PRZEWIDYWANIE LOJALNOŚCI KLIENTÓW

Tomasz Demski
StatSoft Polska Sp. z o.o.

Wprowadzenie

Podstawowe znaczenie terminu „lojalność klienta" oznacza jego skłonność do wielokrotnych zakupów u jednego dostawcy. Często stosuje się również szersze rozumienie tego pojęcia: lojalny klient nie tylko często dokonuje zakupów, ale ponadto ma dobrą opinię o dostawcy i jest gotów dzielić się nią z innymi. Dyskusję zagadnienia lojalności klienta przedstawiono w artykule „Lojalność, satysfakcja - ich znaczenie i pomiar" w [1] (artykuł jest dostępny również na stronie www.statsoft.pl/czytelnia/marketing/wprowmarketing.html).

Zagadnienie lojalności klientów jest bardzo istotne w wielu dziedzinach gospodarki, zwłaszcza tam, gdzie z jednej strony potrzebne są duże nakłady, aby pozyskać klienta, a z drugiej strony klienci nie przywiązują się do dostawcy. Wzorcowym przykładem jest telefonia komórkowa. Wystarczy powiedzieć, że z badań przeprowadzonych w USA wynika, że 90% posiadaczy telefonów komórkowych zmieniło operatora w ciągu ostatnich 5 lat (por. [2] i [3]).

Rolę lojalności podsumowuje poniższy wzór (tzw. reguła sukcesu CRM, ang. *CRM formula for success*, [4]):

Przychód = (Liczba klientów) • (Przeciętna wartość) • (Lojalność)

Jest to oczywiście duże uproszczenie skomplikowanych zależności, jednakże unaocznia rolę lojalności klientów.

Analiza danych a lojalność klientów

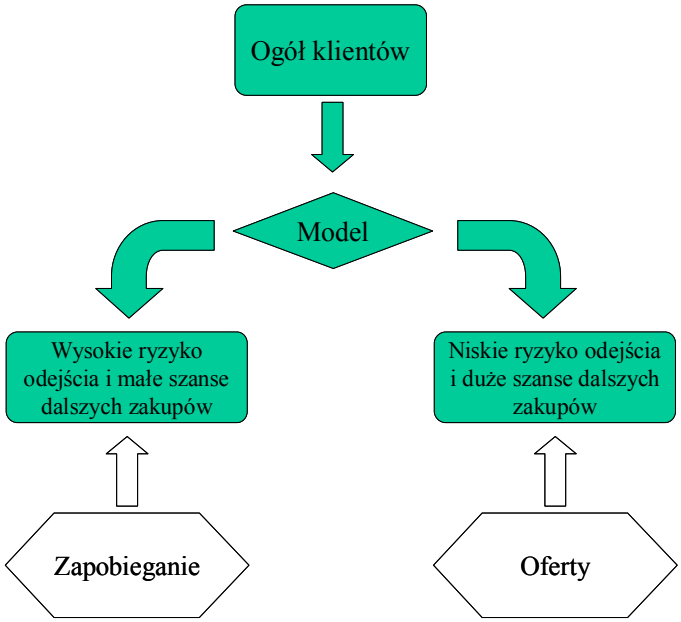
Analizy danych stosowane są w kontekście lojalności klientów do badania i monitorowania satysfakcji klienta, wykrywania przyczyn decydujących o zadowoleniu klienta i jego lojalności, monitorowania zmian stopnia satysfakcji klientów. Przegląd i przykłady zastosowań tego typu przedstawiono w pracy [1].

Inna grupa zastosowań analizy danych to metody odkrywania wzorców zachowań klientów, np. jakie są typowe sekwencje zdarzeń i zakupów (co poprzedza odejście klienta).

Ważną rolę odgrywa również segmentacja klientów, umożliwiająca znalezienie jednorodnych grup podobnych nabywców (zob. artykuł poświęcony segmentacji w niniejszej publikacji).

Często poszukuje się reguł decydujących o tym, że ktoś przestaje kupować nasze produkty lub usługi (tzw. model *churn*). Przykład takiej analizy znajduje się w pracy [5] oraz podręcznikach [2] i [3]. My zajmiemy się odwrotnym zagadnieniem: przewidywaniem, kto pozostanie naszym klientem, a mówiąc bardziej technicznie, zajmiemy się tworzeniem modeli lojalności klienta.

Modele lojalności klienta (nazywane też modelami utrzymania klienta, w języku angielskim używa się wymiennie terminów *loyalty*, *attrition* i *retention models*) służą do przewidywania prawdopodobieństwa, że dana osoba (lub firma) pozostanie naszym klientem. Model budujemy metodami analizy danych (statystyki i *data mining*) – wskazuje na to choćby przewidywanie **prawdopodobieństwa**. Oczywiście nie jesteśmy w stanie z całkowitą pewnością przewidzieć, że dany klient będzie lojalny. Możemy jednak podzielić klientów na grupy o różnym poziomie zagrożenia odejściem z jednej strony, a różnymi szansami na dalsze zakupy z drugiej strony. Działanie modelu przedstawia poniższy schemat (dodatkowe informacje można uzyskać w artykule [6]).



W stosunku do obu grup klientów wydzielonych przez model podejmujemy odmienne działania. W szczególności nielojalni klienci mogą być celem działań zapobiegawczych (typu oferta zniżki na abonament). Natomiast do lojalnych klientów będziemy kierować oferty zakupu nowych produktów (*cross-selling*), czy też rozszerzenia obecnej współpracy (*up-selling*). Zauważmy, że wysyłanie propozycji zapobiegających odejściu do lojalnych

klientów jest szkodliwe (bo np. obniżymy abonament komuś, kto wcale nie zamierzał odejść).

Model lojalności tworzymy, wykorzystując dane z przeszłości:

- o zachowaniu klientów (zakupy i kontakty z dostawcą),
- o cechach demograficznych klienta (wiek, wykształcenie, region itp.),
- dane zewnętrzne (np. o poziomie bezrobocia w miejscu zamieszkania klienta).

Zwróćmy uwagę, że model da nam informację, które cechy klientów wpływają na poziom lojalności. Na jej podstawie możemy np. kierować kampanie marketingowe do tych potencjalnych klientów, którzy najprawdopodobniej na długo pozostaną naszymi klientami. Zauważmy, że z reguły sukcesu CRM (str. 51) wynika, że pozornie nieinteresujący klient (np. kupujący za niewielkie kwoty) przyniesie największe dochody (ponieważ będzie kupował przez długi czas).

Ponadto model lojalności może wskazać również pewne wydarzenia, które poprzedzają odejście lub wpływają na utrzymanie klienta.

W kolejnej części zajmiemy się budową modelu lojalności klientów (przy przygotowaniu artykułu korzystaliśmy z podręcznika [3]).

Przykład

Zbudujemy model przewidujący, którzy klienci dokonają ponownego zakupu. Model będzie stosowany niedługo po pierwszym zakupie, a więc w jego budowie musimy ograniczyć się do danych dostępnych po realizacji pierwszego zamówienia.

Do dyspozycji mamy dane:

- o zamówieniu (wartość pierwszego zakupu, liczbę i kategorię nabytych produktów, sposób dostawy, formę płatności i czy zakup był zakupem na raty),
- o kliencie (regionie i wielkości miejscowości zamieszkania klienta, jego wieku, płci i wykształceniu, liczbie osób w gospodarstwie domowym i kategorii dochodu),
- czy klient przed lub tuż po zakupie kontaktował się z dostawcą.

Część informacji o kliencie nie była konieczna do realizacji transakcji i niektórzy klienci ich nie podawali. Ponieważ wszystkie dane o kliencie zakodowano jako zmienne skategoryzowane (informujące o przynależności do grupy, np. osób w średnim wieku), problem brakujących danych rozwiązano, wprowadzając specjalną klasę: *Brak odpowiedzi*. Dodatkowo wprowadzono zmienną informującą, że klient nie udzielił odpowiedzi na jedno z pytań.

Łącznie w analizie uwzględniliśmy 21 predyktorów, a dysponowaliśmy 2527 przypadkami. Dane zostały wcześniej przygotowane do analizy i sprawdzone.

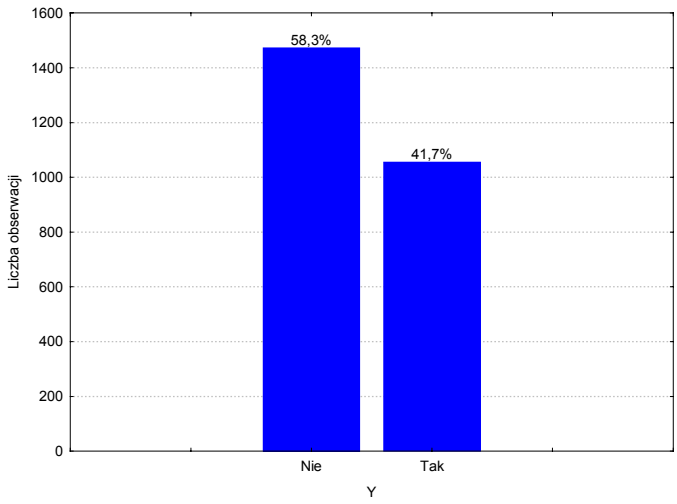
Eksploracja danych

Pierwszym etapem analizy jest eksploracja danych, której celem jest wstępne poznanie ogólnych właściwości zmiennych i ich wzajemnych relacji. Ponadto powinniśmy wykryć obserwacje nietypowe (odstające) i dokonać selekcji zmiennych (np. pominąć zmienne, które praktycznie zawsze przyjmują jedną i tę samą wartość, zawierają dużo braków danych, nie wpływają na badane zjawisko).

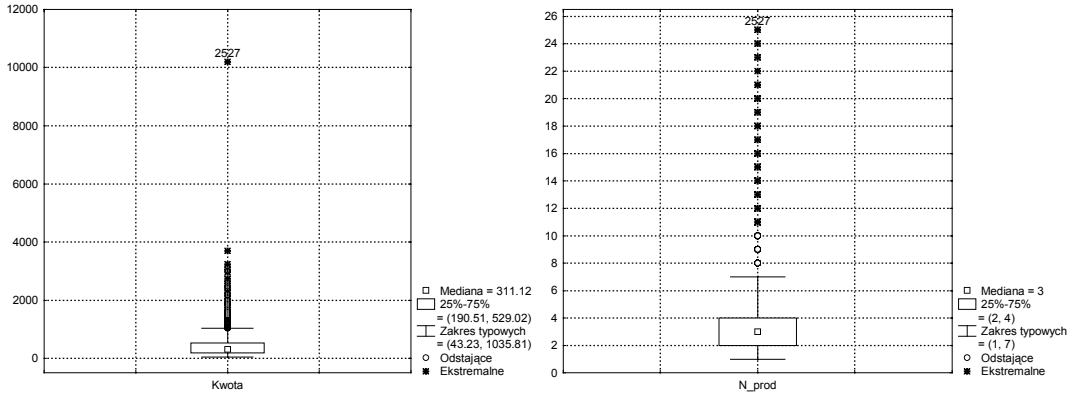
Badanie związków między zmiennymi dotyczy nie tylko relacji predyktorów i modelowanej wielkości, ale również powiązania predyktorów. Wbrew pozorom dosyć często zdarza się, że w danych występują w pełni redundantne zmienne, z których tylko jedną należy uwzględnić w analizie.

Na etapie eksploracji danych sprawdzamy także, jak często występują braki danych i decydujemy, jak z nimi postępować. W wyniku eksploracji danych możemy również uzyskać informacje podpowiadające przekształcenia zmiennych, które ułatwią przeprowadzenie analizy. Więcej informacji o eksploracji danych i przykłady znajdują się w podręczniku [3].

Na początek zobaczmy, jak wygląda rozkład zmiennej zależnej. Na rysunku poniżej widzimy, że około 58% klientów nie dokonało ponownego zakupu (*Y = Nie*), a około 42% dokonało (*Y = Tak*).



Rozkład predyktorów ilościowych (kwota i liczba produktów *N_prod*) widzimy na poniższych wykresach ramka-wąsy. Jak widać, obie zmienne mają rozkład niesymetryczny, ze stosunkowo nielicznymi wartościami wyraźnie przewyższającymi „typowe" wartości.

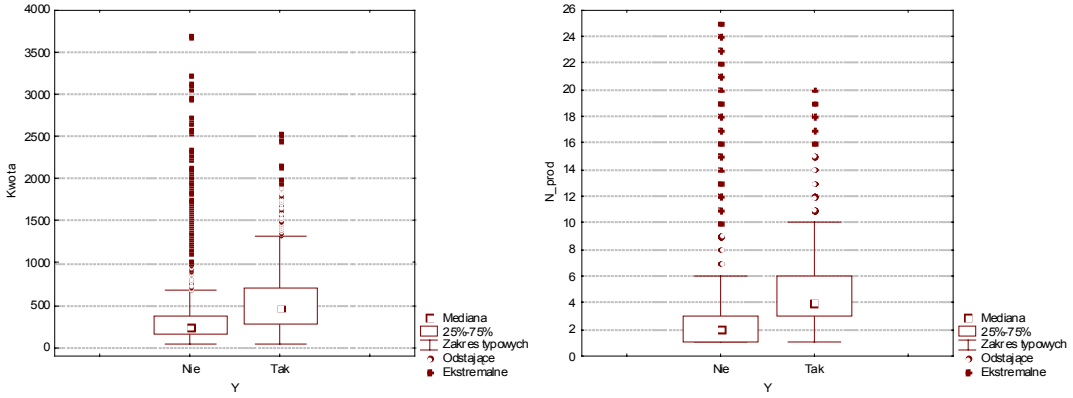


W dalszej analizie będziemy w większości stosować metody odporne na obserwacje odstające, jednak obserwacja numer 2527 wydaje się być zbyt dziwna: kwota dla tego przypadku jest ponad dwukrotnie większa od drugiej co do wielkości. Obserwację numer 2527 wyłączymy z dalszej analizy za pomocą jednej z funkcji graficznej eksploracji danych *STATISTICA*. W przypadku stosowania metod wrażliwych na obserwacje odstające, warto by rozważyć zlogarytmowanie kwoty i liczby produktów.

Predyktory jakościowe możemy przeanalizować, korzystając z histogramów lub tabel liczności, jednak w przypadku naszych danych nie pokazują one nic ciekawego. W szczególności liczności wszystkich klas są rozsądne.

Kolejnym krokiem jest zbadanie zależności między potencjalnymi predyktorami, a zmienną zależną.

W przypadku predyktorów ilościowych i zmiennej zależnej jakościowej dosyć często stosuje się podejście „odwrotne", tzn. sprawdzamy, czy rozkład predyktorów jest różny dla różnych klas modelowanej cechy. Pierwszym sprawdzianem będą wykresy ramka-wąsy pokazujące wzajemne relacje statystyk pozycyjnych (nie stosujemy układu z średnimi i błędami standardowymi, albowiem zmienne *Kwota* i *N_prod* mają niesymetryczny rozkład zdecydowanie odbiegający od normalnego).



Rzuca się w oczy, że klienci lojalni w pierwszym zakupie płacili więcej i kupowali więcej produktów. Jednak jeśli przyjrzymy się bliżej powyższym wykresom, to zauważymy, że dla bardzo dużych kwot i liczb produktów tendencja ta zanika albo wręcz ulega odwróceniu – skrajnie duże wartości częściej pojawiają się dla grupy *Nie*.

Wyniki badania „na oko" możemy potwierdzić, wykonując test U Manna-Whitneya. W przypadku obu zmiennych pozwala on odrzucić hipotezę o tym, że rozkład w grupach jest taki sam na poziomie istotności 1%.

W przypadku predyktorów jakościowych i jakościowej zmiennej zależnej do zbadania zależności możemy użyć tabel krzyżowych i wykresów skategoryzowanych. Dla przykładu zobaczmy, jak forma płatności i to, czy zakup był zakupem na raty, wpływają na lojalność.

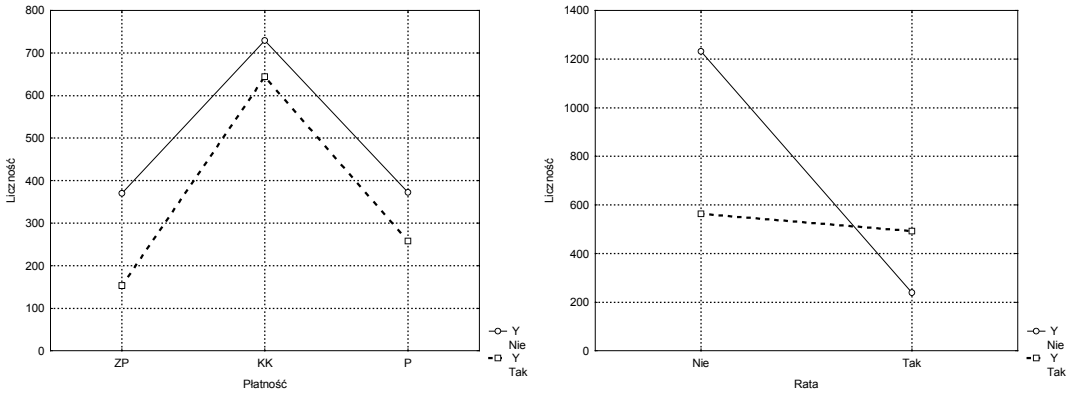
Patrząc na tabele liczności poniżej, wydaje się, że wpływ zmiennej *Płatność* jest niezbyt silny, w szczególności grupy *KK* i *P* niczym się nie różnią.

Y	Płatność ZP	Płatność KK	Płatność P	Wiersz razem
Nie	70.75%	53.10%	59.05%	58.23%
Tak	29.25%	46.90%	40.95%	41.77%
Razem	20.70%	54.35%	24.94%	

Natomiast bardzo silny jest wpływ zmiennej *Rata*: widać, że zakup na raty sprzyja późniejszym, innym zakupom.

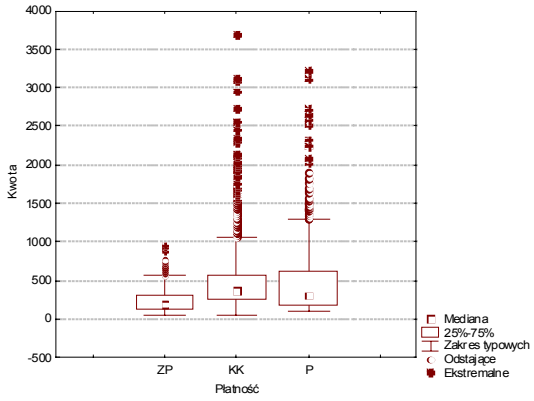
Y	Rata Nie	Rata Tak	Wiersz razem
Nie	68.64%	32.69%	58.23%
Tak	31.36%	67.31%	41.77%
Razem	71.06%	28.94%	

Zależności w tabelkach ilustrują poniższe wykresy interakcji:



W podobny sposób powinniśmy przeanalizować związki między innymi predyktorami a faktem kontynuowania zakupów (tj. zmienną *Y*), zwłaszcza jeśli chcemy użyć metod, które „nie lubią" dużej liczby zmiennych nic nie wnoszących do modelu.

Na etapie eksploracji badamy również związki między zmiennymi niezależnymi. Jak pamiętamy, forma płatności wykazywała słaby związek ze zmienną zależną *Y*. Wydaje się, że forma płatności może być związana z płaconą kwotą. Na poniższym rysunku widzimy rozkład zmiennej Kwota dla różnych form płatności. Wyraźnie widać, że klasa ZP to zdecydowanie niższe kwoty. Jest to najprawdopodobniej przyczyną tego, że mamy inną częstość klas *Tak* i *Nie* zmiennej *Y* dla klasy *ZP* w stosunku do pozostałych wartości zmiennej *Płatność*. Bazując na tym spostrzeżeniu, możemy przypuszczać, że w modelu uwzględniającym wiele predyktorów zmienna *Płatność* się nie pojawi, bo znajdzie się w nim zmienna *Kwota*, która bardzo silnie wpływa na zmienną zależną, a cały związek pomiędzy zmienną *Płatność* a zmienną Y jest skutkiem powiązania *Płatności* i *Kwoty*.



Tworzenie modelu

Po wstępnej analizie danych przejdziemy do tworzenia modelu. Do budowy modelu wykorzystamy kilka różnorodnych technik i zobaczymy, która z nich daje najtrafniejsze przewidywania.

Model będziemy budować i porównywać w przestrzeni roboczej systemu *STATISTICA Data Miner*. Zbiór danych podzielimy losowo na próbę ucząca (80% przypadków) i testową (20%).

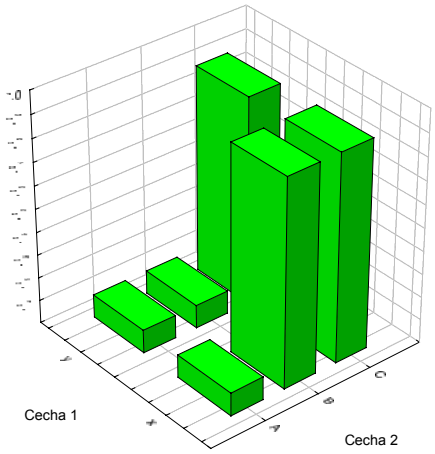
Najpierw zbudujemy model z wykorzystaniem *Uogólnionej analizy dyskryminacyjnej (GDA)*. Procedura ta tworzy podziały liniowe: można powiedzieć, że przestrzeń predyktorów dzielona jest płaszczyznami na segmenty odpowiadające klasom. Aby uzyskać poprawny model, włączymy automatyczny dobór zmiennych metodą krokową postępującą. Polega ona na dodawaniu do modelu kolejnych predyktorów, dla których ryzyko, że nie wpływają one istotnie na zmienną zależną, jest najmniejsze.

Podczas analizy jako pierwszy predyktor do modelu została wstawiona zmienna *Rata*, a jako drugi *N_prod* (liczba produktów). Ostatecznie w modelu znalazły się zmienne: *Rata*, *N_prod*, *Brak_odp* (informująca, że klient nie udzielił odpowiedzi na jedno z pytań), *Gr_wiek* (grupa wieku), *Kontakt* (czy klient kontaktował się z dostawcą), grupa zakupionych produktów, sposób dostawy, miejscowość, płatność, płeć.

Model GDA dla próby testowej myli się w 27% przypadków. Jeśli faktyczną wartością zmiennej *Y* było *Tak*, to stopa błędów wynosi 40%, a w przeciwnym przypadku 17%.

Jak pamiętamy z wstępnej analizy danych, dla dużych wartości kwoty mamy odwrotną tendencję niż dla przeciętnych wartości: tzn. częściej występują wartości *Nie* (nielojalni klienci). Takiego układu powiązań analiza dyskryminacyjna nie jest w stanie wychwycić. Zastosujemy drzewa klasyfikacyjne i regresyjne, które są w stanie znaleźć podziały nieliniowe.

W drzewach klasyfikacyjnych i regresyjnych poszukujemy takich części (segmentów) przestrzeni cech parametrów, w których zmienna zależna przyjmuje tylko pewną określoną wartość. Dla ilustracji rozważmy poniższy wykres. Widzimy na nim częstość występowania klasy *Tak* w zależności od wartości, jakie przyjmują dwie cechy.



Patrząc na wykres, możemy sformułować następującą regułę:

Obiekt należy do klasy *Tak*:

- jeśli *Cecha 2* należy do klasy *C*,
- albo jeśli *Cecha 1* należy do klasy X i *Cecha 2* należy do klasy *B*.

Drzewa klasyfikacyjne poszukują optymalnego podziału na segmenty, wykonując następujące działania:

1. Sprawdzenie dla aktualnie badanego zbioru, czy jest on jednorodny lub czy spełniona jest inna reguła stopu.
2. Zbadanie wszystkich możliwych podziałów na rozłączne części.
3. Wykonanie najlepszego podziału zbioru na rozłączne części.
4. Powtórzenie dla wszystkich zbiorów uzyskanych poprzez wykonanie powyższych czynności.

W problemach klasyfikacyjnych jako wskaźnik jednorodności często stosuje się miarę Giniego, która jest równa prawdopodobieństwu tego, że jeżeli trafimy na obiekt z pewnej klasy, to drugi wylosowany obiekt będzie należał do innej klasy. W przypadku dwóch klas miara ta jest równa *2*f*(f-1)*, gdzie *f* oznacza proporcję występowania jednej z klas. (Miara ta wprowadzona została w biologii i jest tam nazywana wskaźnikiem Simpsona).

Jako reguły stopu stosuje się m.in.: minimalną liczność węzła podlegającego podziałom, minimalną liczność węzła powstającego w wyniku podziałów i maksymalną liczbę poziomów drzewa.

Po zakończeniu podziałów wykonuje się jeszcze operację doboru właściwej wielkości drzewa, np. przycinanie (ang. *pruning*). Przycinanie polega na usuwaniu gałęzi drzewa, co wykonujemy automatycznie lub ręcznie, w oparciu o posiadaną wiedzę o celach analizy, jakości pomiaru poszczególnych cech, ograniczeniach stosowania modelu itp. (jest to wiedza, której nie ma w danych i siłą rzeczy analiza danych nie może jej wydobyć).

Do określenią właściwej wielkości drzewa używa się V-krotnego sprawdzianu krzyżowego. Polega on na podzieleniu danych na V-części, zbudowaniu drzewa dla zbioru uczącego złożonego z V-1 części i ocenieniu go dla pozostałej części. Całą procedurę powtarzamy V razy (każda z części raz pełni rolę zbioru testowego) i na podstawie zależności błędu od złożoności drzewa określamy optymalną wielkość drzewa (więcej informacji można znaleźć w podręczniku [9]).

Uzyskane reguły standardowo prezentuje się w postaci drzewa, dzięki czemu są one stosunkowo przejrzyste, nawet gdy drzewo jest spore. Łatwość zrozumienia wyników i możliwość odkrycia łatwych w interpretacji reguł jest jedną z zalet drzew klasyfikacyjnych.

Inne zalety drzew klasyfikacyjnych i regresyjnych to:

- Prostota algorytmu, ułatwiająca stosowanie go ze zrozumieniem nawet osobom bez dużego doświadczenia w analizie danych,
- Szybkość działania,
- Odporność na nietypowe wartości predyktorów,
- Odporność na nawet dużą liczbę predyktorów faktycznie niewpływających na badaną zmienną,
- Możliwość wychwycenia zależności nieliniowych i interakcji.

Wszystkie te zalety powodują, że dobrze jest zacząć analizę właśnie od tej metody. Z drugiej strony drzewa nie są w stanie opisać tak złożonych zależności jak inne, bardziej skomplikowane matematycznie procedury, np. sieci neuronowe, metoda wektorów nośnych (*Support Vector Machines*) czy drzewa ze wzmacnianiem (*boosted trees*).

W naszym przypadku zastosujemy drzewa klasyfikacyjne C&RT z następującymi parametrami:

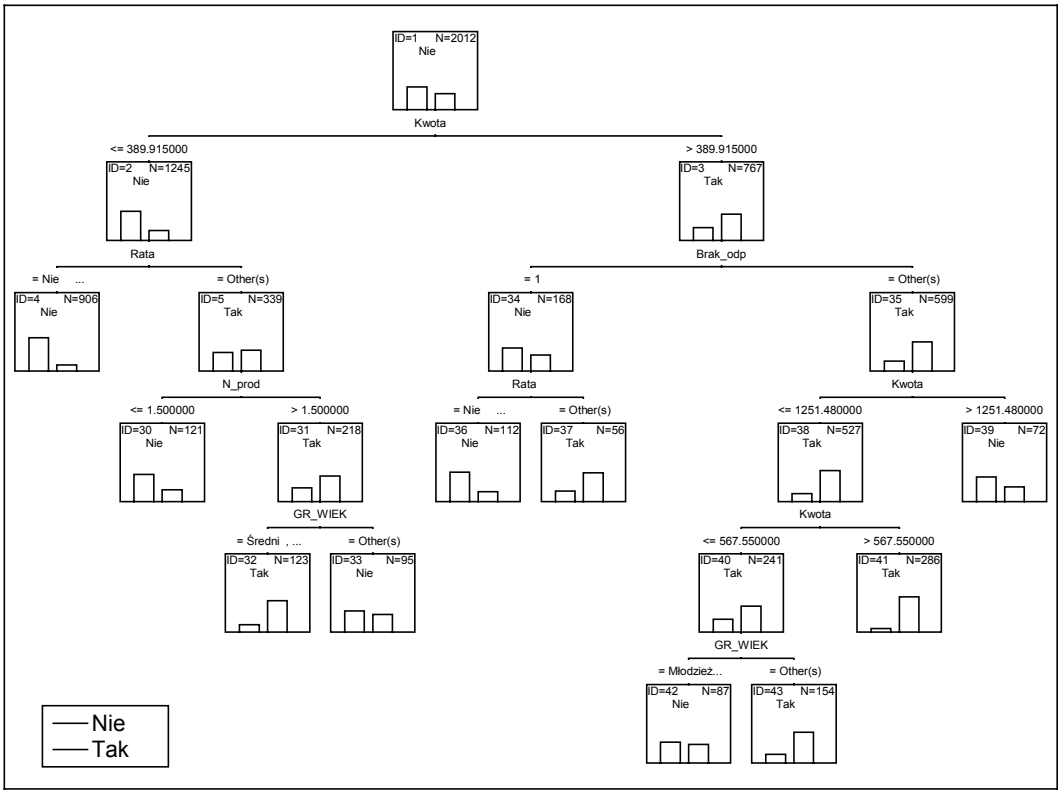
- Jako wskaźnik jednorodności użyjemy miary Giniego,
- Minimalną liczność dzielonego węzła ustawimy równą 125,

- Do doboru optymalnego drzewa wykorzystamy V-krotny sprawdzian krzyżowy z pięcioma złożeniami (V=5).

Na rysunku poniżej widzimy drzewo klasyfikacyjne uzyskane w wyniku analizy. Jest ono dosyć proste: składa się łącznie z 19 węzłów, z czego 10 to liście (węzły końcowe).

Zauważmy, że drzewo przewiduje, że dla dużych wartości kwoty zwiększa się prawdopodobieństwo, że klient nie będzie więcej u nas kupował ($Y = Nie$) – zwróćmy uwagę na podział węzła nr 35.

Drzewo klasyfikacyjne dla próby testowej myli się w 22% przypadkach. Jeśli faktyczną wartością zmiennej Y było *Tak*, to stopa błędów wynosi 38%, a w przeciwnym przypadku 9%. Wszystkie te wskaźniki są wyraźnie lepsze niż dla modelu GDA.



Następny model zbudujemy za pomocą drzew ze wzmacnianiem (*boosted trees*).

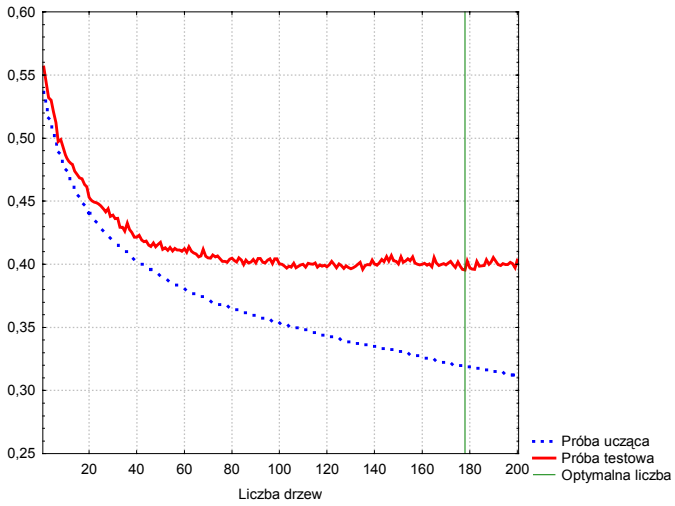
Wzmacnianie polega na budowaniu modelu złożonego z ciągu prostych modeli, przy czym kolejne modele składowe uczone są przy coraz większej wadze „trudnych" przypadków, dla których dotychczasowe modele myliły się.

Mówiąc precyzyjniej, uczenie zaczynamy od opracowania prostego modelu (wagi wszystkich obserwacji są równe). Następnie obliczamy błąd dla modelu jako całości i wagi dla

poszczególnych obserwacji. Wagi możemy wyznaczyć, np. powiększając wagi początkowe z wykorzystaniem wartości błędu całego modelu. Wagi dla poprawnie sklasyfikowanych przypadków nie ulegają zmianie. W następnym kroku dopasowujemy prosty model dla danych ze skorygowanymi wagami. Uzyskane w ten sposób modele cząstkowe łączymy w jeden, wyliczając ich ważoną sumę. Więcej informacji można znaleźć w podręczniku [9].

Model drzew ze wzmocnieniem zbudujemy przy domyślnych ustawieniach *STATISTICA Data Miner*, z jedną modyfikacją: zwiększymy maksymalną liczbę węzłów w składowym drzewie z 3 do 7, aby móc uwzględnić interakcje zmiennych.

Uzyskany model składa się z 178 drzew, a jego uzyskanie zajmuje zdecydowanie więcej czasu niż modelu GDA i drzew klasyfikacyjnych. Jednak trafność modelu jest najlepsza: dla próby testowej frakcja błędów wynosi 19% przypadków. Jeśli faktyczną wartością zmiennej *Y* było *Tak*, to stopa błędów wynosi 21%, a w przeciwnym przypadku 18%.



Pomimo trafniejszych przewidywań drzew ze wzmacnianiem, w praktyce lepszy może okazać się model drzew klasyfikacyjnych ze względu na prostotę i mniej czasochłonne obliczenia.

Literatura

[1]. Analiza satysfakcji i lojalności klientów, 2003, StatSoft Polska.

[2]. Berry M.J.A., Linoff G., Mastering *data mining*, John Willey & Sons, 2000.

[3]. Berson, A., Smith, S., Thearing, K., Building *Data mining* Applications for CRM, McGraw-Hill, 1999.

[4]. Phelan S., "Customer Information as a Strategic Asset", DM Review Online (www.DMReview.com), kwiecień 2002.

[5]. Nowoczesne narzędzia gromadzenia, udostępniania i analizy danych: *STATISTICA Data Miner* i Sybase IQ, StatSoft Polska, 2004 (publikacja dostępna w Internecie na stronie www.statsoft.pl).

[6]. Coppock D. S., "Data Mining and Modeling: Model to Reduce Attrition", DM Review Online (www.DMReview.com), kwiecień 2002.

[7]. Giudici P., "Applied Data Mining. Statistical Methods for Business and Industry", John Wiley & Sons Ltd, 2003.

[8]. Data mining – metody i przykłady, StatSoft Polska, 2002.

[9]. Hastie T, Tibshirani R., Friedman J., The Elements of Statistical Learning, Springer-Verlag, 2002.