METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH
Tom XIII/3, 2012, str. 65 – 78

PROBLEMY MODELOWANIA REZYGNACJI KLIENTÓW W TELEFONII KOMÓRKOWEJ

Krzysztof Gajowniczek, Tomasz Ząbkowski
Katedra Informatyki, Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
e-mail: krzysztof_gajowniczek@sggw.pl, tomasz_zabkowski@sggw.pl

Streszczenie: Przewidywanie i zarządzanie rezygnacjami klientów jest problemem wielu działalności gospodarczych, ale jest ono szczególnie dotkliwe w silnie konkurencyjnym sektorze telefonii komórkowej. Ze względu na wysokie koszty pozyskania nowych klientów i korzyści wynikające z utrzymania istniejących, istotną rolę w tego typu problemach odgrywają modele przewidujące rezygnację klientów. W tym kontekście autorzy artykułu, na podstawie danych empirycznych, zwracają uwagę na takie kwestie jak: (1) przygotowanie danych do analizy (2) problem doboru cech, (3) dobór odpowiednich technik modelowania wraz z oceną ich przydatności w kampaniach utrzymaniowych.

Słowa kluczowe: modelowanie rezygnacji, telefonia komórkowa, modele klasyfikacyjne

WPROWADZENIE

Działania prowadzone dotychczas przez operatorów telekomunikacyjnych, koncentrowały się, w głównej mierze, na dążeniu do zwiększenia udziału w rozwijającym się rynku przez przejmowanie klientów od konkurencji lub pozyskiwaniu zupełnie nowych klientów. Wraz ze wzrostem stopnia dojrzałości rynku przed przedsiębiorstwami telekomunikacyjnymi pojawiały się nowe wyzwania. Na rynkach dojrzałych podstawowym celem działania staje się utrzymanie istniejących klientów, zbudowanie z nimi trwałych relacji oraz pozyskanie ich lojalności i zaufania. W warunkach rosnącej konkurencji, stworzenie grupy stałych klientów wydaje się być podstawowym warunkiem poprawy zyskowności firmy i zachowania dotychczasowej pozycji rynkowej. Na obecnym etapie rozwoju polskich firm telekomunikacyjnych niezbędne staje się więc zachowanie właściwych proporcji między zdobywaniem nowych

a utrzymywaniem istniejących klientów. Zdobywanie nowych klientów jest niewątpliwie ważnym zadaniem. Istnieje jednak potrzeba przewidywania zmian zachodzących na rynku i planowania działań nakierowanych na utrzymanie wartościowych i lojalnych klientów.

Przewidywanie i zarządzanie rezygnacjami klientów jest problemem wielu branż, ale jest szczególnie dotkliwe w silnie konkurencyjnym i obecnie w znacznie zliberalizowanym sektorze telefonii komórkowej. Rezygnacja abonenta w branży telekomunikacyjnej odnosi się do przejścia klienta od jednego operatora telekomunikacyjnego do drugiego. Zjawisko takie w nomenklaturze branżowej określa się z języka angielskiego – churn. Wielu abonentów często rezygnuje z jednego usługodawcy na rzecz innego w poszukiwaniu lepszych stawek, usług lub korzyści wynikających z podpisania nowej umowy (np. otrzymywanie najnowszych aparatów). Ocenia się, że średni wskaźnik rezygnacji dla polskiego sektora telefonii komórkowej wynosi 2,2% miesięcznie, co oznacza blisko 26% utraconych klientów rocznie. Koszty marketingowe potrzebne do pozyskania nowego abonenta (SAC ang. subscriber acquisition cost) w amerykańskim sektorze telefonii komórkowej szacuje się na 300 do 600 dolarów [Wei i Chiu 2002], natomiast w przypadku polskiego sektora te koszty oscylują wokół 132 zł [Grupa Telekomunikacja Polska 2011]. Warto nadmienić, że koszt utrzymania istniejących abonentów jest blisko pięć razy niższy niż koszt ich pozyskania, oraz co istotne mają oni tendencję do generowania większych przychodów, ponieważ są mniej wrażliwi na cenę i często korzystają z dodatkowych usług.

Ze względu na wysokie koszty pozyskania nowych klientów i znacznie korzyści wynikające z utrzymania istniejących, istotną rolę odgrywają modele przewidujące rezygnację klientów. Dają one możliwość wcześniejszego wskazania klientów chcących zrezygnować z usług operatora, dzięki czemu realne staje się wdrożenie w odpowiednim momencie działań, mających na celu ograniczenie tego zjawiska, poprzez różnego rodzaju akcje utrzymaniowe i promocyjne.

W tym kontekście, autorzy na podstawie danych empirycznych [Neslin 2002], zwrócą uwagę na takie kwestie związane z modelowaniem zjawiska rezygnacji klientów jak: (1) problem doboru cech mogących determinować rezygnację klientów w branży telekomunikacyjnej, (2) problematykę odpowiedniego przygotowania danych do analizy, (3) dobór odpowiednich technik modelowania oraz budowę modeli wraz z oceną ich przydatności w kampaniach utrzymaniowych klientów operatora.

PROBLEM REZYGNACJI KLIENTÓW W LITERATURZE

W przypadku rynku usług, na którym znajduje się wielu klientów i wielu dostawców, zachodzi swoboda zawierania umów, w związku z czym, pojawia się możliwość migracji i nabywania usług gdzie indziej.

W praktyce, w usługach abonamentowych można wyróżnić dwa najczęściej stosowane typy rezygnacji [Lu 2002]. Pierwszym z nich jest rezygnacja

niezamierzona (ang. involuntary), wynikająca z przyczyn finansowych. W takim przypadku usługodawca rozwiązuje kontrakt z klientem zalegającym w płatnościach, w celu ograniczenia potencjalnych przyszłych start. W ramach rezygnacji finansowej, wyróżnia się również sytuację, w której konsument nie jest w stanie regulować dalszych zobowiązań pieniężnych [Burez i Van den Poel 2007], [Coussement i Van den Poel 2008]. Drugi typ rezygnacji związany jest z obniżoną satysfakcją z otrzymywanych usług oraz ze skutecznymi działaniami marketingowymi firm konkurencyjnych. Nad drugim typem rezygnacji, nazywanym również rezygnacją komercyjną, skupia się największą uwagę podczas badań nad tym zjawiskiem. Badania prowadzone nad rezygnacja finansową, są bardzo podobne do badań prowadzonych nad skoringiem kredytowym. W powyższej sytuacji, rezygnację finansową jest dużo łatwiej przewidzieć niż rezygnacje komercyjną, lecz jest znacznie trudniej jej przeciwdziałać [Burez i Van den Poel 2008].

Istnieją także alternatywne definicje zakładające, że rezygnacja klienta związana jest ze względnym obniżeniem poziomu usług, dzięki czemu, można przedstawić ją wykorzystując nachylenie krzywej obrazującej popyt klienta na dana usługę lub produkt. [Glady i in. 2009a, 2009b] zauważają, że powyższe terminy odejścia są mocno zależne od definicji i ceny pojedynczego produktu lub usługi, dlatego też proponują zdefiniować konsumenta zagrożonego rezygnacją jako tego, dla którego obniżeniu ulega poziom przyszłych wpływów CLV[1], a dokładniej rzecz ujmując, przez ujemne nachylenie krzywej CLV. [Hwang i in. 2004] idą o krok dalej sugerując, że model CLV powinien brać pod uwagę nie tylko potencjalną wartość przyszłych wpływów, lecz zarówno wartość przeszłych wpływów pieniężnych oraz prawdopodobieństwo rezygnacji z usług.

Poziom rezygnacji mierzony jest zazwyczaj jako odsetek bazy klientów, którzy opuścili danego operatora i podawany jest w odniesieniu do ustalonego okresu czasu, np. w skali miesięcznej lub rocznej. W segmencie telefonii komórkowej odsetek odejść, waha się w zależności od kraju i usługodawcy. W skali rocznej może on sięgać 25-40% w usługach abonamentowych oraz 70-80% w przypadku usług przedpłaconych [Kohs 2006].

Skala problemu odchodzenia klientów uzasadnia potrzebę jego precyzyjnej identyfikacji oraz działania z wyprzedzeniem. Badacze zjawiska podkreślają, że istotną rolę w całym procesie analizy odejść, odgrywa rodzaj danych i ich jakość. Wyróżnia się dane na temat klienta informujące o: demografii, finansach, czasie trwania kontraktu, dodatkowych usług czy wreszcie dane o połączeniach wychodzących i przychodzących [Ahn i in. 2006], [Kim i Yoon 2004], [Sulikowski 2008], [Van den Poel i Larivière 2004].

[1] CLV - (ang. Customer Lifetime Value), analiza mająca na celu ocenę aktualnej wartości przyszłych przepływów pieniężnych, związanych z relacjami z nabywcą, a więc ocenę wartości klienta w czasie.

Najczęściej w praktyce do modelowania rezygnacji klientów stosuje się techniki uczenia nadzorowanego jak np.: logit i probit [Ahn i in. 2006], [Burez i Van den Poel 2007], [Coussement i Van den Poel 2008], [Keramati i Ardabili 2011], [Madden i in. 1999], [Seo i in. 2008], które są rozwinięciem klasycznych metod regresyjnych, zaadaptowanych specjalnie na potrzeby klasyfikacji obiektów, do z góry ustalonej liczy klas. Z kolei [Huang i in. 2012], [Hung i in. 2006], [Karus i Dumas 2011], [Neslin i in. 2006], [Wei i Chiu 2002], stosowali drzewa decyzyjne, będące graficzną metodą wspomagania procesu decyzyjnego, stosowanego w teorii decyzji. Inni, m.in. [Daskalaki i in. 2006], [Huang i in. 2012], [Hung i in. 2006], [Karus i Dumas 2011], [Neslin i in. 2006], [Tsai i Lu 2009], [Waal i Toit 2008], wykorzystują sztuczne sieci neuronowe.

Do modelowania rezygnacji stosowano także metody grupowania modeli, takie jak random forest, bagging i boosting [Burez i Van den Poel 2007], [Burez i Van den Poel 2008], [Coussement i Van den Poel 2008]. Są to techniki, polegające na opracowaniu sekwencji prostych modeli, przy czym każdy kolejny model, przykłada większą wagę do tych obserwacji, które zostały błędnie zaklasyfikowane przez poprzednie modele.

Na koniec należy nadmienić, że skuteczność podawanych przez autorów modeli jest różna, zależy bowiem od przyjętych założeń, jakości posiadanych danych oraz definicji samej rezygnacji (inna dla klientów abonamentowych a inna dla klientów usług przedpłaconych).

OPIS DANYCH

Dane wykorzystane w pracy do modelowania rezygnacji stanowi zbiór „Cell2Cell: The Churn Game" [Neslin 2002], pozyskany z Centrum Zarządzania Relacjami z Klientami Uniwersytetu Duke'a, mieszczącego się w Karolinie Północnej w Stanach Zjednoczonych. Stanowią one reprezentatywny wycinek całej bazy danych, należącej do anonimowej firmy działającej w sektorze telefonii komórkowej w USA.

Dane zawierają 71047 obserwacji, gdzie każda obserwacja odpowiada indywidualnemu klientowi. Każdej obserwacji przypisanych jest 78 zmiennych, z czego 75 potencjalnych zmiennych objaśniających posłuży do późniejszej budowy modeli. Zestaw zmiennych objaśniających zawiera 34 zmienne ilościowe znajdujące się na skali ilorazowej, 38 zmienne jakościowe wyrażone liczbowo znajdujące się na skali dychotomicznej, 2 zmienne jakościowe wyrażone liczbowo znajdujące się na skali nominalnej oraz jedną zmienną jakościową opisaną za pomocą etykiet tekstowych. Wszystkie zmienne objaśniające pochodzą z tego samego okresu czasu, lecz modelowana binarna zmienna objaśniana (przyjmująca wartości 0 oraz 1) oznaczona jako „CHURN", została zaobserwowana w okresie od 31 do 60 dni, od momentu obserwacji pozostałych zmiennych. W zbiorze znajduje się dodatkowo zmienna „CALIBRAT", służąca do identyfikacji próby

uczącej i próby testowej, zawierających odpowiednio 40000 oraz 31047 obserwacji (por. Tab. 1).

Tablica 1. Liczebności próby uczącej i próby testowej

Dane Churn	Liczba obserwacji	Odsetek rezygnacji
Próba ucząca	40000	50%
Próba testowa	31047	1,96%

Źródło: opracowanie własne

Próba ucząca zawiera 20000 przypadków zaklasyfikowanych jako osoby rezygnujące (oznaczone jako 1) oraz 20000 przypadków zaklasyfikowanych jako osoby nierezygnujące (oznaczone jako 0). Wyżej opisany podział zbioru uczącego, w którym odsetek klasy wyróżnionej (osoby rezygnujące) jest taki sam jak odsetek klasy niewyróżnionej, jest mało realistyczny, natomiast prowadzi do ominięcia problemu niezrównoważenia klas [Napierała i Stefanowski 2011], prowadzącego do błędnej specyfikacji modelu oraz słabej generalizacji wiedzy.

W próbie testowej, na której będzie sprawdzana jakość zbudowanego modelu, znajduje się jedynie 1,96% osób rezygnujących. Z tak małym odsetkiem klasy wyróżnionej (oznaczony cyfrą 1), bardzo często można spotkać się w praktyce biznesowej podczas budowy modeli klasyfikacyjnych. Ponadto, wymieniona wartość odsetka, odpowiada rzeczywistej wartości wskaźnika rezygnacji w firmie, z której pochodzą dane.

PRZYGOTOWANIE DANYCH DO ANALIZY

Większość surowych dostępnych w bazach czy hurtowaniach danych jest w postaci danych transakcyjnych, które należy przygotować, aby móc je wykorzystać do analizy. Częstym problem w danych może być również niekompletność (braki danych). Baza danych może również zawierać obserwacje, które są przestarzałe lub zbędne, dane znajdujące się w nieodpowiednim formacie dla modeli czy wreszcie wartości, które są niezgodne z zasadami czy zdrowym rozsądkiem. Aby dane były przydatne do celów eksploracji danych muszą przejść przez wstępną analizę. Nadrzędnym celem staje się minimalizacja GIGO[2] czyli minimalizacja tzw. śmieciowych danych dostających się do późniejszej budowy modelu. Szacuje się [Larose 2006], że sam etap wstępnej obróbki danych może zajmować nawet 60% czasu poświęconego na cały proces eksploracji danych.

W przypadku dostępnego zbioru danych problem dotyczył jedynie braków danych. W bazie danych znajduje się 10 zmiennych, w których takowe braki występują. Tab. 2 przedstawia liczebność braków danych dla poszczególnych zmiennych. Zmienne te dotyczą połączeń głosowych oraz wieku klientów.

[2] GIGO (ang. garbage in – garbage out) – wprowadzisz błędne dane a uzyskasz błędne wyniki.

Tablica 2. Liczebność braków danych dla poszczególnych zmiennych

Cecha	Opis	Braki danych
REVENUE	Średni miesięczny przychód	216
MOU	Średnia liczba minut w miesiącu	216
RECCHRGE	Średnia wartość miesięcznych zobowiązań	216
DIRECTAS	Średnia liczba połączeń bezpośrednich w miesiącu	216
OVERAGE	Średnia liczba minut ponad plan w miesiącu	216
ROAM	Średnia liczba połączeń roamingowych	216
CHANGEM	Procentowa zmiana wykorzystania minut w danych okresach	502
CHANGER	Procentowa zmiana w przychodach w danych okresach	502
AGE1	Wiek pierwszego członka rodziny	1244
AGE2	Wiek kolejnego członka rodziny	1244

Źródło: opracowanie własne

W rekordach z brakami danych ze zbioru uczącego, znajduje się aż 70% klasy wyróżnionej (klientów rezygnujących), podczas gdy stosunek osób rezygnujących i nierezygnujących w całym zbiorze uczącym wynosi 50%. Usunięcie obserwacji z brakami danych w takiej sytuacji mogłoby doprowadzić do zaburzenia wzorców znajdujących się w danych. W związku z tym, w dalszej części zaproponowana została technika uzupełniania tychże braków danych.

Do najczęstszych sposobów radzenia sobie z brakującymi danymi należą: zastąpienie wartości brakującej pewną z góry ustaloną wartością stałą lub zastąpienie ich wartością średnią (dla zmiennych liczbowych) bądź dominantą (dla zmiennych jakościowych). Takie postępowanie nie zawsze jest właściwe, ponieważ w ten sposób mocno ingerujemy w rozkład rzeczywisty danej cechy. Panaceum w takiej sytuacji może okazać się proces generowania liczb losowych z zaobserwowanego rozkładu prawdopodobieństwa danej zmiennej.

Rozkład prawdopodobieństwa dla tej samej cechy w obrębie obydwu klas (rezygnujący vs lojalni klienci) może się diametralnie różnić, w związku z czym, przeprowadzono test do weryfikacji hipotezy o równości wartości średnich badanej cechy w dwóch klasach, który ma postać:

$$u = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{1}$$

gdzie $\bar{x}_1, \bar{x}_2$ są wartościami średnimi obydwu klas, s_1^2, s_2^2 są wariancjami badanych klas oraz n_1, n_2 są licznościami poszczególnych klas. Dla dużej próby statystyka u ma standaryzowany rozkład normalny $N(0,1)$. Wyniki przeprowadzonego testu, w którym zbiór danych na dwie rozłączne podpróby rozdzielała zmienna objaśniana „CHURN", zostały przedstawione w poniższej tabeli (Tab. 3).

Tablica 3. Wyniki testu na równość średnich w dwóch grupach

	Grupująca: CHURN, Grupa 1: 0, Grupa 2: 1			
Zmienna	Średnia 0	Średnia 1	u	p-value
REVENUE	59,2187	57,9548	3,44835	0,000564
MOU	543,2069	482,8304	13,76491	0,000000
RECCHRGE	47,7824	44,6525	15,82461	0,000000
DIRECTAS	0,9180	0,8351	4,55609	0,000005
OVERAGE	39,1729	42,3594	-3,99238	0,000065
ROAM	1,1506	1,3940	-3,23494	0,001217
CHANGEM	-5,3443	-24,4463	9,00406	0,000000
CHANGER	-1,4102	-0,7000	-2,20327	0,027579
AGE1	43,6244	43,4101	1,348094	0,089932
AGE2	44,3017	44,0885	1,692853	0,090786

Źródło: opracowanie własne

Na poziomie istotności $\alpha = 0{,}05$ we wszystkich przypadkach odrzucamy hipotezę zerową mówiącą o tym, że wartości średnie w obydwu klasach są sobie równe. Jedynie w przypadku zmiennych „AGE1" i „AGE2" brak jest podstaw do odrzucenia hipotezy zerowej mówiącej, że wartości średnie w obrębie obydwu klas nie różnią się od siebie.

Proces przygotowania danych do analizy zakończony został etapem generowania danych z zaobserwowanych wcześniej rozkładów i ich uzupełnieniem do zbioru danych. Generowanie liczb losowych w tych przypadkach zostało przeprowadzone za pomocą modułu „Metody symulacji" w pakiecie Statistica. Symulacja danych dla zmiennych „AGE1" i „AGE2", została wykonana za pomocą metody Monte Carlo [Brandt 1998], [Drew i Homem 2012], która jest stosowana do modelowania matematycznego złożonych procesów, tak by można było przewidzieć ich wyniki za pomocą podejścia analitycznego.

W przypadku zmiennych informujących o wskaźnikach na temat połączeń głosowych, należy uwzględnić fakt, że występuje pomiędzy nimi ścisła współzależność funkcyjna. Związek ten został zmierzony za pomocą współczynnika korelacji rang Spearmana. W świetle tych wyników do symulacji braków danych w tych zmiennych została wykorzystana metoda uwzględniająca zależności funkcyjne, gdyż w przeciwny razie uzyskane wyniki mogłyby stać ze sobą w sprzeczności. W tym celu do losowania liczb z zachowaniem współzależności funkcyjnej , została wykorzystana Metoda Imana-Conovera [Haas 1999], wykorzystująca fakt istnienia związku między zmiennymi, mierzonego za pomocą współczynnika korelacji rang Spearmana.

DOBÓR ZMIENNYCH DO MODELU

W modelach opracowywanych w ramach koncepcji data mining jednym z zadań jest dobór zmiennych objaśniających spośród listy kandydatów. Dla

przykładu, budowanie zmiennych może być procesem zautomatyzowanym, w efekcie czego powstaje duża liczba cech sięgająca tysięcy lub nawet setek tysięcy zmiennych. Niestety, wykorzystywane metody analityczne, w obliczu tak dużych wolumenów danych mogą stać się bezużyteczne.

Dobór zmiennych objaśniających za pomocą testu χ^2 Pearsona

W niniejszej pracy dobór zmiennych został oparty m.in. o nieparametryczną statystykę zgodności χ^2 Pearsona oraz wartość p-value dla każdej zmiennej niezależnej [Lu 2002]:

$$\chi_d^2 = \sum_1^k \frac{(n_i - np_i)^2}{np_i} \tag{2}$$

gdzie n_i są doświadczalnymi liczebnościami poszczególnych przedziałów, natomiast p_i są prawdopodobieństwami przynależności do tych przedziałów. W przypadku zmiennych ciągłych zakres wartości zmiennej dzielony jest na k przedziałów, natomiast zmienne jakościowe nie są przekształcane. Trzeba ponadto zaznaczyć, iż wyników tej metody należy traktować, jako pewne rozwiązane heurystyczne będące pomocnym do wyodrębnienia tych zmiennych, które można uwzględnić na etapie budowy modeli.

W kolejnym kroku, za pomocą analizy macierzy współczynników korelacji dokonano wyboru takich zmiennych objaśniających, które są silnie skorelowane ze zmienną objaśnianą i jednocześnie słabo skorelowane między sobą. Pozwala to ominąć często występujący w analizach problem współliniowości zmiennych objaśniających [Maddala 2006]. Wobec tego, ze zbioru danych objaśniających zostały wybrane wszystkie te pary zmiennych, dla których współczynnik korelacji wynosił powyżej 0.7, aby ostatecznie wyeliminować tę zmienną, która charakteryzowała się mniejszą wartością statystyki χ^2 obliczoną powyżej. Ostatecznie z całego zbioru danych pozostało 38 zmiennych objaśniających, w których 22 zmienne ma charakter ilościowy i 16 zmiennych ma charakter jakościowy.

Dobór zmiennych objaśniających za pomocą analizy składowych głównych

Na tym etapie wykorzystano również analizę składowych głównych (PCA), która jest techniką redukcji wymiaru zaliczaną do technik uczących się bez nadzoru. Celem analizy jest poszukiwanie takiego zbioru zmiennych, mniej licznego od zbioru zmiennych oryginalnych, dla którego można z pewnym, ale możliwie najmniejszym błędem, odtworzyć wartości zmiennych oryginalnych. Aby taka redukcja była możliwa, między oryginalnymi zmiennymi muszą zachodzić zależności statystyczne. Metoda ta przekształca oryginalne, skorelowane zmienne na nowe nieskorelowane zmienne, tzw. składowe główne, które

wyjaśniają w maksymalnym stopniu całkowitą wariancję z próby zmiennych pierwotnych.

Za pomocą tej analizy, do dalszego etapu budowy modeli wybrano 6 zmiennych, utworzonych poprzez poszczególne składowe główne, które wyjaśniały całkowitą zmienność w 97,38%.

Podejście eksperckie do doboru zmiennych objaśniających

Jako trzecią strategię doboru zmiennych objaśniających, zastosowano subiektywny ekspercki wybór zmiennych do modelu. W tym celu, spośród wszystkich zmiennych wybrano dziesięć najlepszych predyktorów ilościowych, charakteryzujących się największą wartością statystyki χ^2.

MODELOWANIE REZYGNACJI

Do modelowania rezygnacji klientów wykorzystanych został szereg technik. W szczególności, celem rozwiązania postawionego problemu klasyfikacyjnego zastosowano sztuczne sieci neuronowe, drzewa klasyfikacyjne, drzewa klasyfikacyjne ze wzmacnianiem, regresję logistyczną oraz analizę dyskryminacyjną.

Do oceny modeli wykorzystywane były dwie miary: wykres przyrostu oraz pole pod krzywą ROC.

Wykres przyrostu (ang. lift chart) jest graficznym sposobem podsumowania użyteczności modeli do przewidywania wartości zmiennej objaśnianej przyjmującej dwie wartości, natomiast w przypadku, gdy modelowana zmienna przyjmuje więcej wartości, można stworzyć wykres przyrostu i wykres zysku oddzielnie dla każdej z klas. Jak pokazuje [Larose 2006], przyrost można zdefiniować następująco:

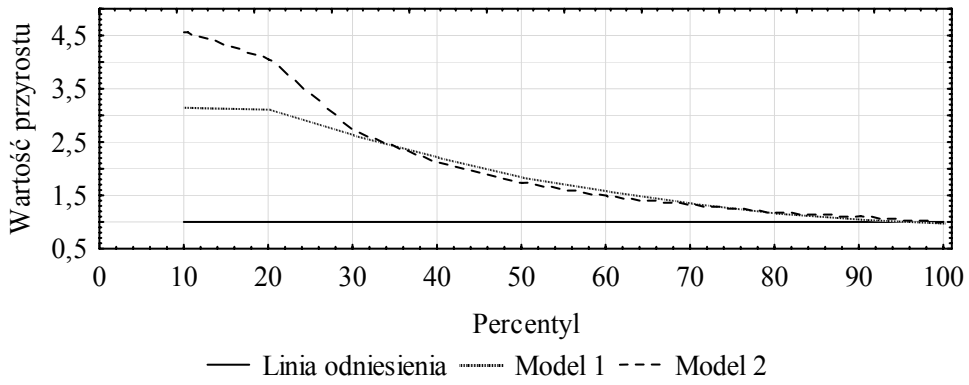
$$przyrost = \frac{procent\ \ pozytywnych\ trafień\ w\ zbiorze\ \ pozytywnych\ klasyfikacji}{procent\ \ pozytywnych\ trafień\ w\ całym\ zbiorze\ danych} \quad (3)$$

Podczas obliczania przyrostu najpierw sortuje się przypadki według prawdopodobieństwa przynależności do danej klasy. Następnie wartości przyrostu oblicza się dla kolejnych percentyli całej zbiorowości tzn. dla 10%, 20% itd. przypadków z największym przewidywanym prawdopodobieństwem przynależności do danej klasy. Uzyskane w ten sposób punkty łączy się linią, która zazwyczaj wolno spada, aż do osiągnięcia wartości 1,0 przy wyborze wszystkich przypadków. Dodatkowo na wykres nanosi się linię odniesienia odpowiadającą losowemu wyborowi 10%, 20% itd. przypadków ze zbioru danych.

Końcowa postać wykresu przedstawiona jest na Rys. 1, na którym widać, że największy przyrost osiągnięty jest dla najmniejszych percentyli. Z wykresu tego typu można odczytać, iż biorąc 10% przypadków (oś x) najpewniej zaklasyfikowanych do odpowiedniej klasy, otrzymamy próbkę, w której co najmniej 4,5 razy więcej przypadków (oś y) należy do wybranej klasy. Innymi

słowy, zastosowanie modelu zwiększa prawdopodobieństwo dotarcia do pożądanej klasy w stosunku do losowego wyboru obserwacji.

Rysunek 1. Graficzne porównanie wartości przyrostu wybranych modeli



Źródło: opracowanie własne

W niniejszej pracy, do oceny modeli klasyfikacyjnych wykorzystywana jest wartość przyrostu dla pierwszego percentyla bazy danych klientów. Jest to podyktowane względami ekonomicznymi, gdyż operator telekomunikacyjny nie kieruje kampanii utrzymaniowej do szerokiej bazy klientów, lecz skupia się na niewielkim odsetku np. 1 – 2% bazy klientów w skali miesiąca, o największym prawdopodobieństwie rezygnacji z usług. Przykładowo, wobec całkowitej liczby klientów u danego operatora sięgającej ok. 10 mln, 1% stanowi grupę ok. 100 tyś. klientów w skali miesiąca, którym zostanie przedstawiona oferta utrzymaniowa.

Jeśli chodzi o ogólną jakość klasyfikatora, konstruuje się miarę (ściśle związaną z indeksem Giniego), będącą polem powierzchni pomiędzy krzywą modelu a osią poziomą (AUC ang. Area Under Curve), którą można w przybliżeniu obliczyć poprzez całkowanie numeryczne metodą trapezów [Fortuna i in. 2002]:

$$AUC \approx \frac{2 - \sum_{i=1}^{n} \left[(Fx_{i+1} + Fx_i)(Lx_{i+1} - Lx_i) \right]}{2} \tag{4}$$

gdzie Fx_i są wartościami prostej utworzonej poprzez dystrybuantę linii odniesienia, natomiast Lx_i są wartościami krzywej utworzonej z wartości zysku obliczonego dla każdego percentyla. Ponadto $AUC \in \langle 0.5,1 \rangle$ oraz $indeks\ Giniego = 2AUC - 1$.

W wyniku przeprowadzonych eksperymentów numerycznych okazało się, że najlepszą metodą doboru zmiennych objaśniających do modelu opisującego

rezygnacje klientów, okazała się metodą χ^2 połączona z późniejszą eliminacją skorelowanych zmiennych w celu uniknięcia współliniowości. Modele charakteryzowały się nie tylko dużym przyrostem dla pierwszego percentyla, ale w szczególności znacznie większą jakością klasyfikacji na całym zbiorze testowym mierzoną polem pod krzywą ROC. Zachowanie jak największej ilości odpowiednich zmiennych pozwala zawrzeć w modelu nawet te najdrobniejsze wzorce i nie traci się cennych informacji zawartych w danych. Poza tym nie są one w żaden sposób przekształcone tak jak w przypadku metody składowych głównych, która okazała się metodą najgorszą.

Najlepszymi modelami w grupach zmiennych w $3/3$ przypadków (przez przypadki rozumiane są trzy różne metody selekcji zmiennych), okazały się algorytmy wzmacniania drzew klasyfikacyjnych, gdzie dla najlepszego modelu przyrost wyniósł 3,93 oraz pole $AUC_{Boosting} = 0{,}6600$. Na kolejnym miejscu uplasowały się modele drzew klasyfikacyjnych C&RT. Następnie, bardzo podobne wyniki uzyskały modele sieci neuronowych i modele dyskryminacyjne, natomiast najgorsze wyniki dawał model regresji logistycznej (por. Tab. 4).

Tablica 4. Zbiorcze wyniki dla opracowanych modeli

Lp.	Model (nazwa)	Przyrost w 1	Pole AUC
Dobór zmiennych za pomocą testu χ^2 i eliminacji skorelowanych zmiennych			
1	Sieci neuronowe (SANN)	2,29	0,6217
2	Drzewa klasyfikacyjne (C&RT)	3,11	0,6204
3	Wzmacniane drzewa klasyfikacyjne (Boosting)	3,93	0,6600
4	Regresja logistyczna (Logit)	2,13	0,6165
5	Analiza dyskryminacyjna (GDA)	2,78	0,6166
Dobór 10 zmiennych ilościowych za pomocą podejścia eksperckiego			
6	Sieci neuronowe (SANN)	2,62	0,6140
7	Drzewa klasyfikacyjne (C&RT)	2,78	0,6192
8	Wzmacniane drzewa klasyfikacyjne (Boosting)	3,93	0,6352
9	Regresja logistyczna (Logit)	1,96	0,5877
10	Analiza dyskryminacyjna (GDA)	2,13	0,5879
Dobór zmiennych za pomocą Analizy Składowych Głównych			
11	Sieci neuronowe (SANN)	2,45	0,6078
12	Drzewa klasyfikacyjne (C&RT)	1,63	0,5930
13	Wzmacniane drzewa klasyfikacyjne (Boosting)	3,44	0,6083
14	Regresja logistyczna (Logit)	1,63	0,5868
15	Analiza dyskryminacyjna (GDA)	1,63	0,5866

Źródło: opracowanie własne

Na podstawie przeprowadzonych analiz zaobserwowano, że na rezygnacje klienta najbardziej wpływają następujące zmienne: „EQPDAYS" – im dłuższy czas posiadania bieżącej taryfy tym większe ryzyko odejścia, „MONTHS" – im dłuższy czas pobytu w sieci tym większe prawdopodobieństwo rezygnacji. „RETCALS" – większa ilość połączeń do sekcji utrzymania klienta wpływa na rezygnację, „MOU" i "OVERAGE" – mniejsze średnie i całościowe miesięczne wykorzystywanie minut sprzyja rezygnacji oraz zmienna „UNIQSUBS" mówiąca o tym, że klient posiadający mniej unikalnych usług częściej rezygnuje.

PODSUMOWANIE

Celem pracy było przedstawienie możliwości modeli klasyfikacyjnych w problemie rezygnacji klientów telefonii komórkowej.

W tym kontekście autorzy artykułu, na podstawie danych empirycznych zwrócili uwagę na takie kwestie jak: (1) problem doboru cech mogących determinować rezygnację klientów w branży telekomunikacyjnej, (2) problematykę odpowiedniego przygotowania danych do analizy (3) dobór odpowiednich technik modelowania oraz budowę modeli wraz z oceną ich przydatności w kampaniach utrzymaniowych klientów operatora.

W szczególności badania wykazały, że:

(i) Najlepszą techniką doboru zmiennych skutkującą największą wartością przyrostu dla pierwszego percentyla oraz polem powierzchni AUC, była metoda oparta o statystykę χ^2 i eliminację skorelowanych zmiennych;

(ii) Algorytm wzmacniania drzew klasyfikacyjnych wykazywał największą trafność klasyfikacji (wyrażoną przyrostem i polem pod krzywą ROC) spośród testowanych metod;

(iii) Modele przewidujące rezygnację klientów, zbudowane za pomocą przedstawionych technik są dobrym narzędziem do osiągnięcia celów strategicznych firmy, nakierowanych na przeciwdziałanie zjawisku rezygnacji.

BIBLIOGRAFIA

Ahn J-H., Han S-P., Lee Y-S. (2006) Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry, Telecommunications policy 30, str. 552-568.

Brandt S. (1998) Analiza danych. Metody statystyczne i obliczeniowe, Wydawnictwo Naukowe PWN, Warszawa.

Burez J., Van den Poel D. (2007) CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services, Expert Systems with Applications 32(2), str. 277-288.

Burez J., Van den Poel D. (2008) Separating financial from commercial customer churn: A modeling step towards resolving the conflict between the sales and credit department, Expert Systems with Applications 35 str. 497-514.

Coussement K., Van den Poel D. (2008) Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques, Expert Systems with Applications 34(1), str. 313-327.

Daskalaki S., Kopanas I., Avouris N. (2006) Evaluation of classifiers for an uneven class distribution problem, Applied Artificial Intelligence Vol. 20 (5), str. 381-417.

Drew S., Homem T. (2012) Some Large Deviations Results for Latin Hypercube Sampling, Methodology and Computing in Applied Probability 14 (2), str. 203-232.

Fortuna Z., Macukow B., Wąski J. (2002) Metody numeryczne, Wydawnictwo Naukowo-Techniczne, Warszawa.

Glady N., Baesens B., Croux C. (2009a) Modeling churn using customer lifetime value, European Journal of Operational Research 197, str. 402-411.

Glady N., Baesens B., Croux C. (2009b) A modified Pareto/NBD approach for predicting customer lifetime value, Expert Systems with Applications 36, str. 2062-2071.

Grupa Telekomunikacja Polska. (2011) Sprawozdanie Zarządu z działalności Grupy Kapitałowej Telekomunikacja Polska w pierwszym półroczu 2011 roku,. str. 19-21, Pozyskano z : http://www.telix.pl/images/sprawozdania/TP_Grupa_2q2011.pdf.

Haas C. (1999) On Modeling Correlated Random Variables in Risk Assessment, Risk Analysis 19 (6), str. 1205-1214.

Huang B., Kechadi M., Buckley B. (2012) Customer churn prediction in telecommunications, Expert Systems with Applications 39, str. 1414-1425.

Hung S.Y., Yen D.C., Wang H.Y. (2006) Applying data mining to telecom churn management, Expert Systems with Applications 31, str. 515–524.

Hwang H., Jung T., Suh E. (2004) An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry, Expert System with Applications 26, str. 181-188.

Karus S., Dumas M. (2011) Predicting the maintainability of XSL transformations, Science of Computer Programming 76, str. 1161-1176.

Keramati A., Ardabili S. (2011) Churn analysis for an Iranian mobile operator, Telecommunications Policy 35, str. 344-356.

Kim H., Yoon C. (2004) Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market, Telecommunications Policy 28, str. 751–765.

Kohs G. (2006) Comparison of Churn Rates, Inside Market Research, June 2006, Pozyskano z : http://insidemr.blogspot.com/2006/06/comparison-of-churn-rates.html.

Larose D.T. (2006) Odkrywanie wiedzy z danych. Wprowadzenie do eksploracji danych, Wydawnictwo Naukowe PWN, Warszawa.

Lu J. (2002) Predicting customer churn in the telecommunications industry – An applications of survival analysis modeling using SAS, Proceedings of SUGI 27, Orlando, Florida, Paper 114.

Maddala G. S. (2006) Ekonometria, Wydawnictwo Naukowe PWN, Warszawa.

Madden G., Savage S., Coble-Neal G. (1999) Subscriber churn in the Australian ISP market, Information Economics and Policy 11, str. 195-207.

Napierała K., Stefanowski J. (2011) BRACID: a comprehensive approach to learning rules from imbalanced data, Journal of Intelligent Information Systems, Pozyskano z : http:// www.springerlink.com/content/d484131415313k17/fulltext.pdf .

Neslin S. (2002) Cell2Cell: The churn game. Cell2Cell Case Notes. Hanover, NH: Tuck School of Business, Dartmoth College.

Neslin S., Gupta S., Kamakura W., Lu J., Mason C. (2006) Defection detection: Measuring and understanding the predictive accuracy of customer churn models, Journal of Marketing Research 43(2), str. 204-211.

Seo D., Ranganathan C., Babad Y. (2008) Two-level model of customer retention in the US mobile telecommunications service market, Telecommunications Policy, Volume 32, Issue 3-4, str. 182-196.

Sulikowski P. (2008) Mobile Operator Customer Classification in Churn Analysis, Proceedings of the SAS Global Forum Conference, San Antonio, Texas, paper 344.

Tsai C-F., Lu Y-H. (2009) Customer churn prediction by hybrid neural networks, Expert Systems with Applications 36, str. 12547-12553.

Waal de D., Toit du J. (2008) Gaining Insight into Customer Churn Prediction using Generalized Additive Neural Networks, Proceedings of SATNAC - South Africa Telecommunication Networks and Applications.

Wei C., Chiu I-T. (2002) Turning telecommunications call details to churn prediction: a data mining approach, Expert Systems with Applications 23, str.103-112.

PROBLEMS OF CHURN MODELLING AT CELLULAR TELECOMMUNICATION

Abstract: Managing of customer churn is a serious problem at many businesses but is particularly important in the highly competitive and liberalized the cellular telecommunication sector. Due to the high costs of acquiring new customers and significant benefits of keeping existing ones, the predictive models for churn classification play an important role in this business. In this context, based on empirical data, the authors will tackle such issues as: (1) the problem of variable selection to determine the churn, (2) the problem of data preparation for the analysis; (3) selection of appropriate modeling techniques and the construction of models with their evaluation to support the retention campaigns.

Keywords: churn modelling, cellular telecommunication, classification models