PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 107
Taksonomia 17 2010

Marcin Owczarczuk
Szkoła Główna Handlowa w Warszawie

PROGNOZOWANIE ODCHODZENIA DO KONKURENCJI KLIENTÓW TELEFONII KOMÓRKOWEJ

Streszczenie: Celem tego artykułu jest analiza możliwości wykorzystania modeli ekonometrycznych zbudowanych z wykorzystaniem uogólnionych estymatorów *maximum score* (maksymalizacji wyniku) do prognozowania odchodzenia do konkurencji klientów telefonii komórkowej. Z naszego badania wynika, że wykorzystanie *maximum score* daje modele o dokładniejszych prognozach oraz bardziej stabilne w czasie niż drzewa klasyfikacyjne oraz o znacznie prostszej strukturze niż modele regresji logistycznej. W analizie wykorzystano dane pochodzące od jednego z operatorów telefonii komórkowej w Polsce.

1. Wstęp

Rynek telefonii komórkowej w Polsce jest obecnie nasycony, pozyskanie nowych klientów jest możliwe właściwie tylko przez przejęcie klientów konkurencji. Jednocześnie klienci często zmieniają operatora. Pozyskanie nowych klientów jest drogie, a efektywność akcji pozyskaniowych jest niewielka w związku z nasyceniem rynku. Znacznie tańsze jest utrzymanie, czyli takie działania, które powstrzymają klientów od odejścia (*churn*).

Głównym celem w takiej sytuacji jest prognozowanie, którzy klienci są skłonni odejść, oraz zastosowanie wobec nich odpowiednio wcześniej odpowiedniej akcji zatrzymaniowej. Naturalnym modelem do prognozy takiego zjawiska jest model dwumianowy. Zmienną objaśnianą jest fakt odejścia lub pozostania klienta w sieci, a zmiennymi objaśniającymi są jego charakterystyki użyciowe, takie jak liczba wykonanych połączeń głosowych czy SMS, liczba i kwota doładowań czy aktualna taryfa klienta. Jednocześnie operator chciałby kierować swoje akcje zatrzymaniowe tylko do pewnego wąskiego grona klientów rzeczywiście zagrożonych odejściem. Żeby jeszcze bardziej podkreślić rolę odejść w działalności operatora, rozważmy następującą sytuację. Niech miesięczny poziom odejść wynosi 6%. Jest to w przybliżeniu rzeczywista wartość pochodząca z przedstawionego w dalszej części pracy przykładu empirycznego. Zatem po roku w przypadku niepodjęcia akcji zatrzymaniowych pozostałoby $0{,}94^{12} = 47{,}59\%$ klientów, czyli blisko połowa odeszłaby w ciągu zaledwie roku.

Problem prognozowania odejść na rynku telefonii komórkowej i wykorzystania w tym celu narzędzi ilościowych jest dobrze opisany w literaturze. Godne uwagi są prace [Hung, Yen, Wang 2006; Pendharkar 2009; Wei, Chiu 2002]. Autorzy tych prac wykorzystują z dobrym skutkiem takie narzędzia data miningowe, jak drzewa klasyfikacyjne, sieci neuronowe oraz algorytmy genetyczne. Wei, Chiu [2002] analizują dane na temat klientów kontraktowych, czyli tzw. postpaid, pochodzące od jednego z tajwańskich operatorów telefonii komórkowej. Wykorzystują przy tym dane pochodzące z aktywacji, takie jak rodzaj płatności, a także informacje pochodzące z billingów, takie jak liczba minut wykonanych połączeń. Ponadto używają podejścia wielomodelowego polegającego na estymowaniu wielu modeli, a następnie uśrednianiu ich wyniku jako końcowego rezultatu. Z kolei Pendharkar [2009] wykorzystał sieci neuronowe w połączeniu z algorytmami genetycznymi w celu prognozy odchodzenia klientów kontraktowych pewnego operatora telefonii komórkowej. Hung, Yen i Wang [2006] wykorzystali sieci neuronowe oraz drzewa klasyfikacyjne. Ich dane pochodziły od tajwańskiego operatora sieci komórkowej, a zmienne objaśniające były demograficzne, takie jak płeć, billingowe, takie jak kwota abonamentu czy liczba rozmów wewnątrz sieci, oraz dotyczące serwisu, takie jak liczba zmian numeru telefonu. Przedmiotem badań w wymienionych publikacjach są klienci płacący abonament, czyli tzw. klienci typu postpaid. Liczba potencjalnych zmiennych objaśniających użytych podczas modelowania nie przekracza kilkudziesięciu.

W tym artykule omówimy przydatność takich narzędzi statystycznych, jak regresja logistyczna, drzewa klasyfikacyjne oraz modele dwumianowe estymowane metodą *maximum score* [Owczarczuk 2009] do prognozowania zjawiska odchodzenia do konkurencji. Wykorzystamy w tym celu dane na temat klientów „na kartę", pochodzące od jednego z polskich operatorów telefonii komórkowej. Omówimy strategię działań zatrzymaniowych, schemat generacji danych, w tym definicję odejścia oraz wykorzystane zmienne objaśniające. Następnie przedstawimy wyniki eksperymentu obliczeniowego. Przede wszystkim zbadamy precyzję prognoz każdego z modeli oraz zanalizujemy problem stabilności modeli w czasie. Pokażemy, że klasyfikatory liniowe są stabilniejsze w czasie niż regułowe.

2. Schemat generacji danych

Naturalnym wyborem modelu służącego do opisu odchodzenia do konkurencji jest model dwumianowy. Zmienną objaśnianą definiujemy w następujący sposób

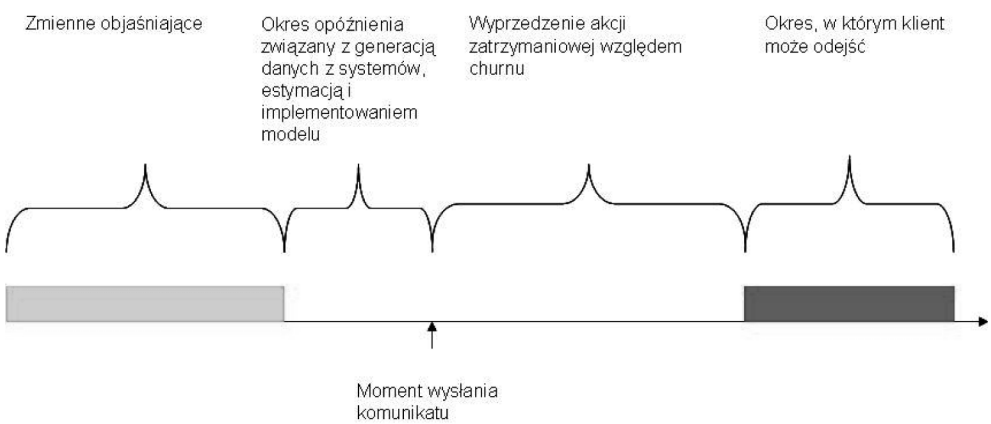
$$y = \begin{cases} 1, \text{ gdy klient odszedł z sieci} \\ 0, \text{ gdy klient pozostał w sieci} \end{cases}.$$

Zmienne objaśniające są określane na podstawie okresu przed obserwacją zmiennej y i mogą dotyczyć właściwie wszystkich charakterystyk, jakie ma do dyspozycji operator. Potencjalna lista zmiennych jest bardzo duża i z reguły obej-

muje od kilkuset do nawet kilkudziesięciu tysięcy pozycji. Typowy schemat generacji danych jest przedstawiony na rys. 1.

Dysponując już gotowym modelem, oszacowanym na danych historycznych, można przejść do fazy jego wykorzystania na danych bieżących. Wykorzystanie polega na znalezieniu grupy klientów o określonej liczności, wśród której będzie jak najwięcej osób zagrożonych odejściem. Oczywiście rozmiar tej grupy można wyrazić przez jej frakcję w całej bazie i oznaczyć przez τ. Najczęściej wykorzystanie polega na następujących czynnościach: (1) dla każdego klienta wyznacz prognozę $P(y_i = 1)$, czyli prawdopodobieństwo odejścia, (2) posortuj klientów ze względu na wartość $P(y_i = 1)$, (3) wybierz τ klientów o największej wartości tego prawdopodobieństwa, (4) wyślij komunikat marketingowy tylko do tych klientów.

Zauważmy, że tak przedstawione postępowanie ma intuicyjne uzasadnienie – wśród klientów o największej wartości prawdopodobieństwa odejścia powinno być rzeczywiście najwięcej klientów, którzy odejdą.



**Churn* to odejście do konkurencji.

Rys. 1. Schemat generacji danych

Źródło: opracowanie własne.

3. Opis danych do badania empirycznego

W naszym badaniu empirycznym wykorzystano dane pochodzące od jednego z operatorów telefonii komórkowej w Polsce. Do dyspozycji są dwa zbiory danych – jeden pochodzący z pewnego miesiąca roku 2007, a drugi z pewnego miesiąca roku 2008, przy czym czas zebrania obu zbiorów dzieli pół roku odstępu. Zbiór z roku 2007 ma 122 098 obserwacji, a zbiór z roku 2008 ma 45 497 obserwacji. Dane dotyczą klientów typu prepaid, czyli na kartę. Nasze badanie, w porównaniu z wcześniejszymi pracami, jest nowatorskie w następujących obszarach: modelujemy

zachowanie klientów na kartę, czyli prepaid, a nie na abonament, czyli postpaid. Modelowanie klientów typu prepaid jest o wiele trudniejsze, gdyż ich użycie jest o wiele mniej regularne. Nie płacą oni comiesięcznego abonamentu i mogą wykonywać doładowania, kiedy chcą. Ponadto brakuje precyzyjnej definicji odejścia. Ten problem jest omawiany dokładnie w dalszej części pracy. Ponadto nasz zbiór potencjalnych zmiennych objaśniających jest o wiele większy od wykorzystanych w innych badaniach – w naszym badaniu jest 1381 potencjalnych zmiennych objaśniających. Ich opis znajduje się w dalszej części pracy. W poprzednich badaniach zmiennych objaśniających było co najwyżej kilkadziesiąt.

4. Definicja odejścia

Poprzednie badania wykorzystywały dane na temat klientów kontraktowych. Klient kontraktowy podpisuje umowę z operatorem telefonii komórkowej na świadczenie usług. Zobowiązuje się też do comiesięcznego płacenia abonamentu. Odejściem jest tutaj właśnie fakt złożenia pisemnego wypowiedzenia, a więc jest to zdarzenie dobrze zdefiniowane. W przypadku telefonu na kartę klient nie podpisuje kontraktu, a co za tym idzie – nie musi go rozwiązywać na piśmie. Kluczowe zatem staje się posiadanie odpowiedniej definicji odejścia. W tej pracy używamy następującej: „klient odszedł, jeżeli miał sześciotygodniowy okres bez połączeń wychodzących i przychodzących". Dodatkowo, ponieważ chcemy, aby akcja marketingowa została wysłana na krótko przed tym momentem, prognozujemy fakt, czy klient odejdzie cztery tygodnie po momencie analizy. Innymi słowy – chcemy prognozować, z wyprzedzeniem czterech tygodni, moment, w którym klient rozpocznie 6 tygodni braku aktywności. Po pierwsze, definicja odejścia powinna umożliwić szybką weryfikację – chcemy czekać tak krótko, jak to tylko możliwe, aby się przekonać, że klient rzeczywiście zaprzestał korzystania z usług. Po drugie – definicja powinna być pewna, chcemy, aby prawdopodobieństwo zdarzenia, że klient po takim okresie nieaktywności nie wykona połączenia aż do momentu dezaktywacji karty SIM, było maksymalnie wysokie. Okres 6 tygodni jest kompromisem pomiędzy tymi dwoma celami – spośród klientów, którzy nie korzystali z telefonu przez 6 tygodni, niewielu wykonywało połączenia po tym okresie. Tymczasem było bardzo wiele „wybudzeń" po 3, 4 i 5 tygodniach nieaktywności.

5. Zmienne objaśniające

W naszym badaniu dysponujemy trzema grupami zmiennych objaśniających: (1) zmienne pochodzące z doładowań, (2) zmienne pochodzące z użycia oraz (3) zmienne pochodzące z taryf i pakietów oraz inne nieużyciowe. Poniżej przedstawiamy przykłady zmiennych. Zmienne pochodzące z doładowań: średnia liczba doładowań, średni nominał doładowania, największy nominał doładowania, liczba dni, jakie upłynęły od ostatniego doładowania. Zmienne pochodzące z użycia:

średnia liczba wykonanych połączeń głosowych, średnia liczba odebranych połączeń głosowych, średnia liczba wysłanych wiadomości SMS, średnia wielkość w kilobajtach danych pobranych w ciągu ostatniego miesiąca, średni czas trwania odebranych połączeń z numeru, z którym najczęściej się klient łączył w ciągu ostatniego miesiąca. Zmienne pochodzące z taryf i pakietów oraz inne zmienne nieużyciowe: taryfa, której obecnie używa klient (szczególnie można na jej podstawie odtworzyć stawkę za połączenia), zmienne zero-jedynkowe określające, czy klient aktywował określone pakiety, np. pakiet tańszych minut do wybranych klientów czy pakiet tańszych SMS-ów, ilość środków na koncie w momencie analizy, czas życia klienta rozumiany jako liczba dni, jakie upłynęły od momentu aktywacji, zmienna zero-jedynkowa określająca, czy klient zmieniał w przeszłości taryfę.

6. Uogólnione estymatory *maximum score*

W przypadku estymatorów *maximum score* (maksymalizacji wyniku) zakłada się następujący sposób generacji danych oparty na zmiennej ukrytej.

$$y^* = \beta_0 + \beta^T x + u$$

$$y = \begin{cases} 1, \text{ gdy } y^* \geq 0 \\ 0, \text{ gdy } y^* < 0 \end{cases}.$$

Obserwowane są (y, x). Zmienna y jest objaśniana, x to wektor zmiennych objaśniających, a u to składnik losowy. Uogólnione estymatory *maximum score* [Owczarczuk 2009] polegają na maksymalizacji średniej wartości zmiennej objaśnianej w pewnym podzbiorze o ustalonej mierze $\tau \in (0,1)$. Geometrycznie polegają na podziale próby, za pomocą płaszczyzny $b^T x = b_0$, na dwa podzbiory o mierze τ i $1-\tau$ w taki sposób, aby podzbiór o mierze τ miał maksymalną średnią zmiennej objaśnianej. Wektor b opisujący tę płaszczyznę daje zgodny wektor oszacowań wektora β z dokładnością do multiplikatywnej stałej [Owczarczuk 2009]. Poszukiwanie tej płaszczyzny rozdzielającej sprowadza się to do rozwiązania następującego zadania maksymalizacji

$$[\beta_N, \beta_{0N}] = \underset{[b,b_0]:[\![b,b_0]\!]=1}{\arg\max} \frac{1}{N}\sum_{i=1}^{N} y_i I(b^T x_i \geq b_0) - \mu \frac{1}{N}\sum_{i=1}^{N} (I(b^T x_i \geq b_0) - \tau)^2,$$

gdzie $\tau \in (0,1)$, a $\mu > 0$ to dostatecznie duża stała. W badaniu przyjęto $\tau = 0{,}05$, ale algorytm jest raczej niewrażliwy na dobór spośród małych wartości tego parametru. Metoda *maximum score* (maksymalizacji wyniku) jest złożona z oblicze-

niowego punktu widzenia. Polega na znalezieniu maksimum pewnej wielomodalnej funkcji wielu zmiennych. Dlatego przyjęto szereg uproszczeń obliczeniowych, które umożliwiają jej praktyczną implementację. Standardowo w przypadku optymalizacji funkcji wielomodalnych stosuje się strategię polegającą na wyborze wielu wektorów startowych, a następnie dla każdego z tych wektorów zastosowaniu pewnej iteracyjnej metody optymalizacji znajdującej maksimum lokalne. Innym popularnym podejściem jest wykorzystanie stochastycznych technik optymalizacji, które znajdują maksimum globalne z pewnym prawdopodobieństwem. W tym badaniu wykorzystano tylko jeden wektor startowy – wektor oszacowań pochodzący z regresji logistycznej. Uzasadnienie jest takie, że regresja logistyczna dostarcza dobrego pierwszego przybliżenia szukanych optymalnych współczynników, które następnie mogą zostać poprawione w wyniku iteracyjnej procedury optymalizacji. Jako technikę optymalizacji wykorzystano tutaj symulowane wyżarzanie (*simulated annealing*). Wykorzystano metodę selekcji zmiennych typu „w przód", tzn. modelowanie rozpoczęto od modelu tylko ze stałą, a następnie krokowo dodawano zmienne, których dodanie najbardziej podwyższało wartość optymalizowanej funkcji. Ustalono próg pół punktu procentowego jako minimalny przyrost kryterium optymalizacyjnego, który uzasadnia dodanie zmiennej. Procedura dodawania zmiennych kończyła się, gdy nie można już było dodać zmiennej tak, aby podwyższyć kryterium optymalizacyjne o co najmniej wartość progową.

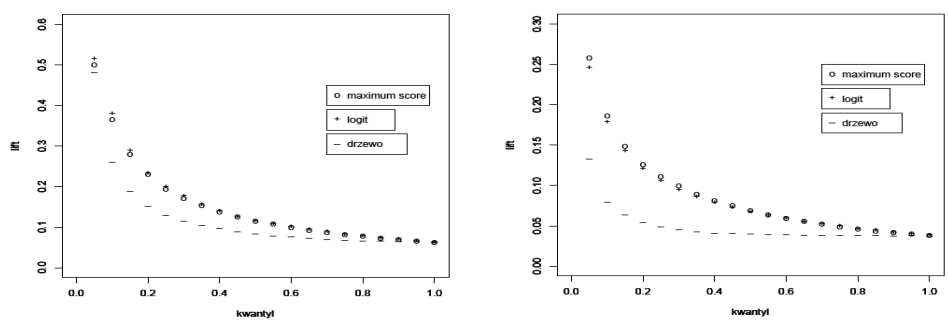
7. Opis i wyniki eksperymentu obliczeniowego

Celem tej części pracy jest zbadanie skuteczności estymatorów *maximum score* (maksymalizacji wyniku) do prognozowania zjawiska *churn* na przykładzie danych opisanych w poprzedniej części. Do porównań wykorzystano model drzew klasyfikacyjnych zbudowany metodą CART oraz regresję logistyczną z selekcją zmiennych typu „w przód" polegającą na krokowym dodawaniu zmiennych istotnych w sensie testu Walda (przy poziomie istotności $\alpha = 0,05$). Obliczenia wykonano w pakiecie R. Drzewa klasyfikacyjne były często wykorzystywanym narzędziem do tego celu w innych pracach na temat modelowania zjawiska odchodzenia klientów sieci komórkowej. Z kolei regresja logistyczna daje model o tej samej postaci funkcyjnej co *maximum score*. Zatem ciekawe może być porównanie skuteczności tych metod. Schemat eksperymentu był następujący: zbiór pochodzący z roku 2007 podzielono w stosunku 7:3 na część uczącą i testową. Modele szacowano na części uczącej. Część testowa oraz zbiór z roku 2008 służyły do pomiaru skuteczności. Zbiór testowy posłużył do zbadania, czy modele są skuteczne w krótkim okresie, tzn. ich wykorzystanie jest zasadne dla okresu, w którym zostały zbudowane. Zbiór z roku 2008, który został zebrany pół roku później niż zbiory uczący i testowy, posłużył do zbadania, czy modele są stabilne w czasie, a co za tym idzie, czy można je z powodzeniem wykorzystywać nawet długo po tym, jak zostały zbudowane. Jest to ważny aspekt organizacji akcji zatrzymaniowych, gdyż modele, które się

szybko starzeją, wymagają częstych i kosztownych aktualizacji. W zbiorze zebranym w roku 2007 było 5,76% klientów, którzy odeszli, a w zbiorze z roku 2008 było 3,75% takich klientów.

Ze względu na fakt, iż potencjalnych zmiennych objaśniających jest bardzo dużo, zastosowano wstępną selekcję cech. Wybrano 50 zmiennych, które charakteryzowały się największym modułem statystyki w teście *t*-Studenta na różnicę średnich w grupie klientów, którzy odeszli, oraz tych, którzy nie odeszli. Idea takiego postępowania jest następująca: te zmienne, które mogą potencjalnie dobrze odróżniać tych klientów, powinny mieć duże różnice w wartościach pomiędzy tymi dwiema grupami. Test *t*-Studenta w pewnym sensie mierzy istotność tych różnic (tutaj akurat w sensie średniej). Następnie szacowano modele już tylko z wykorzystaniem tych 50 zmiennych. Dodatkowo w celu dalszej redukcji liczby zmiennych zastosowano selekcję typu „w przód" dla *maximum score* (maksymalizacji wyniku) i regresji logistycznej. Drzewo klasyfikacyjne ma selekcję niejako wbudowaną w swój algorytm.

Rysunek 2 przedstawia podsumowanie precyzji poszczególnych modeli zebrane jako krzywe lift. Oś OX reprezentuje rząd kwantyla, czyli wielkość grupy docelowej wyrażoną jako frakcja całej populacji, a oś OY frakcję obserwacji z klasy $y=1$ w tej grupie docelowej. Im wyższe wartości funkcji lift dla danego rzędu kwantyla, tym lepiej.



Rys. 2. Krzywe lift dla zbioru z roku 2007 (lewy rysunek) i zbioru z roku 2008 (prawy rysunek)

Źródło: opracowanie własne.

Możemy zauważyć, że *maximum score* i regresja logistyczna osiągają znacznie lepsze wyniki i to zarówno dla zbioru testowego zebranego w tym samym okresie co zbiór uczący, jak i dla zbioru zebranego pół roku później, czyli w 2008 r. Przede wszystkim świadczy to o tym, że modele oszacowane metodą *maximum score* i metodą regresji logistycznej są stabilniejsze w czasie niż drzewa klasyfikacyjne. Można to łatwo wyjaśnić. Drzewa klasyfikacyjne wykorzystują reguły w postaci $x_i \le C_i$, gdzie x_i to zmienna objaśniana, a C_i to stała rzeczywista.

Rynek telekomunikacyjny zmienia się bardzo szybko, np. stawki za połączenia ulegają obniżeniu, co powoduje, że klienci wykonują coraz więcej połączeń. Powoduje to, że rozkłady zmiennych zmieniają się w czasie. W efekcie do niektórych liści drzewa dostaje się coraz więcej obserwacji, podczas gdy do innych dostaje się coraz mniej, co obniża precyzję prognoz. Reguły liniowe w postaci $a_1x_1 + ... + a_kx_k$ są bardziej odporne na tego rodzaju zmiany w danych. Nawet jeśli rozkłady zmiennych ulegną przesunięciu, to całe wyrażenie $a_1x_1 + ... + a_kx_k$ również się przesuwa i nie zmienia to znacznie posortowania obserwacji, a zatem i krzywej lift. Wykorzystanie regresji logistycznej generuje modele o znacznie większej liczbie zmiennych (29 zmiennych) niż modele szacowane metodą *maximum score* (4 zmienne). Wykorzystanie *maximum score* daje modele, które mają o wiele prostszą strukturę przy tej samej precyzji prognoz. W tym konkretnym przypadku w modelu znalazły się dwie zmienne określające sposób, w jaki dany klient się doładowuje, i dwie zmienne określające, jak często używa telefonu. Użycie regresji logistycznej i testu Walda do selekcji zmiennych prowadzi do zbyt dużej liczby zmiennych w modelu.

Literatura

Hung S.Y., Yen D.C., Wang H.Y., *Applying data mining to telecom churn management*, „Expert Systems with Applications" 2006, 31.

Koronacki J., Ćwik J., *Statystyczne systemy uczące się*, WNT, Warszawa 2005.

Owczarczuk M., *Maximum score type estimators*, „Central European Journal of Economic Modelling and Econometrics" 2009, 1.

Pendharkar P.C., *Genetic algorithm based neural network approaches for predicting churn in cellular wireless network services*, „Expert Systems with Applications" 2009, 36.

Wei C.P., Chiu I.T., *Turning telecommunications call details to churn prediction: a data mining approach*, „Expert Systems with Applications" 2002, 23.

PREDICTING CHURN IN THE MOBILE TELECOMMUNICATION SECTOR

Summary: The aim of this article is the analysis of the application of the maximum score estimators to the problem of churn prediction in the telecommunication sector. Our research shows that models built using maximum score give more precise and more stable models than classification trees and simpler than logistic regression. In our analysis we used data gathered from one of the Polish mobile operators.