MODEL DATA MINING PRZEWIDUJĄCY ODPOWIEDŹ KLIENTÓW NA OFERTĘ

Tomasz Demski, StatSoft Polska Sp. z o.o.

Jednym z najpopularniejszych zadań *data mining* jest przewidywanie, którzy klienci najchętniej odpowiedzą na naszą ofertę. Zajmiemy się właśnie rozwiązaniem takiego zadnia, przy czym duży nacisk położymy na przygotowanie danych i ich wstępną analizę.

Opis problemu i danych

Naszym zadaniem będzie przewidzenie, którzy klienci odpowiedzą pozytywnie na ofertę zakupu karty kredytowej. Dysponujemy danymi o decyzjach klientów z przeszłości. Na podstawie tych danych zbudujemy model *data mining*, który będziemy mogli zastosować dla nowych klientów. Wykorzystamy nieco zmieniony zbiór CREDIT dostępny z podręcznikiem [1].

Dane zawierają informacje o potencjalnych klientach zakupione w biurze informacji kredytowej. Bank kontaktuje się z tymi osobami za pośrednictwem poczty elektronicznej lub telefonu i proponuje im zakup karty kredytowej. W zbiorze danych mamy informacje, którzy klienci pozytywnie odpowiedzieli na ofertę. Naszym zdaniem jest stworzenie modelu, który na podstawie cech klientów będzie w stanie przewidzieć ich odpowiedź na ofertę. Interesuje nas nie tylko samo przewidywanie decyzji klientów, ale również wiedza dotycząca czynników najmocniej wpływających na odpowiedź na ofertę oraz wzajemnych związków między zmiennymi; innymi słowy chcemy wychwycić wzorce zachowań klientów.

Nasze zadanie należy do dziedziny predykcyjnego *data mining* (ukierunkowanego lub uczenia z nauczycielem) i jest to problem klasyfikacyjny: przewidujemy, do jakiej klasy należy dana osoba: tych, co kupią kartę, czy tych, co jej nie kupią. Ponieważ chcemy zrozumieć dane i interpretować model, duży nacisk położymy na przygotowanie danych oraz wstępną analizę i zastosujemy metody, które są łatwe w interpretacji.

Dysponujemy danymi o 13 996 osobach, którym zaproponowano kartę kredytową. W zbiorze znajduje się 39 zmiennych (cech potencjalnych klientów), na podstawie których będziemy chcieli przewidywać odpowiedź na ofertę. Zmienne te są predyktorami w naszej

analizie. Zmienną zależną jest zmienna *Buyer* przyjmująca dwie wartości 'T' (klient zakupił kartę) i 'N' (negatywna odpowiedź na ofertę).

Dane wymagają wstępnej obróbki i przygotowania do właściwej analizy. Co prawda możemy liczyć na to, że zawierają prawdziwe informacje, ale istnieje niebezpieczeństwo, że występują w nich problemy uniemożliwiające lub bardzo utrudniające wykonanie skutecznej i użytecznej analizy.

Problemy z danymi, które należy rozwiązać przed właściwą analizą, to:

- braki danych (puste zmienne),
- zmienne niewykazujące zmienności i niewpływające na zmienną zależną,
- anachroniczne informacje (dane, które są wpisywane po zarejestrowaniu wartości zmiennej zależnej i są z nią ściśle związane),
- obserwacje nietypowe.

Wiele metod (lepiej lub gorzej) radzi sobie z powyższymi trudnościami, przykładowo drzewa klasyfikacyjne nie są wrażliwe na problem predyktorów niezwiązanych ze zmienną zależną. Pewne podstawowe operacje na danych są również wykonywane przez praktycznie wszystkie procedury *STATISTICA* – dotyczy to m.in. obsługi braków danych. W programie *STATISTICA Data Miner* możemy skorzystać np. z *Automatycznego projektanta sieci* (ang. *Intelligent Problem Solver*). Jednak w naszym przypadku chcemy dowiedzieć się jak najwięcej o związkach między zmiennymi i przygotowanie danych przeprowadzimy ręcznie.

Wstępna analiza danych i przekształcenia

Statystyki opisowe i rozkłady

Zaczniemy od statystyk opisowych zmiennych występujących w pliku danych. Poniżej widzimy tabelę ze statystykami opisowymi dla predyktorów z naszego pliku danych. (Uwaga: statystyki wyznaczone są dla wszystkich zmiennych bez względu na ich typ i część z nich nie ma jasnej interpretacji, chodzi nam jednak o to, aby mieć w jednej tabeli zebrane podstawowe informacje o wszystkich zmiennych).

	N ważnych	Średnia	Min.	Maks.	Odch.Std.	Skośność
BEACON	13996	749,66	670,00	804,0	24,87	-0,34
DAS	13996	154,70	-202,00	524,0	120,06	0,16
CRITERIA	13996	1,00	1,00	1,0	0,00	
ROPEN	13996	3,54	0,00	17,0	2,17	1,08
RBALNO	13996	2,25	0,00	16,0	1,61	1,42
LST_R_OPEN	13676	27,01	1,00	99,0	28,14	1,43
RBAL	13676	3295,70	0,00	78928,0	4586,61	3,10
RLIMIT	13676	14853,11	0,00	999999,0	13040,21	32,48

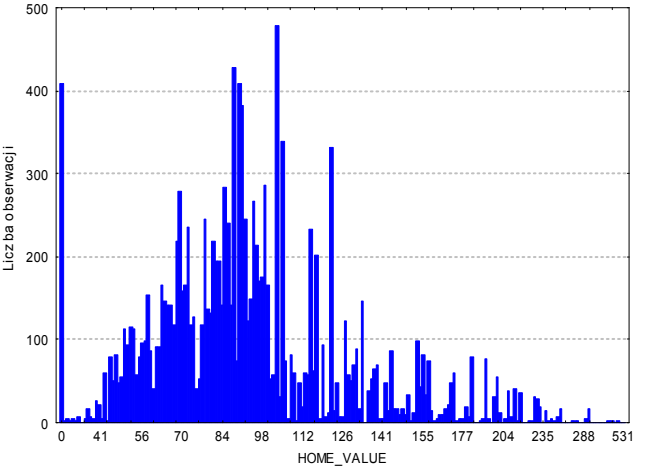
	N ważnych	Średnia	Min.	Maks.	Odch.Std.	Skośność
TOPEN	13996	3,95	0,00	17,0	2,20	1,04
TBALNO	13905	12899,98	0,00	999999,0	24003,05	10,00
MOF	13996	177,02	0,00	976,0	85,75	1,26
RBAL_LIMIT	13996	0,60	0,00	9,0	0,96	2,26
EQLIMIT	128	27021,09	20000,00	219000,0	18162,20	9,54
EQBAL	128	11354,84	0,00	67950,0	9087,98	1,84
EQHIGHBAL	277	47739,38	20000,00	237000,0	34501,09	2,05
EQCURBAL	277	39496,95	0,00	220000,0	34247,14	1,93
BCLIMIT	11481	9660,33	0,00	46435,0	6286,96	1,24
BCBAL	11481	2080,74	0,00	24251,0	2837,80	2,58
IHIGHBAL	7292	18347,60	0,00	116545,0	12191,20	1,62
ICURBAL	7290	10852,12	0,00	171000,0	9339,57	2,44
UNSECLIMIT	6117	6797,15	0,00	39395,0	5158,84	1,13
UNSECBAL	6117	2082,05	0,00	23917,0	3338,58	2,45
MTHIGHBAL	7721	83935,58	0,00	614000,0	44561,68	2,06
MTCURBAL	7721	77314,59	0,00	613000,0	45207,11	1,78
BCOPEN	0					
YEARS_RES	13411	5,66	0,00	15,0	4,55	0,88
CHILDREN	13996	0,30	0,00	1,0	0,46	0,90
EST_INC	13996	66349,46	42999,50	87499,5	17164,01	0,02
OWN_HOME	13996	0,00	0,00	0,0	0,00	
HOME_VALUE	13996	94,49	0,00	531,0	41,18	1,41
HOME_INC	13996	41,76	0,00	150,0	15,89	0,56
HOME_ED	13996	127,97	0,00	160,0	25,85	-3,28
PRCNT_WHIT	13996	90,15	0,00	99,0	17,42	-4,18
PRCNT_PROF	13996	32,88	0,00	86,0	12,92	-0,30
DOB_MONTH	1139	1,13	0,00	12,0	2,88	2,54
DOB_YEAR	9448	51,91	30,00	70,0	9,39	-0,36
AGE_INFERR	13996	43,79	35,00	57,0	7,26	0,91
SEX	13996	101,14	101,00	103,0	0,39	2,70
MARRIED	13996	101,46	101,00	102,0	0,50	0,16

Najpierw zobaczmy, czy któraś ze zmiennych nie jest pusta. W pierwszej kolumnie powyższej tabeli mamy liczbę poprawnych wartości dla każdej ze zmiennych. Zmienna *Bcopen* jest całkowicie pusta – nie mamy dla niej żadnych wartości; na pewno musimy ją wykluczyć z dalszych analiz. Cztery zmienne od *Eqlimit* do *Eqcurbal* zawierają bardzo mało poprawnych wartości – zaledwie klika procent komórek jest w nich wypełnione. Te

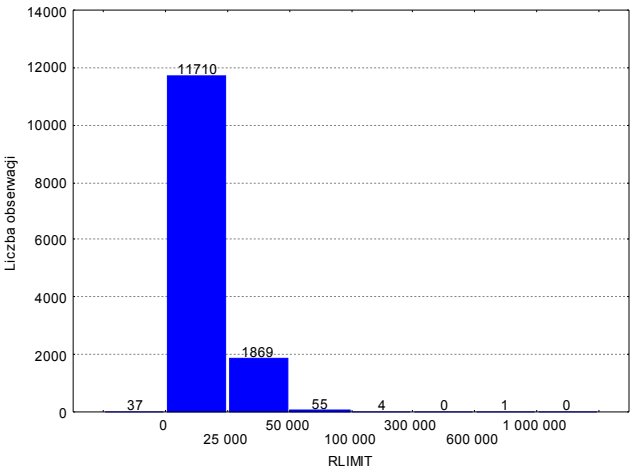
zmienne również powinniśmy pominąć w budowie modelu. Mało poprawnych informacji mamy również dla zmiennej *Dob_month*: jest ona wypełniona w poniżej 10% przypadków – później przyjrzymy się tej zmiennej bliżej, teraz możemy stwierdzić, że uwzględnienie jej w modelu bez wcześniejszego przekształcenia może być trudne. Pozostałe zmienne mają poprawne wartości w kilku tysiącach przypadków, jednak 6 z nich (*Ihighbal*, *Icurbal*, *Unseclimit*, *Unsecbal*, *Mthighbal*, *Mtcurbal*) ma udział braków danych na poziomie od 40% do 60% i w dalszej analizie powinniśmy na nie zwrócić uwagę. W rozdziale „Braki danych" (str. 106) zbadamy dokładniej wpływ wystąpienia braku danych na zmienną zależną.

Zmienne, które przyjmują tylko jedną wartość dla wszystkich obiektów, na pewno nie będą użyteczne w modelu (poszukujemy cech różnicujących obiekty, pozwalających rozpoznać te, które trafiają do klasy 'T' i 'N" zmiennej zależnej). Właśnie takie są zmienne *Criteria* oraz *Own_home*: przyjmują one tę samą wartość (odpowiednio 1,0 i 0,0) dla wszystkich przypadków i mają odchylenie standardowe równe zeru.

Powinniśmy sprawdzić, czy zakres wartości cech jest sensowny. Nieodpowiedni zakres ma zmienna *Home_value* (wartość domu) – jej minimum wynosi 0. Przyjrzyjmy się rozkładowi tej zmiennej (wykres poniżej). Jak widać, 0 występuje bardzo często, przy czym bliskie mu liczby są bardzo rzadkie. Raczej na pewno 0 nie oznacza tu bezwartościowego domu, lecz sygnalizuje brak danych (np. osoba nie posiada domu, a wynajmuje mieszkanie). Przy tworzeniu modelu można zastąpić 0 zmiennej *Home_value* brakami danych lub ją skategoryzować, przypisując zeru osobną kategorię – tak właśnie postąpimy.



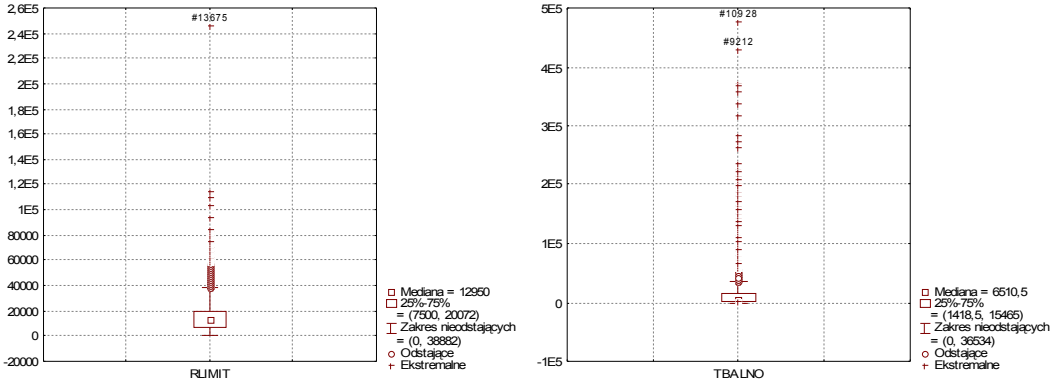
Zwróćmy uwagę na skośność. Duże wartości tej statystyki oznaczają, iż rozkład zmiennej bardzo odbiega od rozkładu symetrycznego i sugerują występowanie wartości nietypowych. Największą wartość skośności ma zmienna *Rlimit*: jej skośność przekracza 32. Jej średnia wynosi ok. 14 583, odchylenie standardowe ok. 13 040, a maksimum aż 999 999 – czyli jest odległe od średniej aż o 75 odchyleń standardowych – takie statystyki są bardzo podejrzane.



Po sporządzeniu tabeli liczności lub histogramu (widocznego na rysunku powyżej) dla zmiennej *Rlimit* jasne jest, że wartość 999 999 należy pominąć przy budowie modelu. Występuje ona tylko raz, a druga co do wielkości wartość jest poniżej 300 000! Biorąc pod uwagę, że 999 999 jest „nieprzypadkową" liczbą, mamy prawo podejrzewać, iż ktoś użył jej jako kodu braku danych, tym bardziej, że podobną sytuację mamy dla zmiennej *Tbalno*. W przypadku obu tych zmiennych wartość 999 999 zamienimy na brak danych.

Po zastąpieniu 999 999 kodem braku danych skośność *Rlimit* spada do 2,15, a dla *Tbalno* do 6,12 – jest to druga co do wielkości wartość tej statystyki.

Bardzo dużą wartość skośności ma również zmienna *Eqlimit*, ale ponieważ jest ona praktycznie pusta i pominiemy ją w modelu, nie musimy się nią przejmować.



Występowanie obserwacji odstających możemy sprawdzić, stosując wykresy ramka-wąsy. Na rysunkach powyżej widzimy wykresy ramka-wąsy dla zmiennych *Rlimit* i *Tbalno* (już po usunięciu wartości 999 999). Jak widać, rozkłady tych zmiennych są bardzo niesymetryczne, a ponadto występują obserwacje nietypowe. Tym razem decyzja o usunięciu

dziwnych wartości nie jest tak oczywista, jak w przypadku 999 999. Być może powinniśmy zastąpić brakiem danych wartość przypadku nr 3675 dla zmiennej *Rlimit* (jest ona ponad dwukrotnie większa niż druga co do wielkości wartość tej zmiennej). Na pewno przy doborze metody musimy uwzględnić występowanie nietypowych wartości i skośność rozkładów albo w znacznym stopniu przekształcić dane (por. [1]).

W przypadku cech jakościowych (informujących nas o przynależności do klasy) powinniśmy utworzyć tabele liczności. Nasz zbiór danych zawiera 3 zmienne jakościowe, które możemy użyć w modelu *Children*, *Married* i *Sex*. Poniżej widzimy tabele liczności dla tych zmiennych; szczęśliwie tym razem nie wykazują one nic niepokojącego. Warto jednak zwrócić uwagę na płeć (zmienna *Sex*): przyjmuje ona trzy wartości - oprócz męskiej (*M*) i żeńskiej (*F*) występuje również kod braku informacji (*U*).

	Tabela liczności: CHILDREN (Credit.sta)			
Kategoria	Liczność	Skumulow. Liczn.	Procent	Skumulow. Procent
N	9863	9863	70,47013	70,4701
T	4133	13996	29,52987	100,0000
Braki	0	13996	0,00000	100,0000

	Tabela liczności: SEX: Płeć (Credit.sta)			
Kategoria	Liczność	Skumulow. Liczn.	Procent	Skumulow. Procent
M	12183	12183	87,04630	87,0463
F	1631	13814	11,65333	98,6996
U	182	13996	1,30037	100,0000
Braki	0	13996	0,00000	100,0000

	Tabela liczności: MARRIED (Credit.sta)			
Kategoria	Liczność	Skumulow. Liczn.	Procent	Skumulow. Procent
M	7557	7557	53,99400	53,9940
U	6439	13996	46,00600	100,0000
Braki	0	13996	0,00000	100,0000

Na koniec sprawdźmy rozkład zmiennej zależnej. Widzimy go w tabeli poniżej. Warto zwrócić uwagę, że negatywna odpowiedź na ofertę występuje znacznie częściej niż pozytywna. Nierówne częstości klas utrudnią uzyskanie użytecznego modelu: zazwyczaj modele będą miały tendencję do faworyzowania częstszej klasy. Zauważmy, że „bezmyślny" model zawsze przewidujący klasę *N* będzie miał stopę błędów około 21%, co wcale nie jest złą wartością (oczywiście jednocześnie model będzie całkowicie bezużyteczny).

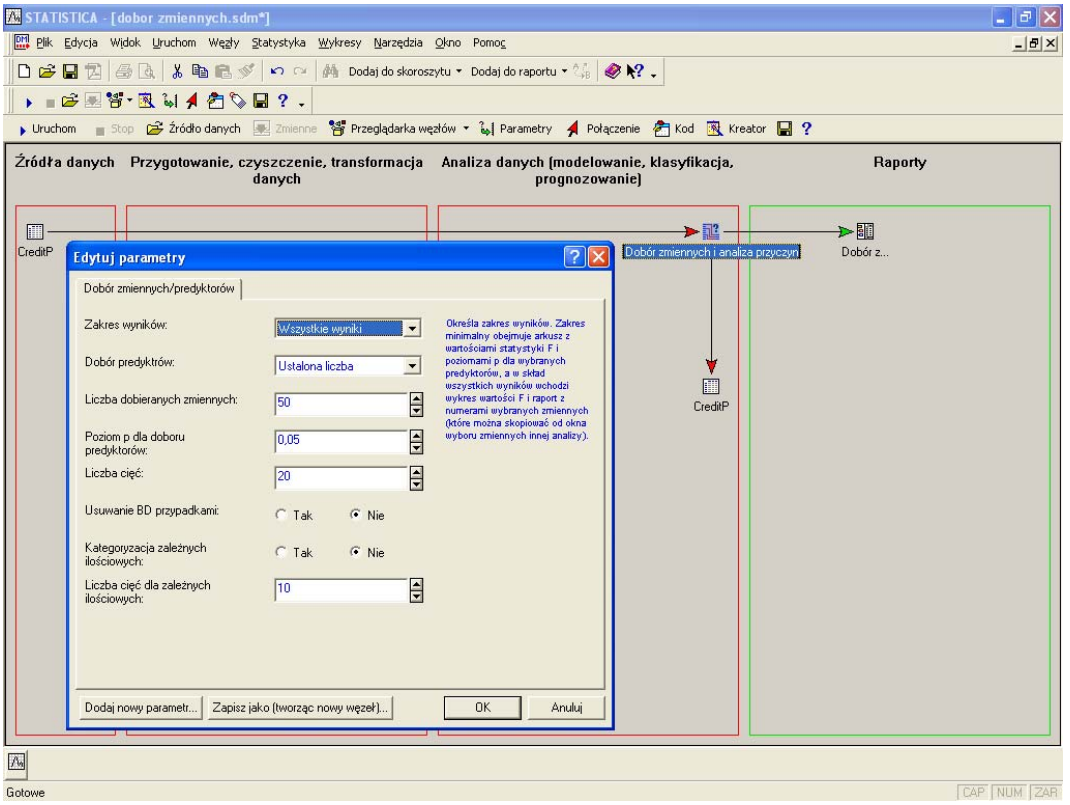
	Tabela liczności: BUYER: Zakup (Credit.sta)			
Kategoria	Liczność	Skumulow. Liczn.	Procent	Skumulow. Procent
N	10999	10999	78,58674	78,5867
T	2997	13996	21,41326	100,0000
Braki	0	13996	0,00000	100,0000

Problem niezrównoważonych częstości możemy rozwiązać, modyfikując dane (stosując *oversampling* lub inną technikę, więcej informacji można znaleźć w pracach [2] i [3]) lub korzystając z metod pozwalających ustawić różne koszty błędnych klasyfikacji dla różnych kategorii zmiennej zależnej.

Badanie wpływu poszczególnych zmiennych na odpowiedź na ofertę

Kolejnym etapem wstępnej analizy danych jest zbadanie wpływu potencjalnych predyktorów na zmienną zależną.

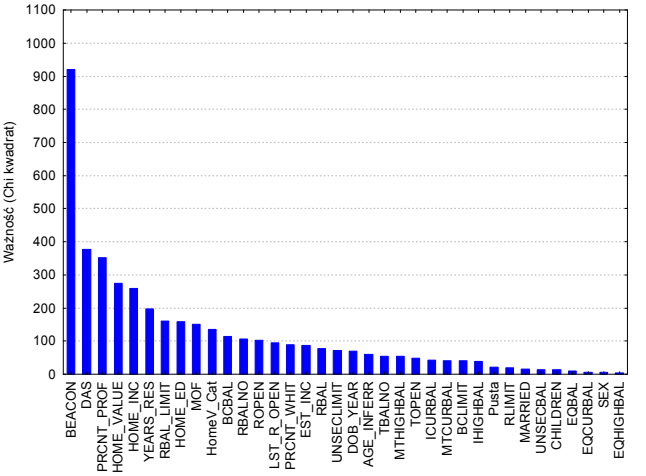
Najpierw badamy, czy poszczególne zmienne wpływają na modelowaną cechę. Do tego celu często wykorzystuje się wykresy i techniki statystyczne: analizę korelacji liniowej (gdy zmienna zależna i predyktory są ilościowe), korelacje nieparametryczne i tabele krzyżowe (gdy zmienna zależna i predyktory są jakościowe), analizę przekrojową i testy t lub ANOVA (gdy zmienna zależna jest jakościowa, a predyktory ilościowe, albo gdy zmienna zależna jest ilościowa, a predyktory jakościowe). Można również zastosować proste techniki modelowania dające wyniki łatwe w interpretacji, takie jak regresja liniowa, analiza dyskryminacyjna lub drzewa (np. algorytm C&RT lub CHAID, który zastosowano np. w [2]).



My zastosujemy wbudowane w *STATISTICA Data Miner* narzędzie do wstępnego doboru zmiennych i eliminacji „pustych" właściwości. Moduł o nazwie *Dobór zmiennych i analiza przyczyn* (ang. *Feature selection and variable screening*) jest w stanie wykrywać zależności nieliniowe i automatycznie stosuje odpowiednią metodę dla zmiennych jakościowych i ilościowych. Ponadto wykrywa puste zmienne i zmienne wypełnione jedną wartością.

Z modułu *Dobór zmiennych i analiza przyczyn* korzystamy w trybie interakcyjnym lub za pośrednictwem węzła w przestrzeni roboczej *data mining*. W tym drugim przypadku automatycznie tworzone jest źródło danych z dobranymi predyktorami. Na rysunku powyżej widzimy przestrzeń roboczą *data mining* z otwartym oknem parametrów dla węzła *Dobór zmiennych i analiza przyczyn*.

Chcemy się dowiedzieć, jak silny jest wpływ zmiennych na decyzję potencjalnego klienta, dlatego *Liczbę dobieranych zmiennych* ustawiamy jako równą 50. Na rysunku poniżej widzimy miarę wpływu poszczególnych zmiennych na decyzję klienta.



Zmienna zależna jest najsilniej związana z dwoma wskaźnikami zdolności kredytowej *Beacon* i *Das*, przy czym wskaźnik *Beacon* zdecydowanie góruje nad innymi cechami. Na pewno ważne są również dane demograficzne kandydata *Prcnt_prof* (stopień profesjonalizmu określający stanowisko pracy), *Years_res* (czas zamieszkiwania w aktualnej lokalizacji), *Home_ed* (poziom wykształcenia w rodzinie) oraz zmienne majątkowe *Home_inc* (przychód gospodarstwa domowego) i *Home_value* (wartość domu).

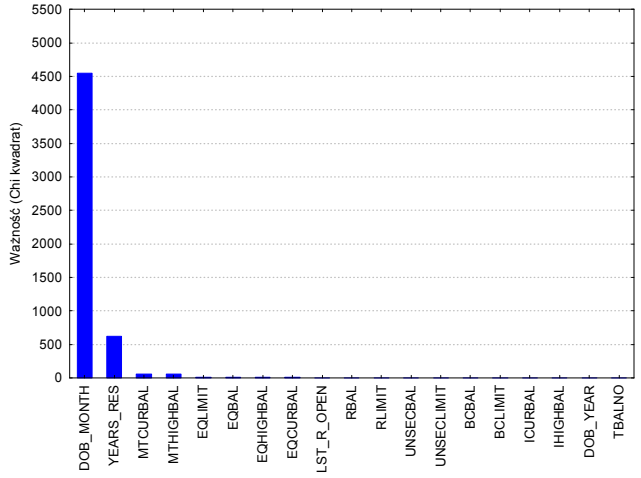
Do pliku danych dołączyliśmy zmienną *Pusta* o losowo wygenerowanych wartościach, nie związanych ze zmienną zależną. Ważność dla tej zmiennej daje nam pewien poziom odniesienia, pozwalający ocenić ważność potencjalnych predyktorów.

Braki danych

Wcześniej zajmowaliśmy się wpływem wartości poszczególnych predyktorów na zmienną zależną. W praktyce dosyć często użyteczna jest informacja, że dana osoba nie

odpowiedziała na jedno z pytań lub informacja jest niedostępna z innego powodu. Do zbadania wpływu występowania braków danych w poszczególnych zmiennych na zmienną zależną zastosujemy tę samą metodę co w rozdziale „Badanie wpływu poszczególnych zmiennych na odpowiedź na ofertę". Tym razem jednak jako potencjalne predyktory zastosujemy przekształcone zmienne: wszystkie wartości zostaną w nich zastąpione przez 0, a braki danych przez 1 (takie podejście zaproponowano w pracy [1]).

Na rysunku poniżej widzimy wyniki procedury dla przekształconych zmiennych (informujących, czy w oryginalnej zmiennej wystąpił brak danych). Rzuca się w oczy bardzo duża ważność dla zmiennej *Dob_month* (miesiąc urodzenia) – jest ona prawie 5 razy większa niż ta statystyka dla wskaźnika *Beacon* (najsilniejszego predyktora wśród oryginalnych zmiennych). Nie ma rozsądnego uzasadnienia dla aż tak silnego związku *Dob_month* i *Buyer* – chyba, że bardzo głęboko wierzymy w astrologię..

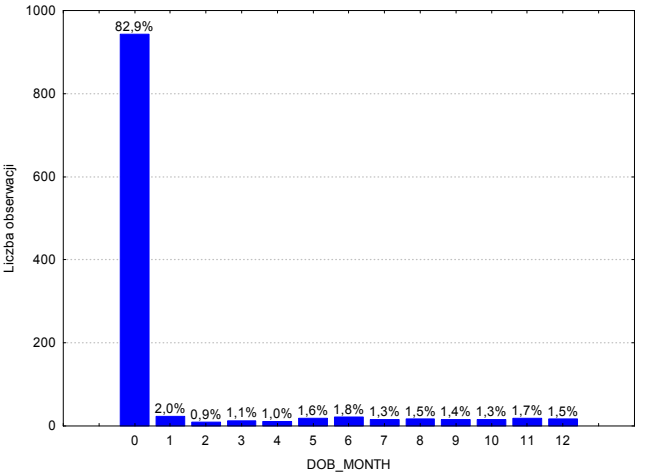


Przyjrzyjmy się bliżej licznościom klasy *T* i *N*, gdy wartość zmiennej *Dob_month* jest i nie jest podana. W tabeli poniżej widzimy, że wszyscy klienci, których miesiąc urodzenia znamy, kupili kartę kredytową (klasa *T*), a ci którzy nie kupili, zawsze mają brak danych w zmiennej *Dob_month*.

Tabela liczności (CreditP_ABD.sta)

BUYER	DOB_MONTH Podana wartość	DOB_MONTH Brak danych	Wiersz Razem
N	0	10999	10999
T	1139	1858	2997
Ogół grup	1139	12857	13996

Zobaczmy, jak wygląda rozkład wartości zmiennej *Dob_month* (rysunek poniżej). Jak widać, przyjmuje ona 13 wartości, z czego najczęstsza jest 0. Jak się okazuje, oznacza tak naprawdę brak danych.



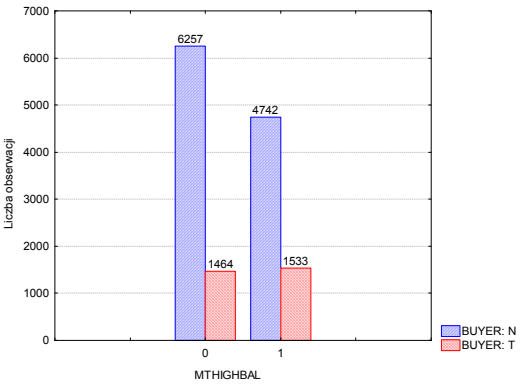
Wartości zmiennej *Dob_month* były wpisywane dla osób, które zdecydowały się zakupić kartę kredytową. Jest to przykład anachronicznej zmiennej: stąd jej dokładna zależność z wartościami zmiennej zależnej. Oczywiście zmienna taka zepsuje nam model. Przede wszystkim może uniemożliwić jego stosowanie dla nowych klientów, dla których będzie ona zawsze nieznana. Oznacza to, że jeśli wystąpi w formule modelu, to odpowiedź modelu będzie również nieznana. Poza tym jej uwzględnienie w modelu obciąży jego przewidywania.

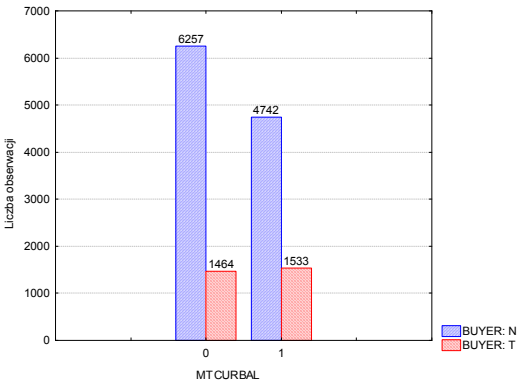
Wróćmy do wykresu ważności braków danych (str. 107). Wskazuje on, że wystąpienie braku danych w zmiennych *Years_res*, *Mtcurbal* i *Mthighbal* również ma pewien wpływ na odpowiedź klienta, przy czym znacznie ważniejsze są braki danych w zmiennej *Years_res*. Ponadto zmienna ta ma dużą ważność (jest piąta pod tym względem, zob. wykres ważności na str. 106), natomiast cechy *Mtcurbal* i *Mthighbal* mają niewielki wpływ na odpowiedź na ofertę.

Najpierw przyjrzymy się bliżej wpływowi braku danych dla *Years_res* na odpowiedź klienta. Jak widać, wpływ jest bardzo silny, ale nie ma pewnej reguły, tak jak w przypadku zmiennej *Dob_month*. Jest trochę dziwne, że osoby, dla których nie mamy informacji o czasie zamieszkania w obecnym miejscu, chętniej podejmują decyzję o zakupie karty kredytowej, niemniej na pewno powinniśmy uwzględnić ten fakt w budowanym modelu. Ponieważ również oryginalna zmienna *Years_res* jest ważna (patrz str. 106), to skategoryzujemy tę zmienną i wprowadzimy specjalną kategorię „wystąpił brak danych" (podobnie jak w przypadku zmiennej *Home_value*, patrz str. 102).

	Tabela liczności (CreditP_ABD.sta)			
	BUYER	YEARS_RES 0	YEARS_RES 1	Wiersz Razem
Liczba	N	10781	218	10999
% z kolumny		80,39%	37,26%	
Liczba	T	2630	367	2997
% z kolumny		19,61%	62,74%	
Liczba	Ogół gr up	13411	585	13996

Wpływ braków danych w zmiennych *Mtcurbal* i *Mthighbal* na decyzję klienta przedstawiają histogramy widoczne poniżej. Tym razem związek jest zdecydowanie słabszy niż dla *Years_res*, jednak ze względu na dużą liczbę przypadków może być ważny. Zauważmy, że liczby decyzji pozytywnych i negatywnych w grupach „wystąpił brak danych" (1) i „nie wystąpił brak danych" dla obu zmiennych są identyczne. Nasuwa to podejrzenie, iż braki danych dla zmiennych *Mtcurbal* i *Mthighbal* pojawiają się zawsze razem.





W poniższej tabeli widzimy, że istotnie tak jest: *Mtcurbal* jest pusta wtedy i tylko wtedy, gdy *Mthighbal* jest pusta.

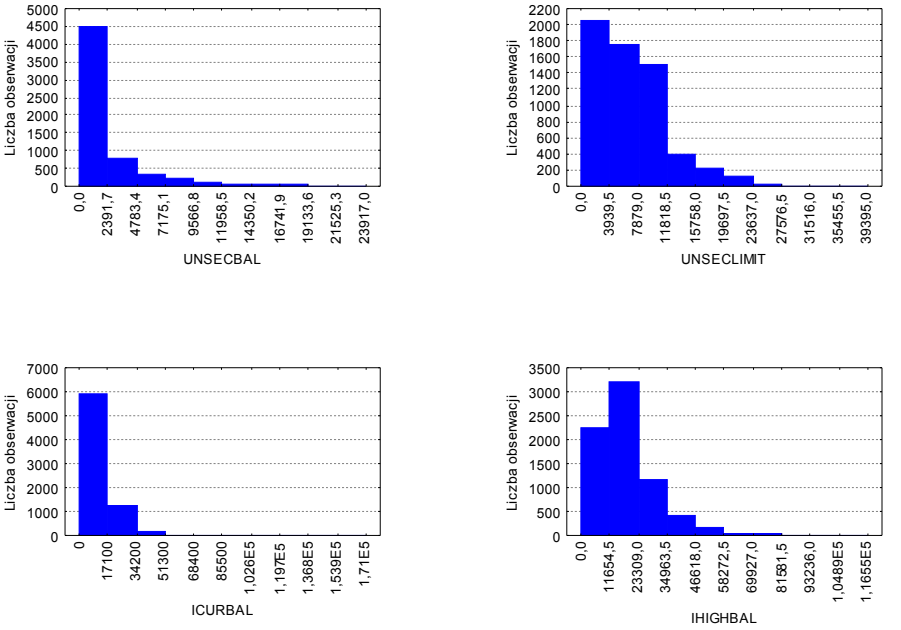
Tabela liczności (CreditP_ABD.sta)

MTHIGHBAL	MTCURBAL 0	MTCURBAL 1	Wiersz Razem
0	7721	0	7721
1	0	6275	6275
Ogół grup	7721	6275	13996

Zmienne: *Mtcurbal* i *Mthighbal* mają dosyć mały wpływ na odpowiedź klienta i są w ponad 40% wypełnione brakami danych. W przypadku stosowania metod modelowania wrażliwych na nadmierną liczbę predyktorów powinniśmy zamiast nich użyć jednej zmiennej informującej, że w *Mtcurbal* i *Mthighbal* wystąpił brak danych. My jednak skategoryzujemy je, wprowadzając osobną kategorię brak danych.

Mamy jeszcze 4 zmienne (*Ihighbal*, *Icurbal*, *Unseclimit*, *Unsecbal*) z dużą ilością braków danych, dla których nie ustaliliśmy sposobu postępowania z pustymi wartościami. Mają one udział braków danych na poziomie 40% do 60% i w dalszej analizie powinniśmy na nie zwrócić uwagę. Zmienne te nie mają kluczowego wpływu na odpowiedź klienta (jak widać na wykresie ważności na str. 106, najważniejsza spośród nich (*Unseclimit*) jest 18 pod względem ważności). Wystąpienie braku danych w tych zmiennych ma całkowicie nieistotny wpływ na zmienną zależną (por. str. 107). Te dwa fakty pozwalają nam bezpiecznie zastąpić braki danych w tych zmiennych wartością liczbową, pomimo tego, że taka procedura może obciążyć wyniki. W naszym przypadku to obciążenie powinno być nieznaczne.

Najczęściej braki danych zastępuje się średnią lub medianą. Przy wyborze jednej z tych opcji uwzględnimy rozkład zmiennych. W rozdziale „Statystyki opisowe i rozkłady" (str. 100) wyznaczyliśmy skośność wszystkich zmiennych. Interesujące nas zmienne mają stosunkowo dużą skośność i ich rozkłady są silnie asymetryczne, widać to wyraźnie również na poniższych histogramach. W takim wypadku rozsądniejsze wydaje się zastąpienie braków danych medianami.

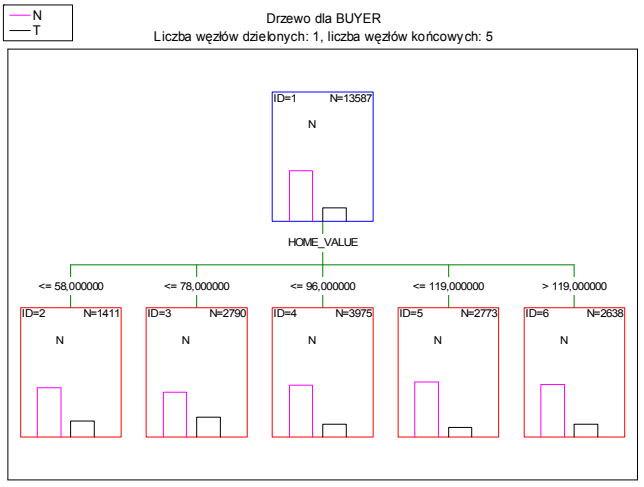


Przekształcenie zmiennych ilościowych na jakościowe (kategoryzacja)

Na podstawie dotychczasowego badania danych doszliśmy do wniosku, że ze względu na występowanie braków danych i ich wpływ na zmienną zależną powinniśmy skategoryzować zmienne *Home_value* (str. 102), *Year_res* (str. 108) oraz zmiennych *Mtcurbal* i *Mthighbal*.

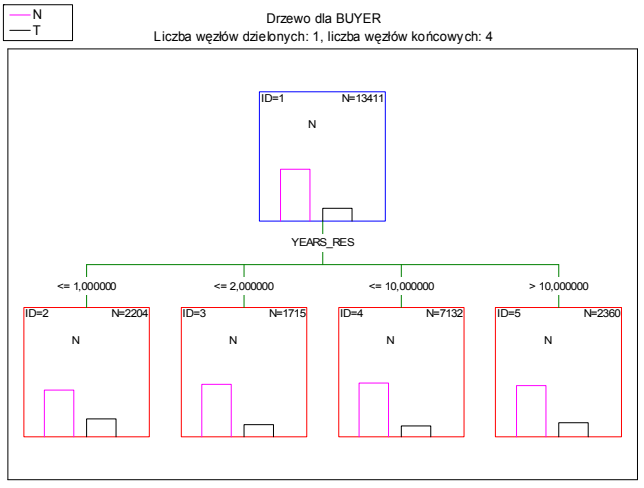
Istnieje wiele sposobów przekształcania zmiennych ilościowych na jakościowe. W pewnych sytuacjach dysponujemy wiedzą, jaki podział jest najlepszy (np. na podstawie wcześniejszych badań, teorii opisujących zjawisko) lub jaki podział na klasy jest powszechnie przyjęty. Czasami dzielimy zmienną jakościową na grupy, bazując na jej rozkładzie; więcej informacji o kategoryzacji zmiennych przedstawiono w [5]. My wykorzystamy metodę CHAID (opisaną w [4]) do dobrania podziałów tak, aby rozkład zmiennej zależnej był najbardziej zróżnicowanych w wybranych kategoriach przekształcanej zmiennej.

Zaczniemy od podziału na klasy zmiennej *Home_value*. Przed samym podziałem wszystkie wartości 0 potraktujemy jako brak danych, zgodnie z wyciągniętymi wcześniej wnioskami (str. 102).



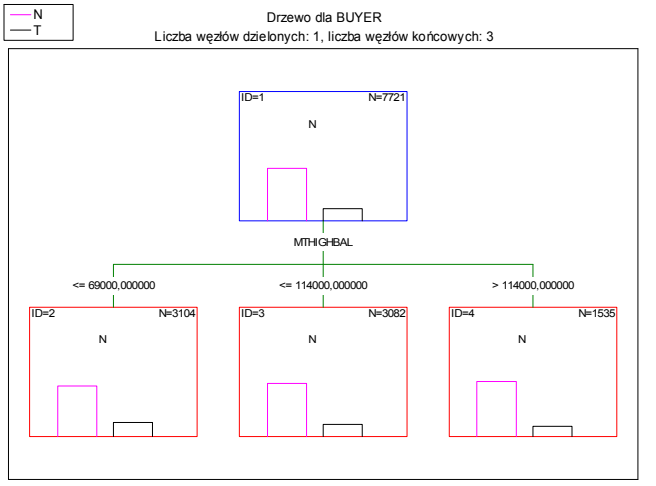
Na rysunku powyżej widzimy wyniki analizy uzyskane przy domyślnych ustawieniach procedury CHAID w systemie *STATISTICA Data Miner*. Otrzymane w wyniku podziału klasy mają odpowiednio duże liczności i możemy je zaakceptować. Pamiętajmy, że oprócz uzyskanych metodą CHAID 5 klas, utworzymy jeszcze klasę odpowiadająca kodowi braku danych, tj. 0.

Tak samo postępujemy dla zmiennej *Year_res* – tym razem otrzymujemy cztery klasy (piątą będzie stanowił brak danych o tej cesze).



Również kategoryzację zmiennych *Mtcurbal* i *Mthighbal* spróbujemy przeprowadzić za pomocą algorytmu CHAID. W przypadku zmiennej *Mthighbal* uzyskujemy rozsądny

podział na 3 klasy, natomiast w przypadku drugiej z nich procedura nie znajduje żadnego podziału. Oznacza to, że CHAID nie jest w stanie wykryć zależności między zmienną *Mtcurbal* a zmienną zależną. Podobny wynik uzyskaliśmy, stosując procedurę *Dobór zmiennych i analiza przyczyn*: na wykresie na str. 106 zmienna ta jest daleko z tyłu pod względem ważności.



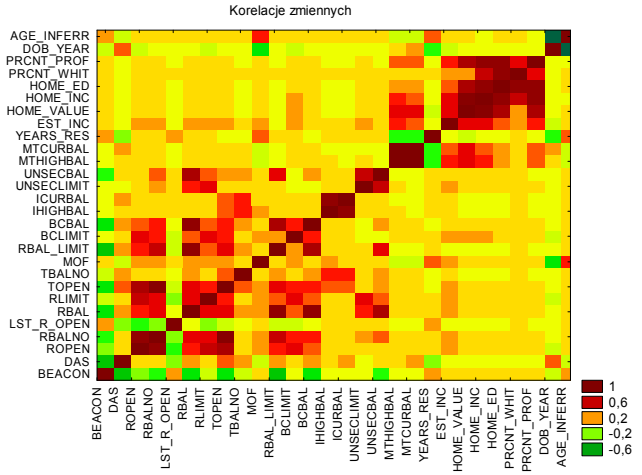
Biorąc pod uwagę dużą liczbę braków danych w zmiennej *Mtcurbal* i jej związek ze zmienną *Mthighbal* (opisany powyżej na str. 109, ponadto współczynnik korelacji liniowej tych zmiennych wynosi 0,97), możemy bezpiecznie pominąć *Mtcurbal* w procesie modelowania.

Związki między zmiennymi niezależnymi

W rzeczywistych danych prawie zawsze potencjalne predyktory są w mniejszym lub większym stopniu ze sobą związane. Zależności między predyktorami są dla wielu metod modelowania dużym utrudnieniem, dotyczy to zwłaszcza tradycyjnych metod statystycznych, np. regresji i analizy dyskryminacyjnej, ale każdej procedurze mogą zaszkodzić bardzo silnie powiązane predyktory (choćby dlatego, że pewną informację wielokrotnie bierzemy pod uwagę bez żadnego uzasadnienia). Z drugiej strony związki miedzy predyktorami mogą nam pomóc radzić sobie z brakami danych i wybrać odpowiednie zmienne do modelu. Ponadto pewne zmienne powinny być ze sobą związane, np. wiek i czas wykształcenia w latach – jeśli takiego związku nie będą wykazywały nasze dane, to powinniśmy sprawdzić dlaczego, albowiem nasuwa to podejrzenia co do poprawności danych.

Zaczniemy od analizy korelacji liniowej pomiędzy predyktorami ilościowymi (w analizie pomijamy „puste zmienne" wykryte wcześniej w rozdziale „Statystyki opisowe i rozkłady" str. 100). Na wykresie poniżej duże wartości bezwzględne współczynnika korelacji liniowej oznaczone są ciemnym kolorem. Jak widać wiele zmiennych jest silnie powiązanych ze sobą. Największy współczynnik korelacji liniowej (0,97) występuje

między zmiennymi *Mtcurbal* i *Mthighbal* (pisaliśmy już o tym wcześniej). Bardzo silnie związane ze sobą są również zmienne *Age_inferr* i *Dob_year*: współczynnik korelacji dla nich wynosi -0,9. Tylko dla 3 predyktorów brak innej zmiennej, takiej że moduł współczynnika korelacji jest większy od 0,5. Przy modelowaniu będziemy musieli pamiętać o silnych związkach miedzy predyktorami.



Zmienne *Dob_year* (rok urodzenia) i *Age_inferr* (szacowany wiek) przynoszą tę samą informację: o wieku potencjalnego klienta. Powinny być one bardzo silnie, ujemnie skorelowane (im większy rok urodzenia danej osoby, tym jest ona młodsza). Jak wspomnieliśmy wcześniej, jest tak faktycznie: współczynnik korelacji wynosi aż -0,9.

Zauważmy, że analiza ważności zmiennych (str. 106) dała bardzo podobne wyniki dla cech *Dob_year* i *Age_inferr*. Jednak zmienna *Dob_year* może być kłopotliwa przy budowie modelu, a zwłaszcza jego stosowaniu. Mianowicie jest ona pusta w ponad 30% przypadków. Biorąc pod uwagę wszystkie te fakty, możemy bezpiecznie pominąć w modelowaniu zmienną *Dob_year*.

Modelowanie

Wykorzystując zdobytą do tej pory wiedzę, zbudujemy teraz model przewidujący odpowiedź klientów na ofertę.

Ponieważ w naszych danych występują zmienne różnego typu o skośnych rozkładach z nietypowymi obserwacjami, zastosujemy metodę odporną na takie problemy, czyli drzewa klasyfikacyjne C&RT, a konkretnie ich implementację w module *STATISTICA Drzewa interakcyjne*.

Przed wykonaniem analizy dane losowo podzielimy na próbę uczącą i testową, tak że próba testowa będzie zawierała około 30% przypadków z oryginalnego pliku.

Dobór kosztów błędnych klasyfikacji

Ponieważ mamy niezrównoważone liczności klas, ustalimy różne koszty błędnych klasyfikacji. Jest to bardzo ważny parametr modelu, zasadniczo wpływający na trafność przewidywań; aby dobrać odpowiednie koszty błędnych klasyfikacji, dopasujemy wstępne modele dla kosztów błędnych klasyfikacji od 1:1 do 1:3,5.

W ocenie modelu uwzględnimy trzy wielkości:

1. Frakcję błędnych przewidywań dla obu klas.
2. Procent przypadków, które w rzeczywistości należały do klasy 'T', a model zakwalifikował je do klasy 'N' (udział będzie wyliczany względem liczby wszystkich obserwacji w klasie 'T').
3. Procent przypadków, które w model zakwalifikował do klasy 'T', a w rzeczywistości należały one do klasy 'N' (udział będzie wyliczany względem liczby wszystkich obserwacji zaklasyfikowanych jako 'T' przez model).

Najlepiej byłoby uzyskać jak najmniejsze wartości wszystkich trzech wskaźników, niestety zazwyczaj jest to nieosiągalne (mamy tu podobną sytuację jak z używanym samochodem, który powinien być ładny, dobry i tani, ale dostępne są tylko samochody mające dwie te cechy). To, który ze wskaźników jest kluczowy, zależy od przeznaczenia modelu.

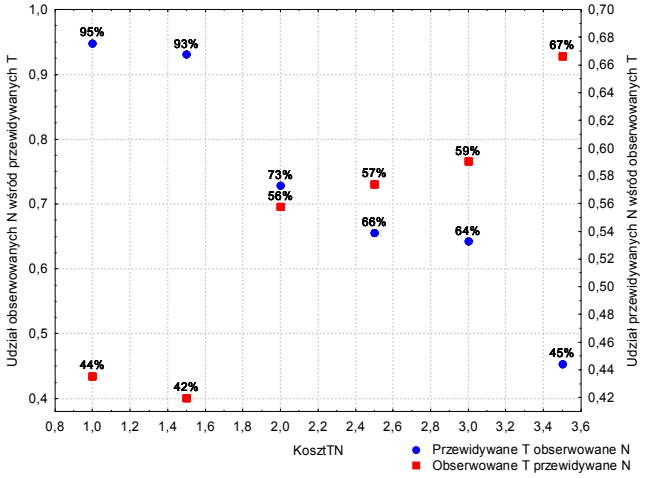
Jeśli naszym celem jest poznanie klientów, to zazwyczaj najważniejsze będzie wychwycenie przez model możliwie jak największej części klientów pozytywnie odpowiadających na ofertę (minimalizujemy wskaźnik 2) przy zachowaniu rozsądnej ogólnej frakcji błędnych przewidywań (wskaźnik 1). Podobnie będzie w sytuacji, gdy planujemy kampanię i jeśli zysk z jednej przeprowadzonej transakcji uzasadnia zwiększenie liczby osób, z którymi się kontaktujemy. Zdarza się również, że celem jest wychwycenie wszystkich osób lub zdarzeń należących do pewnej klasy – wtedy również najważniejszy będzie wskaźnik 2.

W sytuacji, gdy ustalona jest liczba kontaktów w ramach kampanii (ponieważ np. telecentrum jest w stanie skontaktować się z określoną liczbą osób) i chcemy z dużej grupy potencjalnych klientów wybrać podzbiór o największej gęstości osób zainteresowanych ofertą, to najważniejszy będzie wskaźnik 3. Oczywiście model musi być na tyle liberalny, aby wygenerował odpowiednią liczbę osób do kontaktu.

Na wykresie poniżej widzimy stopy błędów w próbie testowej uzyskiwane dla drzewa C&RT dla różnych kosztów błędnych klasyfikacji (inne parametry modelu to: minimalna liczność dzielonego węzła 300, minimalna liczność potomka 75). Zasadniczo wraz ze wzrostem dysproporcji kosztów błędnych klasyfikacji dla klas 'T' i 'N' (od 1:1 do 1:3,5) dosyć szybko spada udział obserwacji 'T" zaklasyfikowanych przez model do klasy 'N' (procent wyliczany jest względem wszystkich obserwacji należących do klasy 'T'). Z drugiej strony wraz ze wzrostem dysproporcji kosztów błędnych klasyfikacji rośnie udział błędnych przewidywań 'T' (w stosunku do wszystkich przewidywanych 'T').

Jak już wspomnieliśmy wcześniej, w zależności od uwarunkowań ważność wskaźników błędu jest różna. Wybierzemy koszty błędnych klasyfikacji 1:2,5 – oba błędy są dla nich

w miarę rozsądne i lepsze od całkowicie losowego wyboru (jeśli całkowicie przypadkowo klasyfikowalibyśmy przypadki do klasy ‘T’ i ‘N’ przy zachowaniu proporcji obu klas ze zbioru danych, to oba błędy wynoszą około 80%).



Tworzenie i ocena drzewa klasyfikacyjnego

Po dobraniu kosztów błędnych klasyfikacji równej 1:2,5 zbudujemy właściwy model. Model budujemy dla minimalnej liczności dzielonego węzła równej 300 i minimalnej liczności potomka równej 15. W wyniku działania procedury *STATISTICA Drzewa interakcyjne* uzyskamy drzewo klasyfikacyjne złożone z 81 węzłów, z czego 41 jest węzłami końcowymi. Można powiedzieć, że model dzieli zbiorowość potencjalnych klientów na 41 segmentów o różnej proporcji decyzji ‘Tak’ i ‘Nie’.

Model ma następujące wskaźniki dobroci dopasowania (obliczane względem liczności całych prób):

	Próba ucząca	Próba testowa
Całkowita frakcja błędów	22,83%	24,47%
Frakcja błędnie przewidywanych ‘T’	12,61%	12,44%
Frakcja błędnie przewidywanych ‘N’	10,22%	12,03

Porównanie udziałów błędnych przewidywań w próbie testowej i uczącej nie daje bardzo zdecydowanego sygnału o przeuczeniu modelu, ale rysuje się takie zagrożenie.

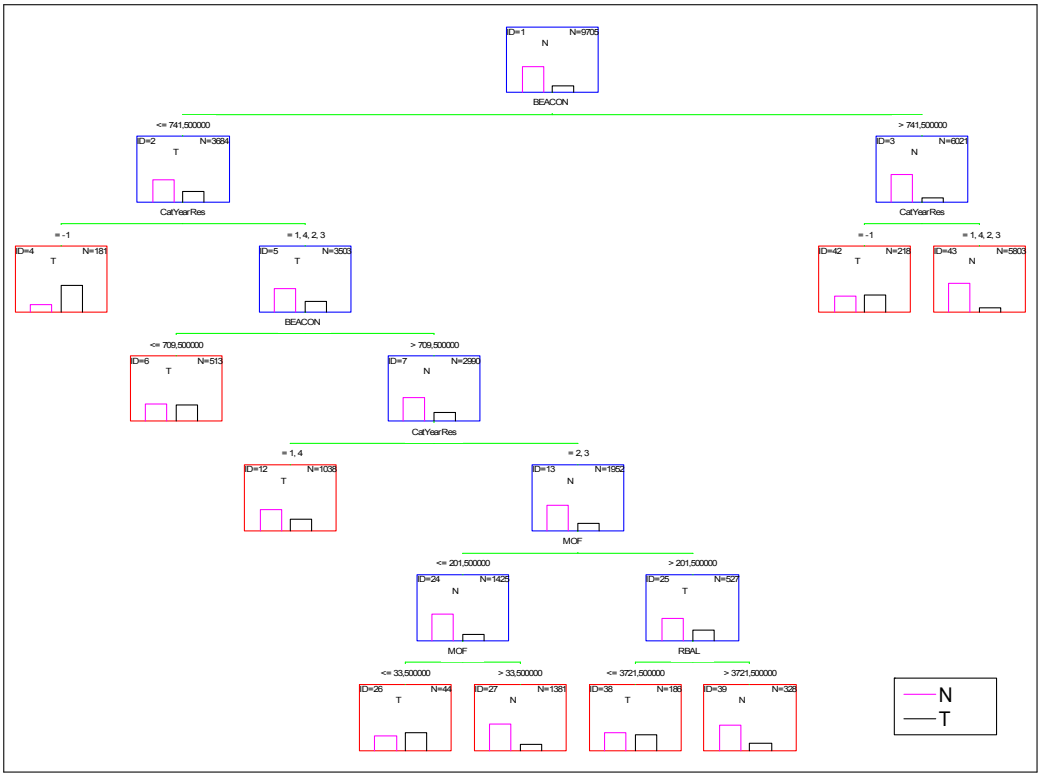
Zauważmy, że dzięki zastosowaniu różnych kosztów błędnych klasyfikacji uzyskaliśmy bliskie błędy dla obu klas – przy równych kosztach frakcja błędów dla klasy ‘T’ jest zdecydowanie wyższa.

Moduł *STATISTICA Drzewa interakcyjne* umożliwia automatyczne przycięcie drzewa, tzn. usunięcie tych podziałów, które faktycznie nic ważnego nie wnoszą (lub są nawet

szkodliwe). Po użyciu tej procedury otrzymamy drzewo złożone tylko z 17 węzłów. Zobaczmy, jak ono działa:

	Próba ucząca	Próba testowa
Całkowita frakcja błędów	23,43%	23,31%
Frakcja błędnie przewidywanych 'T'	11,61%	10,30%
Frakcja błędnie przewidywanych 'N'	11,82%	13,01%

Jak widać, przycięte drzewo ma nieco gorszą całkowitą frakcję błędów w próbie uczącej, ale lepszą w próbie testowej. Co ważne wskaźniki błędów w obu próbach są bliższe, a więc i ryzyko przeuczenia mniejsze. Za przyciętym drzewem przemawia też jego prostota.



Przycięte drzewo widzimy na rysunku powyżej. Zgodnie z uzyskanymi wcześniej wynikami pierwsze podziały dyktowane są przez zmienne *Beacon* i *CatYearRes* (pojawiają się one też w dalszych podziałach), w szczególności ważna jest informacja o tym, że dla danej osoby nie znamy czasu zamieszkiwania w obecnej lokalizacji (klasa *-1* zmiennej *CatYearRes*). Pewnym zaskoczeniem jest pojawienie się w drzewie podziałów wyznaczonych przez zmienne *Mof* i *Rbal*, które nie miały wysokiej ważności w badaniu jednowymiarowym (str. 106). Zauważmy jednak, że zmienne te pojawiają się tylko po

jednej stronie drzewa dla określonych kategorii zmiennych *Beacon* i *CatYearRes*; tłumaczy to, dlaczego nie widać ich silnego wpływu w badaniach jednowymiarowych, gdy nie bierzemy pod uwagę wpływu innych predyktorów (interakcji). Natomiast nie powinno nas dziwić, iż wskaźnik *Das* nie występuje w drzewie: przenosi on informację o zdolności kredytowej potencjalnego klienta, podobnie jak *Beacon*, i jest z nim silnie skorelowany.

Sprawdźmy teraz, czy nasz model jest użyteczny. Przyjrzymy się udziałowi przypadków przewidzianych przez drzewo do klasy 'T', które w rzeczywistości należą do klasy 'N'. W naszym przypadku wynosi on ok. 59% dla próby testowej. Patrząc z drugiej strony, uzyskaliśmy ok. 41% poprawnie przewidzianych przypadków należących do klasy 'T'.

Jakie to ma znaczenie praktyczne? Otóż jeśli zastosujemy nasz model dla nowych danych, to wśród wskazanych przez niego osób do kontaktu około 40% przyjmie ofertę. Przy losowym wyborze klientów do kontaktu ze zbioru danych, na którym pracowaliśmy, udział pozytywnych odpowiedzi wyniesie około 20%. Podsumowując, nasz model jest dwukrotnie lepszy od przypadkowego wyboru. Powinniśmy jednak zwrócić uwagę na jedną bardzo ważną sprawę: nasz zbiór danych ma sztucznie zwiększoną proporcję pozytywnych odpowiedzi, w rzeczywistości zaledwie kilka procent osób przyjęło ofertę.

Zastanówmy się chwilę nad zyskownością naszego projektu. Oczywiście nasza symulacja będzie uproszczona, ale pomimo to unaoczni nam pewne fakty i zależności. Przyjmijmy, że kampania ma dotyczyć 20 000 osób – jest to największa możliwa do zrealizowania liczba kontaktów. Dysponujemy dużo większą bazą danych, z której wybieramy potencjalnych klientów. Załóżmy, że całkowity zysk z jednego pozyskanego klienta wyniesie 200 zł (jest to raczej ostrożny szacunek w przypadku kart kredytowych). Jeśli z bazy danych wybieramy osoby do kontaktu losowo, to około 2% z nich zakupi kartę. Nasz model da nam próbę o co najmniej dwukrotnie większej gęstości pozytywnych odpowiedzi na ofertę. Przy takich samych kosztach przeprowadzenia kampanii uzyskamy co najmniej 400 klientów więcej, a więc nasz przychód z kampanii wzrośnie o 80 000 zł. Pojawią się oczywiście dodatkowe koszty: zbudowania modelu i jego wdrożenia: dla takiego zadania jak przedstawiliśmy, zazwyczaj wynoszą one około 25 000 zł (dotyczy to pierwszej realizacji projektu, przy następnych będą one znacznie niższe). Nawet gdybyśmy do kosztów zaliczyli cenę oprogramowania (które będziemy mogli przecież używać wielokrotnie dla następnych projektów), wynosząca dla minimalnej wersji pozwalającej budować tego typu modele około 11 000 zł, a dla pełnej wersji *STATISTICA Data Miner* około 50 000 zł, to i tak koszty będą mniejsze niż osiągnięte korzyści.

Literatura

1. D. Pyle, *Data Preparation for Data Mining*, Morgan Kaufmann, 1999.
2. D. Pyle, *Business Modeling and Data Mining*, Morgan Kaufmann, 2003.
3. M.J.A Berry, G.S. Linoff, *Mastering Data Mining*, Wiley, 2000.
4. M.J.A Berry, G.S. Linoff, *Data Mining Techniques*, Wiley, 1997.
5. J. Han, M. Kamber, *Data Mining. Concepts and Techniques*, Morgan Kaufman, 2001.

