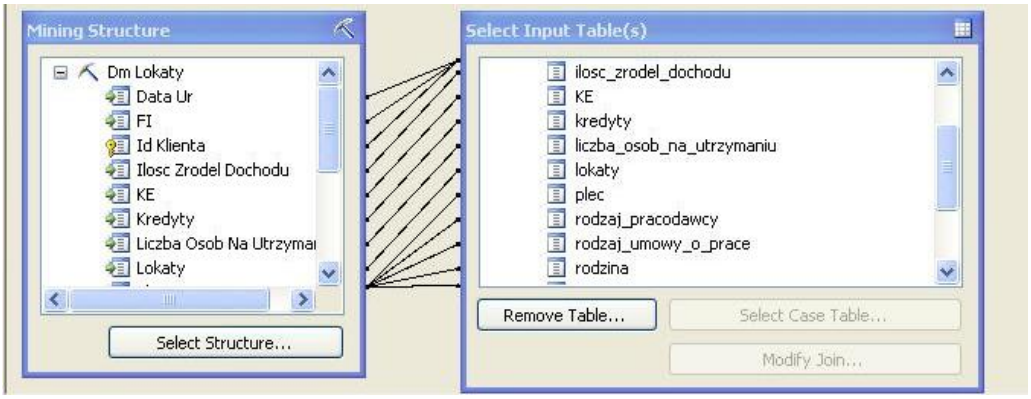


Rys. 7. Sieć zależności w modelu „DM_Lokaty”
Fig. 7. Network model „DM_Lokaty”

4.4. Testowanie modelu

Po opracowaniu modelu należy sprawdzić jego skuteczność. Aby tego dokonać, należy posiadać zbiór danych walidacyjnych, które pokazałyby, jak stworzony model sprawdza się w rzeczywistości. Najwłaściwszą metodą jest przetestowanie zbioru w warunkach rzeczywistych. Jednakże wiąże się to z kosztami (w przypadku niepowodzenia – stratami). W rozpatrywanym przypadku test modelu odbywa się na podstawie danych umieszczonych w hurtowni danych.

Do testowania modelu posłużono się częścią zbioru danych, które nie brały udziału w procesie uczenia. Gdyby użyto zbioru uczącego, model zachowywałby się idealnie, ponieważ byłby testowany na danych, na podstawie których został stworzony. Rysunek 8 prezentuje, w jaki sposób dokonano mapowania zmiennych modelu ze zmiennymi w zbiorze testującym.



Rys. 8. Mapowanie zbioru testującego
Fig. 8. Mapping a set of testing

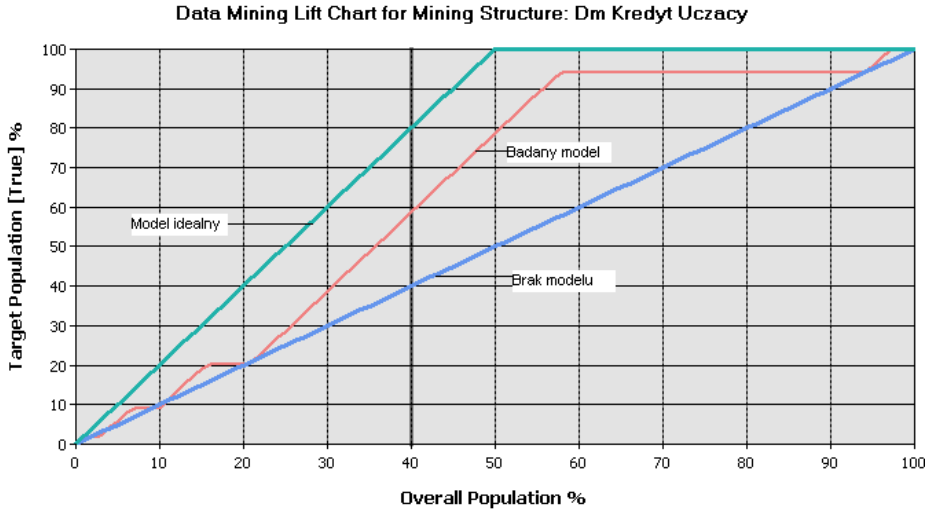
4.5. Badanie skuteczności modelu za pomocą wykresów

Aby móc zbadać skuteczność modelu, jak również porównać stworzone modele pod kątem skuteczności, konieczne jest zobrazowanie tej miary. Zazwyczaj stosuje się graficzną prezentację wyników za pomocą dedykowanych wykresów. Istnieje kilka rodzajów wykresów obrazujących ową skuteczność. Są one pochodnymi dwóch rodzajów wykresów [7]: lift chart (wykres przyrostu), gain chart (wykres korzyści).

Przyrost jest miarą efektywności modelu predykcyjnego, będącą stosunkiem pomiędzy rezultatami działań z użyciem modelu i bez użycia modelu.

Przyrost = (% pozytywnych trafień spośród pozytywnych klasyfikacji) / (% pozytywnych trafień w całym zbiorze). Przyrost jest jednostką bezwymiarową.

Wykresy zysku i przyrostu obrazują graficznie efektywność modeli i pomagają w ocenie ich przydatności. Każdy z wykresów składa się z krzywej przyrostu i tzw. linii bazowej.



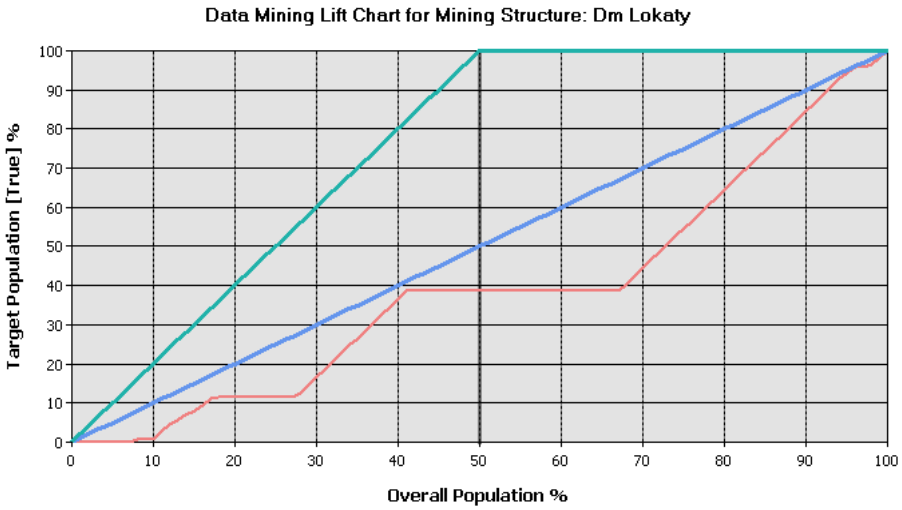
Rys. 9. Skumulowany wykres zysku dla modelu DM_Kredyt
Fig. 9. Cumulative profit chart for model DM_Kredyt

Na rys. 9 zamieszczono model do sprzedaży kredytu. Należy zauważyć, że zastosowanie tak opracowanego modelu przyniesie potencjalne zyski. Wskazuje na to pole zawarte pomiędzy krzywą dolną a środkową.

W tym przypadku pojawiła się kolejna krzywa reprezentująca idealny model. Krzywa górna biegnie narastając liniowo do 50 percentyla, a następnie jest stała aż do 100 percentyla. Punkt przegięcia występuje w 50 percentylu. Dzieje się tak dlatego, ponieważ zbiór testujący składa się w połowie z wartości pozytywnych i negatywnych.

Z porównania zaprezentowanych na rys. 9 i 10 modeli DM_Lokaty i DM_Kredyt wynika, że zastosowanie modelu do lokat okazałoby się niewłaściwe, ponieważ model ten prawdopodobnie przyniesie gorsze rezultaty niż standardowe podejście. Zastosowanie takiego modelu na środowiskach produkcyjnych mogłoby się zakończyć stratami dla korporacji. W takim przypadku do analityka należy decyzja, czy dana grupa zmiennych wejściowych jest poprawna i czy niosą one ze sobą informację, na podstawie której możliwe jest przewidywanie i wybór grupy klientów skłonnych do oszczędzania.

Poniżej przedstawiono model do sprzedaży lokat.



Rys. 10. Skumulowany wykres zysku dla modelu DM_Lokaty
Fig. 10. Cumulative profit chart for model DM_Lokaty

4.6. Macierz klasyfikacji

Kolejnym sposobem na ocenę skuteczności modelu jest macierz błędu klasyfikacji. Zawiera ona informację o aktualnych i przewidywanych klasyfikacjach dokonanych przez system. Skuteczność modelu jest oceniana na podstawie danych zawartych w macierzy, która przypomina tablicę prawdy znaną z algebry Boole'a.

Opierając się na źródłach [3,7] wprowadzono następujące zmienne:

- A – ilość poprawnych klasyfikacji, które są fałszem,
- B – ilość niepoprawnych klasyfikacji, które są prawdą,
- C – ilość niepoprawnych klasyfikacji, które są fałszem,
- D – liczba poprawnych klasyfikacji, które są prawdą.

Tabela 1

Tabela przewidywania klasyfikacji

		Przewidywanie	
		FAŁSZ	PRAWDA
Aktualny stan	FAŁSZ	A	B
	PRAWDA	C	D

Opierając się na źródłach[3,7] wprowadzono wskaźniki:

AC – (accuracy) – dokładność zdefiniowana jako liczba poprawnych klasyfikacji do liczby wszystkich klasyfikacji

$$AC = \frac{A + D}{A + B + C + D} \tag{1}$$

TP – (true positive rate) – współczynnik poprawnej klasyfikacji dodatniej, zdefiniowanej jako stosunek poprawnie zidentyfikowanych przypadków dodatnich do wszystkich pozytywnych przypadków.

$$TP = \frac{D}{C + D} \tag{2}$$

FP – (false positive rate) – współczynnik fałszywej klasyfikacji dodatniej jako stosunek przypadków ujemnych zidentyfikowanych niepoprawnie do wszystkich przypadków ujemnych.

$$FP = \frac{B}{A + B} \tag{3}$$

TN – (true negative rate) – współczynnik poprawnej klasyfikacji ujemnej jako stosunek przypadków ujemnych zidentyfikowanych poprawnie do wszystkich przypadków ujemnych.

$$TN = \frac{A}{A + B} \tag{4}$$

FN – (false negative rate) – współczynnik fałszywej klasyfikacji ujemnej jako stosunek przypadków ujemnych zidentyfikowanych niepoprawnie do wszystkich przypadków dodatnich.

$$FN = \frac{C}{C + D} \tag{5}$$

a.

Counts for Dm Kredyt Uczacy on [Kredyty]:

Predicted	False (Actual)	True (Actual)
False	4802	4914
True	198	86

b.

Counts for DmLokaty DT on [Lokaty]:

Predicted	False (Actual)	True (Actual)
False	8486	9986
True	1514	14

Rys.11. Macierz klasyfikacji dla modelu DM_Kredyt (a) i DM_Lokaty (b)
Fig.11. Classification matrix of model DM_Kredyt (a) and DM_Lokaty (b)

P – (precision) – precyzja – jako stosunek poprawnie zidentyfikowanych przypadków do wszystkich dodatnich przypadków zidentyfikowanych przez model jako poprawne.

$$P = \frac{D}{B + D} \tag{6}$$

Na rys. 11 przedstawiono macierz klasyfikacji dla modelu do sprzedaży produktów kredytowych oraz lokat.

Następnie dokonano obliczenia poszczególnych wskaźników zdefiniowanych powyżej.

Tabela 2
Tabela wskaźników efektywności modeli dla modelu „DM_Kredyt" oraz „DM_Lokaty"

Wskaźnik	Wartość dla modelu „DM_Kredyt"	Wartość dla modelu „DM_Lokaty"
AC	0,49	0,43
TP	0,30	0,01
FP	0,51	0,54
TN	0,49	0,46
FN	0,70	0,99
P	0,02	0,00

Ponieważ w pracy zastosowano tylko jeden z rodzajów modeli, nie jest możliwe porównanie innych rodzajów modeli predykcyjnych zastosowanych do rozwiązania tego samego problemu. W takim przypadku na wykresach korzyści oraz w macierzach klasyfikacji pojawiają się odpowiednie krzywe oraz dodatkowe zmienne, na podstawie których można dokonywać porównania, który model jest właściwszy. Istnieje również metoda zwana champion challenger. W metodzie tej analitycy tworzą kilka modeli predykcyjnych do rozwiązania jednego problemu. Następnie na podstawie narzędzi oceniających skuteczność modeli wybierany jest jeden – o najlepszych wynikach. Staje się on modelem „champion". W miarę napływu danych badane są kolejne modele, także pod kątem skuteczności. W przypadku kiedy okaże się, że jeden z modeli „challenger" osiąga lepszą skuteczność niż aktualnie najlepszy model, zostaje dokonana zamiana i cykl rozpoczyna się od początku.

4.7. Wykorzystanie danych wynikowych

Po opracowaniu modelu predykcyjnego oraz sprawdzeniu go na zbiorze danych testujących, należy wyeksportować dane wynikowe do aplikacji. Nie jest konieczne eksportowanie wszystkich danych na środowiska produkcyjne. Należy określić dolną wartość funkcji prawdopodobieństwa, poniżej której dane nie będą brały udziału w procesach marketingowych. Określenie tej wartości jest uzależnione głównie od budżetu jaki został przeznaczony na daną kampanię marketingową oraz liczności klientów przeznaczonych do kampanii.

Wyniki z modelu pozyskane zostały na podstawie opracowanego zapytania. Do bazy danych zostały wyeksportowane następujące dane:

- Id_Klienta – klucz definiujący jednoznacznie klienta,
- Lokaty – zmienna prawda/fałsz opisująca, czy dany klient będzie zainteresowany danym produktem,
- PredictProbability – wartość funkcji prawdopodobieństwa określająca prawdopodobieństwo wystąpienia wartości Lokaty.

Jako Case Table użyto tabeli z rozszerzeniem _final oznaczającej zbiór danych produkcyjnych, które nie brały udziału w procesie uczenia i testowania modelu. Dodatkowo zbiór wynikowy powinien być tak odfiltrowany, aby wartość zmiennej przewidywanej była zgodna z oczekiwaniami. W analizowanym przypadku wartość zmiennej Lokaty powinna wynosić True.

5. Przykładowa aplikacja z modułem raportującym OLAP

Aplikacja jest warstwą prezentacji całego systemu. Ma ona za zadanie ułatwić użytkownikom posługiwanie się całym systemem oraz zaprezentować dane w przystępnej i czytelnej formie.

Została ona napisana w języku Visual Basic za pomocą środowiska Visual Studio. Aplikacja ma za zadanie prezentowanie wyników działania systemu. Zapewnia ona również interakcję z użytkownikiem. Poniżej przedstawiono założenia, które musi ona spełniać:

- interfejs MDI, pozwalający na jednoczesną pracę z wieloma modułami aplikacji,
- szybki i prosty dostęp do poszczególnych modułów,
- prezentacja możliwie jak największej i obszerniejszej liczby wyników,
- zabezpieczenia przed możliwością nieuprawnionego kopiowania danych,
- proste wyszukiwanie potrzebnych informacji.

5.1. Moduł “Raporty”

Moduł Raporty jest formatką mająca na celu ukazanie obecnego stanu oddziału w graficzny sposób. Raporty takie pozwalają pracownikowi ocenić wyniki pracy całego oddziału korporacji. Dane prezentowane w formie historycznej pozwalają na ocenę pracy danego oddziału nie tylko z bieżących wyników, ale także na długoterminowo. Ponieważ wyniki korporacji zależą w głównej mierze od liczby sprzedanych produktów, zatem taka właśnie wartość jest raportowana.

Do raportowania wykorzystano usługę Sql Server Reporting Services. Wykorzystanie tej usługi daje twórcy raportu możliwość zmiany jego zawartości bez zmiany aplikacji. Innymi słowy, modyfikacja zawartości raportu nie wymaga żadnych ingerencji w aplikację. Drugą istotną sprawą jest to, że wszelkie obliczenia odbywają się na serwerze do tego dedykowanym. Takie rozwiązane daje twórcy dużo swobody w modyfikowaniu treści raportu nie powodując angażowania sporych zasobów na zmiany w kodzie aplikacji i ewentualną instalację na maszynach użytkowników końcowych.

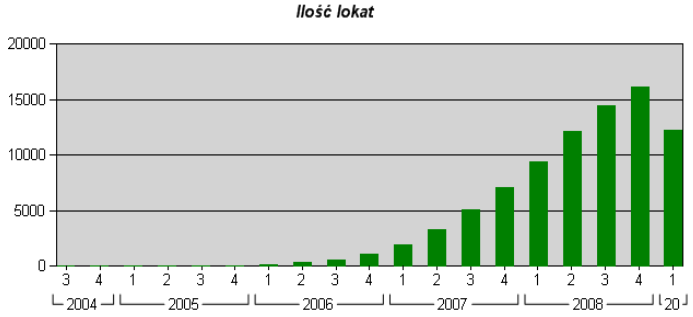
Raporty podzielone są na grupy produktów:

- rachunki bieżące (ROR),
- karty kredytowe,
- kredyty,

- lokaty,
- fundusze inwestycyjne.

Każdy raport zawiera graficzną prezentację w formie wykresu słupkowego sprzedaży danego produktu w określonych kwartałach danego roku. Poniżej wykresu znajduje się tabela, w której przedstawione są konkretne wartości liczbowe.

Dodatkową zaletą wykorzystania tej usługi jest możliwość eksportu danego raportu do innych formatów. Powszechnie wiadomo, że najczęściej wykorzystywaną aplikacją w korporacjach jest arkusz kalkulacyjny. Pracownik chcąc dokonać własnych obliczeń może wyeksportować otrzymany raport np. do formatu Excel i dokonać własnych analiz.



Rok	Kwartał	Miesiąc	Ilość lokat
⊞ 2004			3
⊞ 2005			107
⊞ 2006			2355
⊞ 2007			17497
⊞ 2008			52319
⊞ 2009			12309

Rys. 12. Fragment raportu o ilości lokat
Fig. 12. Report the amount of deposits

5.2. Raport OLAP

Formatka Raporty OLAP zawiera dane przeznaczone tylko dla kadry kierującej daną jednostką. Zawiera ona szczegółowe informacje potrzebne do monitorowania poszczególnych pracowników. Pozwala także na monitorowanie wartości i wolumenów poszczególnych produktów.

Dane takie zarezerwowane są tylko dla wąskiego grona odbiorców gdyż, pokazują wyniki danego oddziału, jak i całej korporacji. Dane te mają charakter poufny i powinny być kierowane tylko do wąskiej grupy odbiorców. Rejestracja wyników sprzedaży każdego z pracowników oraz całej korporacji pozwala na obliczenie ewentualnych premii oraz jest podstawą do nagradzania pracowników. Jeżeli dana korporacja jest notowana na giełdzie papierów wartościowych, jej pracowników obowiązują okresy zamknięte. Wtedy pracownik posiadający akcje spółki, w której pracuje, nie może kupować ani sprzedawać akcji. Wiedza o wyniku finansowym całej korporacji umożliwiłaby nadużycia.

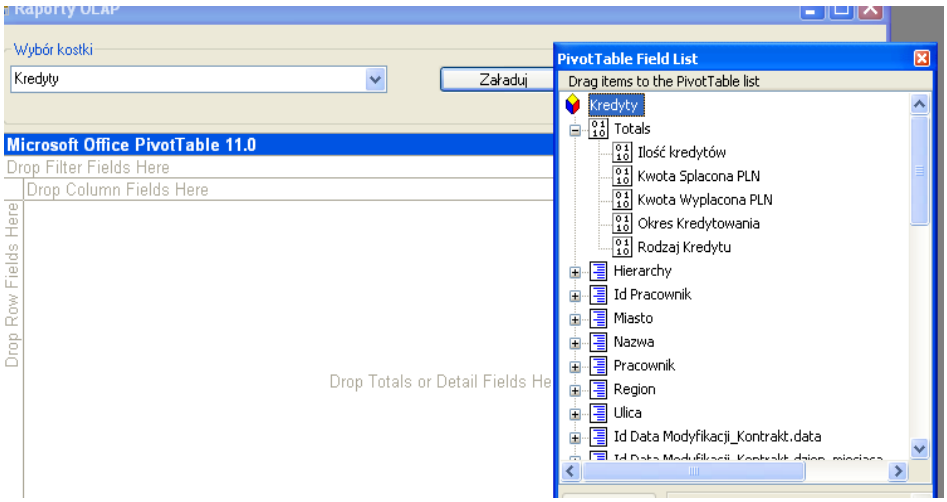
Raporty są podzielone w zależności od danego produktu:

- fundusze inwestycyjne,
- karta kredytowania,
- kredyty,
- lokaty,
- produkty,
- rachunek bieżący,

oraz grupa danych zależnych tylko od oddziału:

- oddziały

Dane te odpowiadają perspektywom zdefiniowanym w kostce. Po wybraniu określonej tematyki raportu użytkownik zobowiązany jest do wciśnięcia przycisku załaduj, który pozwoli na budowę zestawienia.



Rys. 13. Budowanie raportu dostępnych miar i wymiarów
Fig. 13. Building report of available measures and dimensions

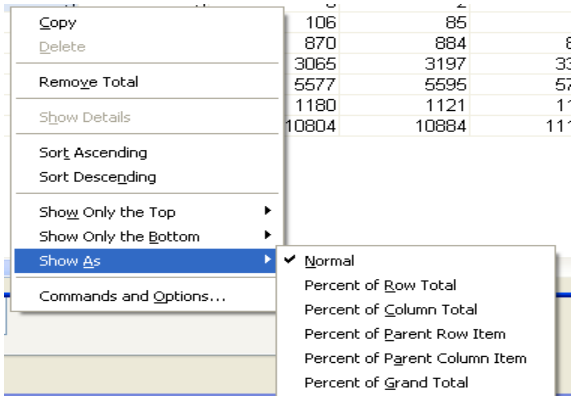
Po lewej stronie ukaże się tabela zawierająca wszelkie wymiary i miary w kostce. Użytkownik przesuwając odpowiednie pola w miejsca wierszy lub kolumn uzyska żądane przez niego zestawienie. Jak widać, rozwiązanie takie pozwala na tworzenie dowolnych zapytań pozwalających na szybki monitoring działań i stanów poszczególnych fragmentów działalności korporacji. Rysunek 14 prezentuje wyniki zapytania. Otrzymany raport ukazuje jak w poszczególnych kwartałach danego roku prezentowała się sprzedaż kredytów w poszczególnych miastach.

	Miasto						
	Gorzów Wlkp.	Katowice	Kielce	Kraków	Lublin	Łódź	Poznań
rok / kwartal	Ilość kredytów	Ilość kredytów	Ilość kredytów	Ilość kredytów	Ilość kredytów	Ilość kredytów	Ilość kredytów
⊞ 2004	3	3	3	6	2	3	10
⊞ 2005	101	164	95	106	85	93	191
⊞ 2006	936	1722	917	870	884	872	1808
⊞ 2007	3381	6298	3407	3065	3197	3300	7060
⊞ 2008	5966	11264	5899	5577	5595	5726	12101
⊞ 2009	1200	2299	1181	1180	1121	1174	2481
Grand Total	11587	21750	11502	10804	10884	11168	23651

Rys.14. Przykładowy raport dotyczący kredytów
Fig.14. Credit report (sample)

Istnieje również możliwość filtrowania danych lub też ich sortowania w trakcie budowy raportu. Użytkownik może zdefiniować także własne podsumowania.

Wszystkie dane prezentowane są przy użyciu kontrolki Microsoft PivotTable 11.0. Jest ona wykorzystywana do prezentacji danych w formie tabeli przestawnej. Ponieważ nie jest ona domyślną kontrolką w środowisku programistycznym Visual Studio, należy zdefiniować referencję do tego obiektu.



Rys. 15. Alternatywne podsumowania w tabeli przestawnej
Fig. 15. Alternative summaries in a PivotTable

6. Moduł aplikacji przeznaczony do sprzedaży produktów

Moduł wspomagający proces sprzedaży produktów jest ściśle powiązany z hurtownią danych oraz narzędziami służącymi do eksploracji danych. Jest to część aplikacji prezentująca pracownikowi wyselekcjonowanych klientów, którzy z dużym prawdopodobieństwem mogą dany produkt zakupić. Selekcja odbywa się przy użyciu modeli drzew decyzyjnych lub też przy pomocy wiedzy eksperta, który wybiera klientów do analizy.

Moduł ten gromadzi i systematyzuje wszelkie prowadzone w korporacji kampanie marketingowe oraz zapobiega nadmiernemu obciążeniu klienta ofertami. Pozwala także kontrolować aktualność oferty, jak i zdolność klienta do nabycia produktu.

Na część odpowiedzialną za oferty marketingowe składają się dwie opcje.

- sprzedaż krzyżowa,
- składanie zamówień.

6.1. Menu„Marketing”

W karcie „Marketing” można wyróżnić trzy obszary:

- kampanie marketingowe,
- kampanie dla wybranego klienta,
- skuteczność działań.

Obszar odpowiedzialny za kampanie marketingowe prezentuje wszelkie aktualnie prowadzone kampanie marketingowe. Podzielone są one na dwie grupy:

- kampanie przeznaczone dla całej korporacji,
- kampanie dla oddziału.

Powyższy podział wynika z faktu, że oddział może przeprowadzać akcje uwarunkowane geograficznie lub też przygotowane we własnym zakresie. W przypadku słabych wyników w sprzedaży danego produktu dany oddział może zamówić indywidualną kampanię mającą na celu poprawienie wyników. Kampanie korporacyjne przeprowadzane są dla całej korporacji i dla wszystkich oddziałów włącznie.

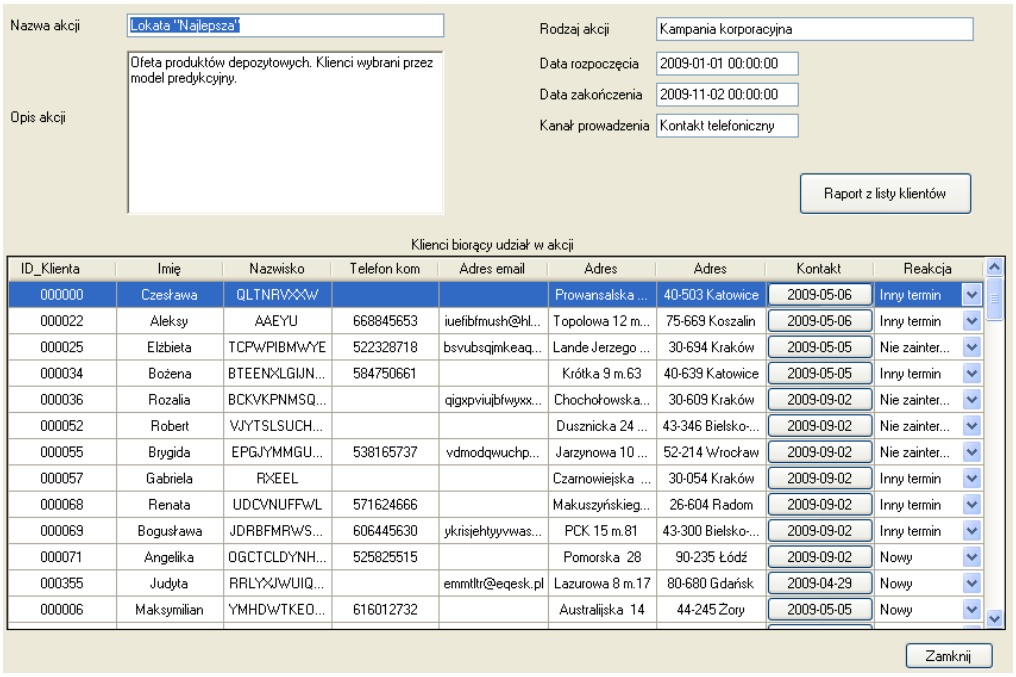
Pracownik może na bieżąco monitorować swoje działania na podstawie modułu obliczającego skuteczność działań. Lista Skuteczność działań zawiera listę wszystkich kampanii, zarówno korporacyjnych, jak i dla oddziału, dla których badana jest skuteczność działań marketingowych.

Jednym z elementów modułu jest opcja „Marketing". Menu „Marketing" zawiera dane szczegółowe dotyczące danej akcji marketingowej. Okno „Marketing" składa się z dwóch zasadniczych części. Pierwsza zawiera szczegółowy opis kampanii oraz podstawowe jej parametry. Pozwala to użytkownikowi przygotować się do pracy z daną bazą klientów. Natomiast drugą część stanowi tabela zawierająca listę klientów biorących udział w danej akcji marketingowej.

Tabela „Klienci biorący udział w akcji" zawiera spis wszystkich klientów, którzy zostali wytypowani przez system do udziału w danej kampanii marketingowej. Prezentuje ona podstawowe dane kontaktowe oraz pozwala określić reakcje klienta na daną ofertę.

Kolumna Reakcja pozwala użytkownikowi określić reakcję klienta na daną akcję marketingową. Po zakończeniu kontaktu, pracownik zobowiązany jest wypełnić to pole. Ma ono na celu określić wstępne wyniki danej rozmowy. Wartości tego pola można wykorzystać do dalszych badań. Możliwa reakcja klienta:

- zainteresowany – klient odpowiedział na ofertę pozytywnie,
- nie zainteresowany – klient odpowiedział na ofertę negatywnie,
- brak kontaktu - podane dane teleadresowe uniemożliwiają kontakt z klientem lub są nieaktualne,
- inny termin – propozycja innego terminu spotkania.



Rys. 16. Okno „Marketing”
Fig. 16. "Marketing" window

Aplikacja ma też możliwość badania skuteczności działań danej kampanii, ma na celu ukazanie pracownikowi efektów jego działań. Pokazuje, jaki sukces odniosła dana kampania i jakie były korzyści z niej płynące, a także statystykę działań pracownika i określa jego wydajność.

Na karcie marketing użytkownik może wpisać w obszarze oznaczonym jako Kampania dla klienta dane jednoznacznie identyfikujące klienta. W przypadku kiedy podane dane są prawidłowe i istnieją w systemie, pracownikowi zostanie wyświetlone okno o nazwie „Akcja dla klienta”.

7. Podsumowanie

W niniejszej pracy starano się zaprezentować, w jaki sposób współczesne korporacje radzą sobie z przetwarzaniem gromadzonych i pozyskiwanych danych. Przedstawiono sposoby raportowania, prezentacji informacji dla kadry zarządczej, mające na celu usprawnienie procesu podejmowania decyzji.

Przy projektowaniu systemu uwzględniono różnorakie aspekty działalności korporacji. Przeprowadzono analizę pod kątem oferowanych produktów i możliwości ewentualnych zmian. Rozpatrzono również strukturę organizacyjną korporacji, zwracając uwagę na hierarchiczność i funkcje pełnione przez poszczególnych pracowników organizacji. Miało to szczególne znaczenie przy budowie modelu uprawnień do poszczególnych obiektów bazy danych oraz modułów aplikacji.

W pracy ukazano różne aspekty prowadzenia kampanii marketingowych uwzględniając możliwe kanały dotarcia do klienta z daną ofertą. Wymagało to również uwzględnienia specyfiki danego kanału i uwarunkowań właściwych dla każdego ze sposobów.

Z uwagi na brak możliwości, nie przeprowadzono szczegółowych badań dotyczących skuteczności przedstawionych w pracy modeli predykcyjnych. Badania takie wiązałyby się z koniecznością wprowadzenia ich do systemu CRM danej korporacji oraz wymagałyby długiego okresu badania skuteczności, tak aby jednoznacznie stwierdzić ich przydatność. Jakkolwiek takie badanie nie odzwierciedla rzeczywistej skuteczności modelu na zbiorze produkcyjnych danych, pozwala na miarodajne określenie skuteczności modelu.

BIBLIOGRAFIA

1. Microsoft: SQL Server Analysis Services Tutorial , http://technet.microsoft.com/enus/ library/ms170208.aspx, 2008.
2. Brown E.L.: SQL Server 2005 - Wyciśnij wszystko. Gliwice 2007.
3. Larose D.T.: Odkrywanie wiedzy z danych. Warszawa 2006.
4. Wikipedia: Eksploracja danych , http://pl.wikipedia.org/wiki/Eksploracja_danych, 2009.
5. SPSS Inc.: Pomoc do programu PASW Modeler, http://www.spss.pl/narzedzia/ clementine.html, 2009.
6. StatSoft: Metody Data Mining, http://www.statsoft.pl/dataminer2.html, 2009.
7. Hamilton H.J.:Cumulative Gains Chart and Lift Chart, http://www2.cs.uregina.ca/ ~hamilton/courses/831/index.html, 2007.
8. Microsoft: Reporting Services Tutorials, http://msdn.microsoft.com/enus/library/ ms170246 (SQL.90).aspx, 2008.

Recenzent: Dr inż. Marcin Gorawski

Wpłynęło do Redakcji 31 stycznia 2010 r.

Abstract

The paper presents the way in which modern corporations store, process and analyse collected data. Created system aims at facilitating selling process, conducting customer selection and estimating the probability of specified product purchase by the client. It also supports marketing campaigns carried out by entire corporation as well as its separate organizational units.

Moreover, it enables customer relation management and provides help for client's financial advisory process.

Adresy

Paweł DRZYMAŁA: Politechnika Łódzka, Instytut Mechatroniki i Systemów Informatycznych, ul. Stefanowskiego 18/22, 90-924 Łódź, Polska, pdrzymal@p.lodz.pl.

Henryk WELFLE: Politechnika Łódzka, Instytut Mechatroniki i Systemów Informatycznych, ul. Stefanowskiego 18/22, 90-924 Łódź, Polska, welfle@p.lodz.pl.

Sławomir WIAK: Politechnika Łódzka, Instytut Mechatroniki i Systemów Informatycznych, ul. Stefanowskiego 18/22, 90-924 Łódź, Polska, wiakslaw@p.lodz.pl.