EKONOMETRIA ECONOMETRICS 2(48) • 2015

ISSN 1507-3866
e-ISSN 2449-9994

Mariusz Kubus
Politechnika Opolska
e-mail: m.kubus@po.opole.pl

IDENTYFIKACJA POTENCJALNYCH NABYWCÓW POLIS UBEZPIECZENIOWYCH W WARUNKACH MOCNO NIEZBILANSOWANEJ PRÓBY UCZĄCEJ

IDENTIFICATION OF POTENTIAL PURCHASERS OF THE INSURANCE POLICIES UNDER HARD UNBALANCED TRAINING SET

DOI: 10.15611/ekt.2015.2.08

Streszczenie: Dysponując zbiorem danych o dokonanych transakcjach i cechach demograficznych klientów, można wykorzystać *scoring* marketingowy w akcjach skierowanych na wspieranie sprzedaży. Stosowane w nim metody dyskryminacji napotykają często problem niezbilansowania próby uczącej oraz zbędnej informacji w postaci zmiennych, które nie mają związku z zakupem produktu. W artykule analizowany jest rynek ubezpieczeń, gdzie do scoringu wykorzystana będzie metoda ważonych *k* najbliższych sąsiadów oraz wielowymiarowe kryteria doboru zmiennych. Selekcja zmiennych znacząco wpłynęła na zwiększenie liczby poprawnie zidentyfikowanych potencjalnych nabywców polisy.

Słowa kluczowe: *scoring* marketingowy, metoda ważonych *k* najbliższych sąsiadów, selekcja zmiennych.

Summary: Having given the data set with executed transactions and customer demographic features one can use marketing scoring to support sales campaign. The discrimination methods used in the scoring often face the problem of imbalance classes and irrelevant variables. In this paper, we analyze the insurance market, where the scoring is performed with a use of the weighted *k* nearest neighbors and multivariate filters. The feature selection significantly contributed to increasing the number of correctly identified potential purchasers of the insurance policy.

Keywords: marketing scoring, weighted *k* nearest neighbours method, feature selection.

1. Wstęp

W obecnej sytuacji nasycenia rynku i ostrej konkurencji proces sprzedaży wspierany jest różnego rodzaju kampaniami marketingowymi. Nowoczesne przedsiębiorstwo

wykorzystuje jako zasób informację o kliencie w postaci bazy danych, która zawiera dokonane w przeszłości transakcje oraz wybrane cechy demograficzne. Jednym z ważnych zadań analizy ilościowej jest dobór grupy docelowej klientów do oferty sprzedażowej. Wyłonienie klientów, którzy z dużym prawdopodobieństwem staną się nabywcami produktu, pozwala maksymalizować wielkość sprzedaży i jednocześnie minimalizować koszty akcji zachęcającej do kupna. Zadanie to realizowane jest przez *scoring* marketingowy, który polega na rangowaniu klientów według prawdopodobieństwa zakupu produktu. W modelowaniu takiego rankingu wykorzystuje się metody dyskryminacji. Najpopularniejsza jest tu regresja logistyczna, a następnie agregowane drzewa czy też sieci neuronowe.

W artykule do rangowania klientów zaproponowano wykorzystanie metody ważonych k najbliższych sąsiadów. Poważną jej wadą jest jednak spadek dokładności klasyfikacji, gdy w zbiorze danych są zmienne, które nie mają wpływu na zmienną objaśnianą (*irrelevant variables*). Stawianym celem jest zatem nie tylko weryfikacja skuteczności proponowanej metody, lecz także pokazanie, że etap selekcji zmiennych może znacząco poprawić jakość modelu. W praktyce etap ten wspierany jest często przez wiedzę ekspercką, tu ograniczamy się do automatycznej selekcji zmiennych z wykorzystaniem wyłącznie metod statystycznych.

2. Opis zbioru danych

Przedmiotem badań będzie rynek ubezpieczeń przyczep kempingowych w Holandii. Informacja o rynku pochodzi ze zbioru danych zgromadzonych przez firmę ubezpieczeniową o blisko 10 000 klientach. Są oni opisani 43 cechami demograficznymi (np. typ klienta, poziom wykształcenia, charakter pracy, związek małżeński, liczba samochodów, przeciętny dochód) oraz 42 cechami opisującymi aktywność na rynku ubezpieczeń (np. liczby różnych polis ubezpieczeniowych czy też wartości ubezpieczeń: OC, łodzi, roweru, samochodu dostawczego, od pożaru itp.). Dwie zmienne charakteryzujące typ klienta są nominalne, a pozostałe ilościowe, które poddano dyskretyzacji. Rolę zmiennej objaśnianej odgrywa binarna zmienna wskazująca, czy klient wykupił ubezpieczenie przyczepy kempingowej. W roku 2000 dane te były przedmiotem konkursu CoIL (Computational Intelligence and Learning) na najlepiej klasyfikujący model i jego interpretację [van der Putten, van Someren 2000]. W konkursie dokonano losowego podziału na zbiór uczący (5822 klientów) i testowy (4000 klientów). Zbiór ma kilka charakterystyk wspólnych dla wielu praktycznych zagadnień, przez co wydaje się, że metody, jakie będą tu przedstawione, znajdą też zastosowanie w warunkach polskich. Mianowicie próba jest silnie niezbilansowana – tylko około 6% klientów, którzy wykupili polisę na przyczepę kempingową. Wymiar przestrzeni cech jest dość wysoki. Predyktory są z sobą skorelowane oraz można przypuszczać, że nie wszystkie mają wpływ na zmienną objaśnianą. Dane są bazą pełnej informacji gromadzonej o klientach, a nie zbiorem z wyselekcjonowanymi przez ekspertów cechami, które powinny mieć związek z badanym zjawiskiem, jakim jest wykupienie polisy ubezpieczającej przyczepę kempingową.

Celem analizy jest identyfikacja przyszłych nabywców polisy ubezpieczeniowej na przyczepę kempingową. Cel będzie osiągany przez budowę modelu dyskryminacyjnego, klasyfikującego klientów do grupy nabywców z oszacowanym prawdopodobieństwem. Model umożliwi wyłonienie grupy docelowej, do której wysłana będzie oferta sprzedaży. Chodzi o to, by obniżyć koszty akcji sprzedażowej, kierując ją jedynie do tych klientów, którzy wykazują największe prawdopodobieństwo pozytywnej odpowiedzi na ofertę.

3. Metody analizy: klasyfikatory i *scoring*

Wykupienie ubezpieczenia na przyczepę kempingową reprezentowane jest przez zmienną dychotomiczną, mamy zatem do czynienia z zadaniem dyskryminacji dwóch klas. Metoda ta polega na zbudowaniu modelu na podstawie zbioru wielowymiarowych obserwacji ze znaną przynależnością do klas (tzw. zbiór uczący). Model taki, nazywany często klasyfikatorem, ma odzwierciedlać wpływ zmiennych objaśniających $X_1, \ldots, X_p$ na zmienną objaśnianą Y, a więc na przynależność do klasy. Jego postać wykorzystuje albo funkcje (np. regresja logistyczna), albo warunki logiczne (np. drzewa klasyfikacyjne). Warto nadmienić, że niektóre klasyfikatory (metody *k* najbliższych sąsiadów lub agregowanych drzew klasyfikacyjnych) mają postać ,,czarnej skrzynki", tzn. nie mają wyraźnych, interpretowalnych reguł klasyfikacyjnych. Model wykorzystuje się do przewidywania klasy nowo zaobserwowanych obiektów, o nieznanej przynależności do klas. Musi zatem cechować się wysoką zdolnością generalizacji, co uzyskuje się przez unikanie nadmiernego dopasowania do danych ze zbioru uczącego w etapie uczenia (zob. np. [Hastie, Tibshirani, Friedman 2009]).

3.1. Metoda ważonych *k* najbliższych sąsiadów wkNN

Jedną z najprostszych ideowo metod dyskryminacji jest metoda *k* najbliższych sąsiadów (kNN). Etap uczenia (estymacji parametrów modelu) jest w niej pominięty. Do obserwacji $\boldsymbol{x}$ poddawanej klasyfikacji wyznacza się *k* najmniej od niej oddalonych obiektów ze zbioru uczącego. W najprostszym przypadku obiekt $\boldsymbol{x}$ przypisany jest do klasy najczęściej występującej wśród *k* najbliższych sąsiadów (tzw. głosowanie majoryzacyjne). Metoda kNN ma związek z bayesowską regułą klasyfikacji, mianowicie można pokazać, że prawdopodobieństwo *a posteriori* dla wybranej klasy może być szacowane jako liczba obiektów tej klasy w stosunku do *k*. Inny sposób głosowania polega na wprowadzeniu wag, które są funkcjami odległości obiektu rozpoznawanego $\boldsymbol{x}$ od najbliższych sąsiadów. Wagi są w odpowiednich klasach sumowane, a obserwacja $\boldsymbol{x}$ przypisana do tej z maksymalną sumą. Wybrane funkcje ważące przedstawiono w tab. 1. Nieco inny sposób ważenia proponują Dudani [1976] oraz Gou i in. [2012]. Modyfikacja ta nazywana jest metodą ważonych *k* najbliższych sąsiadów (wkNN). Kluczową decyzją w stosowaniu metody jest wybór

wartości parametru *k*. Pewne sugestie podali Enas i Choi [1986]. Zwykle jednak proponuje się wyznaczenie optymalnej wartości *k* za pomocą sprawdzania krzyżowego.

Tabela 1. Wybrane funkcje ważące odległości $d(\boldsymbol{x},\boldsymbol{x}_i)$ klasyfikowanego obiektu od *i*-tego najbliższego sąsiada

Inwersja	Funkcja Gaussa	Funkcja Epanechnikova
$K(d)=\frac{1}{d}$	$K(d)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{d^2}{2}\right)$	$K(d)=\frac{3}{4}\left(1-d^2\right)$

Źródło: opracowanie własne.

W literaturze zaproponowano szereg miar oceniających jakość modelu dyskryminacyjnego (zob. np. [Misztal 2014]). Niewątpliwie najpopularniejszą z nich jest błąd klasyfikacji, czyli frakcja obiektów błędnie rozpoznanych. W wielu zadaniach praktycznych błędne klasyfikowanie do różnych klas nie jest jednak jednakowo traktowane. Na przykład zaklasyfikowanie klienta, który nie wykupił polisy, do klasy tych, którzy ją wykupili, wiąże się z pewnym kosztem skierowanej do niego akcji sprzedażowej. W sytuacji przeciwnej – to znaczy, gdy klient, który wykupił polisę, zostaje sklasyfikowany do tych, którzy jej nie wykupili – firma traci zysk, który prawdopodobnie wielokrotnie przewyższa koszty związane z akcją sprzedażową. W odniesieniu do dwóch klas opracowano szereg miar rozróżniających błędne klasyfikacje. Wprowadźmy następujące oznaczenia i nazewnictwo. Jedną z klas nazwijmy wyróżnioną (*positive class*). Reprezentuje ona zjawisko, które jest przedmiotem zainteresowania, u nas jest to klasa klientów, którzy wykupili polisę na przyczepę kempingową. W niektórych zastosowaniach jest to klasa nielicznie reprezentowana w zbiorze danych. Drugą klasę nazwijmy niewyróżnioną (*negative class*). Miary jakości modelu formułuje się za pomocą macierzy klasyfikacji, której elementy oznaczymy przez:

- TP (*true positives*): liczba poprawnie sklasyfikowanych obiektów z klasy wyróżnionej.
- TN (*true negatives*): liczba poprawnie sklasyfikowanych obiektów z klasy niewyróżnionej.
- FP (*false positives*): liczba błędnie sklasyfikowanych obiektów do klasy wyróżnionej.
- FN (*false negatives*): liczba błędnie sklasyfikowanych obiektów do klasy niewyróżnionej.

Najpopularniejszą miarą uwzględniającą rodzaj błędnych oraz prawidłowych klasyfikacji jest pole pod krzywą ROC (*receiver operating characteristic*):

$$AUC = 0{,}5 \cdot \left(1 + \frac{TP}{TP+FN} - \frac{FP}{FP+TN}\right). \tag{1}$$

Z kolei Galar i in. [2013] w przypadkach silnego niezbilansowania klas rekomendują średnią geometryczną dokładność klasyfikacji:

$$GM = \sqrt{\frac{TP}{TP+FN} \cdot \frac{TN}{FP+TN}}. \tag{2}$$

Obie miary przyjmują maksymalną wartość równą jeden w razie braku błędnych klasyfikacji.

Z modelowaniem klasyfikatorów związane są pewne problemy wynikające, ogólnie rzecz ujmując, z jakości danych. Problem może stanowić zawarty w danych szum, braki danych, obserwacje nietypowe, zmienne niemające wpływu na zmienną objaśnianą (*irrelevant variables*), predyktory skorelowane między sobą czy też silne niezbilansowanie klas. W analizowanym zbiorze grupa posiadaczy polisy ubezpieczającej przyczepę kempingową stanowi jedynie około 6% całej próby. W tym przypadku większość klasyfikatorów faworyzuje klasę większościową, inaczej mówiąc, mało obiektów z klasy mniejszościowej – u nas wyróżnionej – jest poprawnie sklasyfikowanych. Ilustruje to tab. 2, gdzie zastosowano metodę kNN na zbiorze danych o ubezpieczeniach. Przez P oznaczono klasę wyróżnioną, a więc klientów, którzy wykupili polisę. Dla danych oryginalnych tylko 10 obiektów klasy wyróżnionej zostało poprawnie sklasyfikowanych, a błąd klasyfikacji nie różni się zbytnio od prawdopodobieństwa *a priori*. Po zastosowaniu bilansowania liczba TP = 159, ale wartości miar *AUC* oraz *GM* maleją. Błąd klasyfikacji około 42% jest w tym przypadku mniejszy od domyślnej reguły klasyfikacji, ale powstaje pytanie, czy wynik taki jest satysfakcjonujący, jeśli wiadomo, że frakcja klientów, którzy wykupili polisę, jest bliska 6%.

Tabela 2. Oceny dokładności klasyfikacji metody kNN dla zbioru ubezpieczenia na próbie testowej

	Dane oryginalne	Dane po zbilansowaniu
Błąd klasyfikacji	0,0697	0,4192
AUC	0,6109	0,5629
GM	0,3930	0,2955
Macierze klasyfikacji	predykcja N P N 3711 51 P 228 10	predykcja N P N 2164 1598 P 79 159

Źródło: obliczenia własne.

Z kolei zmienne nieistotne (*irrelevant variables*) mogą przyczyniać się do nadmiernego dopasowania modelu do danych ze zbioru uczącego, co obniża dokładność klasyfikacji nowych obiektów. Prezentowany wcześniej klasyfikator kNN jest przykładem metody wrażliwej na ten problem.

3.2. *Scoring* marketingowy

Oprócz klasyfikacji obiektu, ważną informacją jest prawdopodobieństwo, z jakim zaobserwowany obiekt należy do klasy wyróżnionej, a więc prawdopodobieństwo *a posteriori*. Posiadanie takiej informacji stwarza możliwość uporządkowania obiektów – u nas klientów – według prawdopodobieństwa bycia w grupie potencjalnych nabywców produktu. Mamy więc do czynienia z zadaniem scoringu marketingowego. Ponieważ celem jest wyłonienie grupy docelowej, do której będzie skierowana akcja zachęcająca, podstawowym problemem jest ustalenie liczby klientów. Chodzi więc o ustalenie wartości progowej dla prawdopodobieństw *a posteriori*, nazywanej punktem odcięcia. W podejściu tym kalkuluje się zyski i koszty akcji sprzedażowej. Dotarcie do dużej liczby klientów wiąże się z kosztami, chodzi więc o to, by w grupie klientów objętych akcją było jak najwięcej tych, którzy kupią produkt. Wybranie zbyt małej grupy docelowej naraża z kolei firmę na nieosiągnięcie potencjalnych zysków. Jeżeli koszty i zyski są łatwe do oszacowania, łatwo wyznaczyć punkt odcięcia, rozwiązując zadanie optymalizacji. Jeśli koszty i zyski nie są łatwe do oszacowania, można arbitralnie ustalić liczbę klientów docelowych na podstawie budżetu przeznaczonego na akcję zachęcającą do kupna produktu.

We wspomnianym konkursie CoIL Challenge 2000 ustalono arbitralnie liczbę 800 klientów. Jako miarę oceniającą jakość modelu scoringowego przyjęto liczbę poprawnie wytypowanych klientów – z klasy nabywców polisy – wśród 800 pierwszych w rankingu w zbiorze testowym. Nadesłano 43 rozwiązania (uczestnicy indywidualni lub zespoły). Najlepszy wynik to 121 zidentyfikowanych klientów, następnie 115 oraz 112. Mediana wynosiła 103, a najczęściej osiągany wynik to 109 zidentyfikowanych klientów (zob. [Elkan 2001]). W dwóch najlepszych rozwiązaniach stosowano naiwny klasyfikator Bayesa.

4. Selekcja zmiennych

Zadanie selekcji zmiennych można sformułować jako problem optymalizacji kombinatorycznej [Tsamardinos, Aliferis 2003]. Przyjmując pewną klasę modeli i kryterium ich jakości, należy wybrać taki podzbiór S zbioru zmiennych objaśniających, by uzyskać optymalną wartość funkcji oceny. Zwykle postuluje się też, by podzbiór S był możliwie najmniejszy. Sprawdzenie wszystkich kombinacji zmiennych jest w dużych wymiarach niepraktyczne. Ponadto wyczerpujące przeszukiwanie przestrzeni wszystkich podzbiorów zmiennych objaśniających może prowadzić do nadmiernego dopasowania modelu do danych ze zbioru uczącego (zob. [Jensen i Cohen 2000]). Zwykle stosuje się więc heurystyczne lub stochastyczne techniki przeszukiwania (zob. np. [Reunanen 2006]). Właśnie ze względu na relację, jaka zachodzi między przeszukiwaniem przestrzeni kombinacji zmiennych a przeszukiwaniem przestrzeni parametrów modelu, metody selekcji zmiennych dzieli się obecnie na trzy grupy (zob. np. [Guyon i in. 2006]). Pierwszą z nich – być może najpopularniej-

szą – stanowią metody doboru zmiennych (*filters*). Druga grupa metod to selekcja zmiennych za pomocą selekcji modeli (*wrappers*). Wreszcie trzecia grupa (*embedded methods*) to metody, gdzie mechanizm selekcji zmiennych wpisany jest w algorytm uczący (np. drzewa klasyfikacyjne). Do analizy potencjalnych nabywców polisy ubezpieczeniowej wykorzystane będą wybrane kryteria doboru zmiennych (*filters*).

Metody doboru zmiennych charakteryzują się tym, że selekcji dokonuje się przed etapem budowania modelu. Wchodzą więc w skład technik przygotowujących dane do etapu modelowania. W metodach tych przyjmuje się kryterium oceniające zmienne ze względu na ich zdolność dyskryminacji. Zaznaczmy, że jest to wybór czysto heurystyczny i nie ma gwarancji, że zmienne optymalne ze względu na to kryterium będą zarazem budowały model optymalny, co zwykle oznacza najdokładniej klasyfikujący. Kryterium może oceniać indywidualny wpływ zmiennej objaśniającej na objaśnianą (jednowymiarowe) lub oceniać zdolność dyskryminacji całego podzbioru predyktorów (wielowymiarowe).

Szeroko stosowaną grupę kryteriów jednowymiarowych stanowią miary bazujące na entropii. Wydaje się, że ich atrakcyjność wynika z możliwości ujęcia nieliniowego związku między zmiennymi, odporności na obserwacje oddalone oraz możliwości stosowania w sytuacji, gdy predyktory są z różnych skal pomiarowych. Przypomnijmy, że w analizowanym zbiorze o ubezpieczeniach mamy dwie zmienne jakościowe oraz zmienne ilościowe. Entropia definiowana jest jako średnia ilość informacji, jaką niesie cecha:

$$H(X) = -\sum_i P(X = x_i) \cdot \log_2 P(X = x_i). \tag{3}$$

Prawdopodobieństwa szacowane są zwykle frakcjami. W przypadku, gdy X jest ilościową zmienną ciągłą, poddaje się ją dyskretyzacji lub rozkłady szacowane są estymatorami jądrowymi [Kwak, Choi 2002]. Związek między predyktorem a zmienną objaśnianą mierzy miara wzajemnej informacji (*mutual information*):

$$MI(Y, X) = H(Y) + H(X) - H(Y, X). \tag{4}$$

Ponieważ jej oceny są obciążone na korzyść zmiennych przyjmujących większą liczbę wariantów, w literaturze zaproponowano kilka sposobów normalizacji. Jedną z nich jest symetryczna niepewność (*symmetrical uncertainty*):

$$SU(Y, X) = 2 \cdot \frac{MI(Y, X)}{H(Y) + H(X)}. \tag{5}$$

Miara ta przyjmuje wartości z przedziału [0;1], gdzie zero oznacza całkowitą niezależność zmiennych. Kryteria jednowymiarowe – oceniające indywidualną moc dyskryminacyjną zmiennych – są powszechnie stosowane w dużych wymiarach, gdyż cechują się szybkością działania. Nie potrafią jednak odkryć niektórych

interakcji między predyktorami (zob. [Kubus 2015]) oraz nie eliminują zmiennych powielających informacje (tzw. zmiennych redundantnych).

Wśród kryteriów wielowymiarowych najpopularniejsze są miary maksymalizujące korelacje zmiennych objaśniających ze zmienną objaśnianą i jednocześnie minimalizujące interkorelacje (tzn. korelacje między predyktorami). Wymienić tu można miarę korelacji grupowej [Ghiselli 1964]:

$$H_1(S) = \frac{k \cdot \overline{r}(X_i, Y)}{\sqrt{k + k(k-1) \cdot \overline{r}(X_i, X_j)}}, \tag{6}$$

czy też pojemność informacji Hellwiga [1969], do której korektę zaproponował Walesiak [1987]:

$$H_2(S) = \sqrt{\det[r_{ij}]} \cdot \sum_{X_i \in S} \frac{r^2(X_i, Y)}{\sum_{X_j \in S} \left| r(X_i, X_j) \right|}. \tag{7}$$

W powyższych formułach $\overline{r}$ oznacza średnią korelację dla zmiennych wchodzących w skład ocenianego podzbioru S, k jest liczbą kardynalną tego podzbioru, a $[r_{ij}]$ macierzą interkorelacji w podzbiorze S. W charakterze korelacji r można przyjąć w powyższych formułach miarę symetrycznej niepewności. W tej postaci Gatnar [2005] stosował kryterium H_2 przy doborze zmiennych do modelu agregowanych drzew klasyfikacyjnych, natomiast kryterium H_1 było wykorzystane przez Halla [2000] w jego algorytmie selekcji zmiennych CFS (*correlation-based feature selection*). Podane kryteria eliminują zmienne redundantne, co powinno wpływać na poprawę stabilności modeli. Z drugiej strony ocena podzbiorów zmiennych wiąże się z problemem kombinatorycznym. Do efektywnego ich stosowania w dużym wymiarze należy stosować techniki przeszukiwania. Na przykład Hall [2000] we wspomnianym algorytmie CFS wykorzystał metodę najpierw-najlepszy (*best-first*).

5. Wyniki analizy

Bezpośrednie zastosowanie metod dyskryminacji do klasyfikacji klientów jako potencjalnych nabywców polisy nie dało zadowalających rezultatów. W każdej z zastosowanych metod błąd klasyfikacji (szacowany na zbiorze testowym) był nieco większy od błędu domyślnej reguły klasyfikacji równego 5,95% (tab. 3). Naiwny klasyfikator Bayesa, ze względu na swą naturę, był stosowany po selekcji zmiennych, gdzie zastosowano kryterium H_2.

W związku z tym, że klasyfikacja nic nie wniosła do zadania identyfikacji potencjalnych nabywców polisy, zadanie to zostanie przeformułowane na rangowanie klientów według prawdopodobieństwa, że wykupią ubezpieczenie, czyli według prawdopodobieństwa *a posteriori*.

Tabela 3. Oceny modeli dyskryminacyjnych w problemie klasyfikacji nabywców polisy ubezpieczeniowej na przyczepę kempingową

	Błąd klasyfikacji	*AUC*	*GM*
Regresja logistyczna	0,0603	0,6961	0,5601
Las losowy	0,0625	0,6994	0,5680
Drzewa wzmacniane	0,0635	0,6734	0,5224
Naiwny klasyfikator Bayesa	0,0663	0,6557	0,4900
Klasyczna kNN	0,0603	0,7111	0,5851
wkNN	0,0670	0,6007	0,3667

Źródło: obliczenia własne.

Scoring potencjalnych nabywców ubezpieczenia przyczepy kempingowej przeprowadzono za pomocą metody wkNN, która zastosowana była po selekcji zmiennych. Zmienne jakościowe charakteryzujące typ klienta zamieniono na binarne zmienne sztuczne (*dummy variables*). W celu doboru optymalnych parametrów metody – liczba sąsiadów oraz postać funkcji ważącej – posłużono się zbiorem walidacyjnym. Z oryginalnego zbioru uczącego wylosowano 2822 obiektów do próby walidacyjnej, a pozostałe 3000 obiektów potraktowano jako próbę uczącą. Progowa liczba klientów w zbiorze walidacyjnym została ustalona na 564, tak by frakcja była taka sama jak w zbiorze testowym, gdzie dla celów porównawczych przyjęto 800 klientów [Elkan 2001]. Sprawdzane były wartości k ze zbioru {1,8,17,34,50,100}. Wartość $k = 8$ to liczba sąsiadów wyznaczona według sugestii Enasa i Choi [1986]. Sprawdzono funkcje ważące z tab. 1 oraz klasyczną metodę kNN. Ponadto sprawdzono dwa kryteria doboru zmiennych: H_1 oraz H_2 . Stosowano je z heurystyczną metodą przeszukiwania najpierw-najlepszy (*best-first*). Najlepszą kombinacją parametrów okazała się liczba sąsiadów 100, funkcja Epanechnikova oraz kryterium Hellwiga z korektą Walesiaka. Na zbiorze walidacyjnym uzyskano 94 poprawne identyfikacje klientów z wykupioną polisą ubezpieczenia przyczepy kempingowej. Z kolei na zbiorze testowym – co stanowi rzeczywiste sprawdzenie przydatności modelu – liczba poprawnych identyfikacji wyniosła 117. Liczba 117 posiadaczy polis wśród 800 pierwszych w rankingu według prawdopodobieństw *a posteriori* daje frakcję 14,6%, a więc znacząco większą od frakcji posiadaczy polisy w całym zbiorze testowym. Bez etapu selekcji zmiennych zidentyfikowano poprawnie jedynie 99 posiadaczy polisy, a więc o 18 mniej. Optymalny podzbiór zmiennych wyznaczony przez kryterium H_2 to:

- rodzina z klasy średniej,
- związek małżeński,
- poziom wykształcenia,
- wynajmowanie domu,
- przeciętny dochód,

- wartość ubezpieczenia samochodu,
- wartość ubezpieczenia od pożaru,
- wartość ubezpieczenia łodzi.

Wyraźnie większe wartości miary *SU* przyjmowały zmienne wyrażające wartości ubezpieczeń.

6. Zakończenie

W artykule zwrócono uwagę na możliwość wykorzystania metody *k* najbliższych sąsiadów w scoringu marketingowym, w jej rzadziej stosowanym wariancie uwzględniającym wagi zależne od odległości. Wydaje się, że informacja, jaką niosą wagi, może mieć szczególne znaczenie w analizach zbiorów silnie niezbilansowanych. W przeprowadzonej analizie optymalna wartość parametru *k* wyniosła 100, co znacząco przekracza wartość $k = 8$, jaką sugerowałyby wyniki teoretyczne Enasa i Choi [1986]. Być może wynika to z tego, że *scoring* wykorzystuje prawdopodobieństwa *a posteriori*, które są dokładniej szacowane dla większych wartości *k*.

Kluczowe znaczenie dla zbudowania dokładnego modelu scoringowego miał etap selekcji zmiennych. Bez niego metoda wkNN identyfikowała tylko 99 klientów, to jest mniej niż mediana wyników uzyskanych w konkursie CoIL Challenge 2000. W badaniu zastosowano jedynie selekcję automatyczną. Takie podejście, o ile jest skuteczne, ma niewątpliwą zaletę, gdyż algorytm selekcji zmiennych staje się narzędziem wydobycia wiedzy z danych, to jest relacji i związków, które nie są znane czy widoczne na pierwszy rzut oka. Z relacji Elkana [2001] wynika, że żaden z uczestników konkursu CoIL Challange 2000 nie ograniczał się do metod selekcji automatycznej. Stosowano różnego rodzaju przekształcenia oryginalnych cech czy też zmniejszano wymiar przestrzeni przez arbitralny wybór predyktorów.

Warto jeszcze zwrócić uwagę na to, że choć modele dyskryminacyjne klasyfikujące potencjalnych nabywców polisy ubezpieczeniowej okazały się nieskuteczne, to analiza prawdopodobieństw *a posteriori* stosowana w scoringu marketingowym potrafiła wydobyć informację cenną dla kampanii sprzedażowej.

Literatura

Dudani S.A., 1976, *The distance-weighted k-nearest-neighbor rule,* IEEE Transactions on Systems, Man and Cybernetics, 6(4), s. 325-327.

Elkan C., 2001, *Magical thinking in data mining: Lessons from CoIL Challenge 2000*, Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining (KDD'01), s. 426-431.

Enas G.G., Choi S.C., 1986, *Choice of the smoothing parameter and efficiency of k-nearest neighbor classification*, Computer and Mathematics with Applications, 12A(2), s. 235-244.

Galar M., Fernandez A., Barrenechea E., Herrera F., 2013, *EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling*, Pattern Recognition, 46(12), s. 3460-3471.

Gatnar E., 2005, *Dobór zmiennych do zagregowanych modeli dyskryminacyjnych*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu, nr 1076, Taksonomia 12, Klasyfikacja i analiza danych – teoria i zastosowania, s. 79-86.

Ghiselli E.E., 1964, *Theory of Psychological Measurement*, McGraw-Hill, New York.

Gou J., Du L., Zhang Y., Xiong T., 2012, *A new distance-weighted k-nearest neighbour classifier*, Journal of Information and Computational Science, vol. 9, no. 6, s. 1429-1436.

Guyon I., Gunn S., Nikravesh M., Zadeh L., 2006, *Feature Extraction: Foundations and Applications*, Springer, New York.

Hall M., 2000, *Correlation-based feature selection for discrete and numeric class machine learning*, Proceedings of the 17th International Conference on Machine Learning, Morgan Kaufmann, San Francisco.

Hastie T., Tibshirani R., Friedman J., 2009, *The Elements of Statistical Learning: Data Mining, Inferance, and Prediction*, 2nd edition, Springer, New York.

Hellwig Z., 1969, *Problem optymalnego wyboru predykant*, Przegląd Statystyczny, nr 3-4.

Jensen D. D., Cohen P. R., 2000, *Multiple comparisons in induction algorithms*, Machine Learning, 38(3), s. 309-338.

Kubus M., 2015, *Feature selection and the chessboard problem*, XXXIII Międzynarodowa Konferencja Wielowymiarowa Analiza Statystyczna, Łódź (w druku).

Kwak N., Choi C.H., 2002, *Input feature selection by mutual information based on Parzen window*, IEEE Transactions. Pattern Analysis and Machine Intelligence, vol. 24, no. 12, s. 1667-1671.

Misztal M., 2014, *Wybrane metody oceny jakości klasyfikatorów – przegląd i przykłady zastosowań*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 328, Taksonomia 23, Klasyfikacja i analiza danych – teoria i zastosowania, s. 156-166.

Reunanen J., 2006, *Search Strategies,* [w:] Guyon I., Gunn S., Nikravesh M., Zadeh L. (red.), *Feature Extraction: Foundations and Applications*, Springer, New York.

Tsamardinos I., Aliferis C.F., 2003, *Towards principled feature selection: relevancy, filters and wrappers*, Proceedings of the Workshop on Artificial Intelligence and Statistics.

Walesiak M., 1987, *Zmodyfikowane kryterium doboru zmiennych objaśniających do liniowego modelu ekonometrycznego*, Przegląd Statystyczny, 1, s. 37-43.

van der Putten, P., van Someren, M., 2000, *CoIL challenge 2000: The insurance company case*, Technical Report 2000-2009, Leiden Institute of Advanced Computer Science, Universiteit van Leiden, http://www.liacs.nl/~putten/library/cc2000.