

*Mariusz Łapczyński**

DATA MINING W BADANIACH RYNKOWYCH I MARKETINGOWYCH – OBSZARY ZASTOSOWAŃ

1. Wprowadzenie

Data mining to zestaw narzędzi analitycznych¹ (m.in.: sieci neuronowych, drzew klasyfikacyjnych i regresyjnych, reguł skojarzeniowych), które pozwalają tworzyć modele deskryptywne i predykcyjne. Te pierwsze pozwalają poznać badane zjawisko za pomocą zestawu reguł, natomiast te drugie umożliwiają budowę prognoz. Termin „data mining” jest kojarzony z terminem „database marketing” (marketing oparty na bazach danych). Jako etap procesu odkrywania wiedzy w bazach danych ma z nimi wiele wspólnego. To właśnie ilość gromadzonych danych i moc obliczeniowa komputerów przyczyniły się do rozkwitu tej dziedziny na początku lat 90. ubiegłego stulecia. Ścisły związek z bazami danych nie oznacza jednak, że narzędzia data mining mają być stosowane wyłącznie do terabajtów informacji w hurtowniach danych. Z powodzeniem zastępują klasyczne statystyczne metody analizy danych ankietowych, gdzie zbiór obserwacji liczy z reguły mniej niż 2 tys. przypadków.

2. Obszary zastosowań data mining

Data mining znajduje zastosowanie w analizie² rentowności klienta, w badaniach związanych z pozyskaniem nowych nabywców, w analizie

* Mgr, Katedra Analizy Rynku i Badań Marketingowych, Akademia Ekonomiczna w Krakowie.

¹ M. Łapczyński, *Wprowadzenie do data mining*, „Zeszyty Naukowe AE w Krakowie” (w druku).

² R. P. Peacock, *Data Mining in Marketing: Part 1*, „Marketing Management” 1998, winter.

sprzedaży krzyżowej, w badaniu przyczyn nielojalności konsumentów, w badaniach segmentacyjnych, w analizie koszykowej, a nawet w analizie danych kwestionariuszowych.

2.1. Analiza rentowności klienta (*customer profitability*)

Podstawowym obszarem zastosowań data mining w praktyce jest analiza rentowności klientów. Jej zdefiniowanie jest punktem wyjścia budowy modelu predykcyjnego. Rentowność może być rozumiana jako wartość współczynnika ROI lub jako wartość życiowa klienta. Ponieważ jej wysokość zmienia się w miarę upływu czasu, istnieje konieczność stałej analizy kontaktów klienta z firmą i wielkości sprzedaży z każdej transakcji. To powoduje, że zamiast o CRM mówi się czasami o CPM (Customer Profitability Management)³.

Narzędzia data mining pozwalają odkryć wzorce i profile konsumentów, na podstawie których menedżerowie są w stanie przewidzieć ich rentowność. Trzeba tylko precyzyjnie zdefiniować zmienną zależną, która w najprostszej postaci może być różnicą między przychodem ze sprzedaży a kosztami produkcji, sprzedaży, obsługi i marketingu. W sytuacji gdy bieżąca i potencjalna wartość klienta są jednakowe, dąży się do minimalizacji kosztów związanych z utrzymaniem sprzedaży na dotychczasowym poziomie. Jeśli z kolei potencjalna wartość klienta jest wyższa od obecnej, to poszukuje się informacji o tym, jak najtaniej spowodować przejście klienta do bardziej dochodowej grupy.

Analizując dane, trzeba zwrócić uwagę na to, by mierzyć wzrost rentowności klienta. Czasem może to być wskaźnik trudny do oszacowania, niemniej jednak jest to najlepsza miara oceniająca skuteczność akcji promocyjnej w marketingu bezpośrednim. Narastająca rentowność klienta (*incremental customer profitability*) to różnica zysku i kosztów promocji. Zysk jest tu traktowany jako przyrost zysku spowodowany akcją promocyjną.

2.2. Pozyskiwanie klientów (*customer acquisition*)

Dotychczasowe podejście do pozyskiwania klientów opierało się na kombinacji marketingu masowego i marketingu bezpośredniego. W marketingu masowym wykorzystywano m.in. reklamę w mediach lub reklamę zewnętrzną. Odpowiednie nośniki reklamy dobierano tak, aby profil słuchacza czy widza

³ A. Berson, S. Smith, K. Thearling, *Building Data Mining Applications for CRM*, McGraw-Hill, USA 2000.

był zgodny z charakterystyką dotychczasowych klientów firmy. W marketingu bezpośrednim używano takich technik, jak telemarketing czy reklama pocztowa. Z bazy danych wybierano te osoby, które spełniały określone kryteria demograficzne. Wraz ze wzrostem liczby cech i, przede wszystkim, objętości baz danych, konieczne było budowanie modeli predykcyjnych. Były to modele statystyczne i w przypadku dużych zbiorów obserwacji modele data mining.

Korzystając z narzędzi data mining, należy odpowiednio zakodować zmienną zależną, którą w tym przypadku jest reakcja odbiorcy na przekaz reklamowy. Jeśli jest to reakcja typu „tak-nie” („kupi-nie kupi”), to stosuje się kod binarny. Jeśli reakcja jest wielowariantowa, to zmienna zależna jest wielokategorialna. Kategoria może oznaczać rodzaj nabytego produktu, jego markę lub przedział cenowy, w jakim się znajduje. Liczba i rodzaj kategorii zależy od problemu badawczego. Ogólnie rzecz biorąc, istnieją 3 możliwe pozytywne reakcje na ofertę kupna:

- a) zapytanie o produkt – klient stara się dowiedzieć czegoś więcej o ofercie firmy, pyta o szczegóły dotyczące ceny, koloru, modelu itp.;
- b) zakup produktu – to klasyczny sukces, klient reaguje zakupem, wybiera z oferty reklamowany produkt;
- c) zakup innego produktu – klient wybiera z oferty firmy produkt inny, niż reklamowany.

Jeśli chodzi o reakcje negatywne, to rozróżnia się 2 jej rodzaje:

- a) klient nie odpowiada na ofertę – brak odpowiedzi nie oznacza odmowy, gdyż może być spowodowany np. nieobecnością klienta lub nieotrzymaniem oferty;
- b) odmowa – to klasyczne niepowodzenie, z badawczego punktu widzenia informacje o osobach nie zainteresowanych ofertą przedsiębiorstwa są równie cenne, pozwalają stworzyć profil takiej osoby i usprawnić ewentualne przyszłe działania marketingowe.

2.3. Sprzedaż krzyżowa (*cross-selling*)

Innym obszarem, w którym znajdują zastosowanie modele data mining, jest sprzedaż krzyżowa. Interakcje między przedsiębiorstwem a klientami są dynamiczne, a zadanie marketerów polega na optymalizacji tej współpracy. Można jej dokonać na 3 sposoby:

- poprzez wydłużenie czasu współpracy;
- poprzez zwiększenie liczby transakcji w czasie trwania współpracy;
- poprzez zwiększenie zysku firmy w kolejnych transakcjach.

Sprzedaż krzyżowa (*cross-selling*) polega na oferowaniu dotychczasowym klientom nowych produktów lub usług. Przykładowo: nabywcom pieluch

proponuje się po pewnym czasie zakup rowerków dziecięcych. Choć ten przykład jest bardzo prosty, to w rzeczywistości narzędzia data mining pozwalają na znalezienie bardziej wyszukanych wzorców zakupów. Jedną z odmian sprzedaży krzyżowej jest sprzedaż uzupełniająca (*up-selling*), która polega na tym, że dodatkowe produkty lub usługi są ściśle związane z produktem (usługą) aktualnie użytkowanym przez konsumenta. Przykładem jest sprzedaż droższego pakietu dla abonenta, który jest już podłączony do sieci telefonicznej.

Przed przystąpieniem do budowy modelu predykcyjnego data mining należy sprecyzować, co będzie stanowić dodatkową ofertę dla klienta. Kolejny krok to zebranie zbioru danych, którym zazwyczaj jest istniejąca w przedsiębiorstwie hurtownia danych. Analizie poddawane są przeszłe transakcje. Proces analizy sprzedaży krzyżowej jest trzyetapowy:

1. Modelowanie zachowań indywidualnych klientów (dla każdej dodatkowej oferty buduje się osobny model predykcyjny).

2. Nadawanie ocen klientom (w najprostszej postaci chodzi o prawdopodobieństwo pozytywnej reakcji na dodatkową ofertę firmy) i konstruowanie macierzy ocen.

3. optymalizacja macierzy ocen – może być dokonana 4 metodami:

- a) naiwną (*naïve*) – opartą na prawdopodobieństwie zakupu,

- b) średnią ekonomiczną (*average economic*) – opartą na średnim zysku z transakcji,

- c) zindywidualizowaną ekonomiczną (*individual economic*) – poszerzoną o dodatkowe informacje finansowe o kliencie,

- d) ograniczoną (*constraint optimization*) – zawierającą dodatkowe zewnętrzne ograniczenia, np. maksymalną liczbą ofert na segment (region); wprowadzenie tych ograniczeń powoduje, że oszacowany zysk przedsiębiorstwa będzie mniejszy lub równy zyskowi oszacowanemu innymi metodami.

2.4. Badania przyczyn nielojalności klientów (*churn analysis*)

Termin „churn” jest stosowany w branży telekomunikacyjnej i oznacza odejście klientów do konkurencyjnych operatorów sieci komórkowych. W praktyce może oznaczać utratę klientów na rzecz konkurencji w dowolnej branży. Istnieją dwa podejścia do szacowania wskaźnika odejścia⁴. Pierwsze opiera się na metodzie RFM (*recency, frequency, monetary*), która w dużym skrócie polega na zbadaniu, kiedy ostatnio dokonano zakupu (*recency*), z jaką częstotliwością dokonywano zakupów w badanym okresie (*frequency*)

⁴ M. A. P. M. Lejeune, *Measuring the impact of data mining on churn management*, „Internet Research: Electronic Networking Applications and Policy” 2001, Vol. 11, No. 5.

i jaka była wartość poszczególnych zakupów (*monetary*). Analizuje się tu wskaźnik *recency*, który jest tożsamy ze wskaźnikiem odejścia. Przykładowa reguła może brzmieć: „jeśli klient dokonał zakupu w ciągu ostatnich 8 miesięcy, to uznaje się go za klienta aktywnego”. To uogólnienie ma tę wadę, że na różnych rynkach i w różnych segmentach inna może być częstotliwość zakupu.

Drugie podejście to metoda VAL (*value, activity, loyalty*), w której szacuje się indywidualne zachowanie klienta na podstawie większej liczby przypadków. O ile w pierwszej metodzie o wysokości wskaźnika odejścia dla pojedynczego klienta decydował wskaźnik *recency* dla tej konkretnej osoby, o tyle w drugiej metodzie wskaźnik odejścia szacowany jest z większej liczby przypadków (np. danego segmentu).

Zadanie data mining polega na predykcji klientów nielojalnych i pośrednio na oszczędności kosztów związanych z pozyskaniem nowych. Najlepszym narzędziem do badania przyczyn nielojalności klientów są drzewa klasyfikacyjne, mimo że równie skuteczne w budowie prognoz są sztuczne sieci neuronowe. Atutem drzew jest ich czytelny i łatwy w interpretacji model graficzny. Narzędzie to pozwala na podejście deskryptywno-predykcyjne.

2.5. Badania segmentacyjne

Segmentacja i selektywność rynku to kolejny obszar zastosowań narzędzi data mining. Przykładowo, drzewa klasyfikacyjne i regresyjne pozwalają analizować zmienne ilościowe i jakościowe. Zwłaszcza te drugie dość często występują w badaniach segmentacyjnych. Identyfikuje się homogeniczne grupy konsumentów na podstawie ich stosunku do marek produktów lub innych instrumentów marketingowego oddziaływania. Coraz popularniejsza staje się segmentacja psychograficzna konsumentów. To właśnie niemierzalny charakter zmiennych opisujących segmenty i niemierzalne cechy produktów w badaniach selektywności powodują, że coraz częściej do profilowania segmentów używa się narzędzi data mining.

2.6. Analiza koszykowa (*market basket analysis*)

Analiza koszykowa opisuje transakcje dokonywane przez klientów supermarketów za pomocą reguł skojarzeniowych (*association rules*). Reguła skojarzeniowa przyjmuje postać: „jeśli zdanie Z_1 , to zdanie Z_2 ” np. „jeśli kupił produkt A , to kupił produkt B ”. Z formalnego punktu widzenia jest to zdanie warunkowe, w którym produkt A jest poprzednikiem, a produkt B następnikiem. Reguły skojarzeniowe dzielimy na jakościowe (Boolowskie)

i ilościowe. W przypadku tych pierwszych uzyskuje się informacje o tym, jaki produkt z następnika został kupiony razem z produktem z poprzednika. Co do ilościowych reguł skojarzeniowych, to poza informacją na temat rodzaju produktu, uzyskuje się informację o nabytej ich liczbie np. „jeśli kupił 2 kg produktu *A*, to kupił 3 sztuki produktu *B*”. Istnieją też inne podziały reguł skojarzeniowych:

- na jednowymiarowe (zawierające jeden poprzednik) i wielowymiarowe (zawierające kilka poprzedników np. „jeśli kupił produkt *A* i produkt *C*, to kupił produkt *B*”),

- na jednopoziomowe i wielopoziomowe (zawierające bardziej szczegółowe informacje o nabytych produktach np. „jeśli kupił produkt *A* marki *Z* w opakowaniu 0,5 l, to kupił produkt *B* marki *Y*”).

Analiza koszykowa znajduje uznanie u specjalistów ds. merchandisingu (pozwala na efektywne rozmieszczenie asortymentu), specjalistów ds. promocji (wiedzą, jakie produkty powinny być promowane razem) i specjalistów ds. logistyki (przewidzą, jak braki w zaopatrzeniu jednych produktów wywołują spadek sprzedaży innych).

2.7. Analiza danych kwestionariuszowych

Alternatywą dla tabel kontyngencji, często stosowanych w badaniach kwestionariuszowych, jest jeden z algorytmów drzew klasyfikacyjnych – CHAID⁵. Tu również wykorzystuje się test niezależności chi-kwadrat, a automatyzacja obliczeń znacznie przyspiesza analizę zbiorów obserwacji z wieloma zmiennymi. CHAID nie mierzy dokładnie siły związku między zmiennymi, nie można porównywać rankingu ważności predyktorów z wartościami współczynnika *V* Cramera. CHAID jest natomiast bardziej przejrzysty, pozwala dokonać profilu kategorii zmiennej zależnej za pomocą zestawu reguł o postaci „jeżeli... to...”, a nie jak w przypadku analizy tabelarycznej za pomocą nieczytelnej tabeli o bardzo dużych rozmiarach.

3. Podsumowanie

Gwałtowny rozwój informatyki na początku lat 90. przyczynił się do popularyzacji dziedziny zwanej data mining. Duża elastyczność narzędzi analitycznych przejawiająca się możliwością analizy zmiennych wyrażonych

⁵ M. Łapczyński, *Analiza porównawcza tabel kontyngencji i metody CHAID*, „Zeszyty Naukowe AE w Krakowie” (w druku).

na różnych poziomach pomiaru spowodowała, że szybko znalazły się one w kręgu zainteresowań badaczy marketingowych. Zaczęto je wykorzystywać nie tylko w marketingu opartym na bazach danych, ale również w analizie mniejszych zbiorów przypadków. Przedstawione w tym artykule obszary zastosowań data mining są tylko krótkim wprowadzeniem w tę interdyscyplinarną dziedzinę.

Mariusz Łapczyński

DATA MINING APPLICATIONS IN MARKETING RESEARCH

The purpose of this article is to describe data mining applications in marketing research. The author outlined shortly main CRM areas: customer acquisition, customer profitability, churn analysis and cross-selling. He also focused on segmentation research, market basket analysis and analysis of questionnaire data with CHAID algorithm.