Mariusz Łapczyński
Uniwersytet Ekonomiczny w Krakowie

MODELE HYBRYDOWE C&RT-LOGIT W ANALIZIE MIGRACJI KLIENTÓW

Wprowadzenie[1]

Celem artykułu jest zastosowanie modelu hybrydowego C&RT-logit w analizie migracji klientów sieci Orange. O analizie migracji klientów w literaturze z zakresu marketingu relacji wspomina się przy okazji zagadnienia utrzymania klienta (model ACURA, cykl życia klienta) oraz satysfakcji i lojalności nabywców. Wydaje się, że brak satysfakcji konsumentów w naturalny sposób wyjaśnia ich odejścia, jednak nie zawsze tak się dzieje. Wyróżnia się kilka możliwych reakcji niezadowolonych konsumentów[2]: a) odejście (*exit*); b) reklamację (*voice*); c) lojalność (*loyalty*) – klient wciąż pozostaje lojalny; d) współpracę (*collaboration*) – klient stara się rozwiązać problem, współpracując z przedsiębiorstwem. Niezadowolenie klienta jest poza tym tylko jedną z kilku możliwych przyczyn zakończenia współpracy. Migracje mogą być bowiem spowodowane[3]: a) nasyceniem – konsument znudzony dotychczasowym produktem poszukuje nowych rozwiązań; b) bardziej atrakcyjną ofertą konkurencji; c) konfliktem z dostawcą lub d) wysokimi barierami wyjścia.

1. Model hybrydowy cart-logit

Do budowy modeli migracji klientów wykorzystuje się najczęściej drzewa klasyfikacyjne, regresję logistyczną, losowy las, wzmacniane drzewa klasyfikacyjne lub metodę wektorów nośnych. W ostatnim czasie obserwuje się tzw. po-

[1] Praca jest częścią projektu sfinansowanego ze środków Narodowego Centrum Nauki przyznanych na podstawie decyzji nr DEC-2011/01/B/HS4/04758.

[2] E. Gummesson: Total relationship marketing. Third Edition. Butterworth-Heinemann, Oxford 2008, s. 105 i nast.

[3] J.N. Sheth, A. Parvatiyar: Relationship Marketing in Consumer Markets. Antecedents and Consequences. W: Handbook of Relationship Marketing. Eds. J.N. Sheth, A. Parvatiyar. Sage Publications, California 2000, s. 191.

dejście hybrydowe, w którym łączy się różne narzędzia analityczne, np. drzewa klasyfikacyjne z analizą skupisk, algorytmy genetyczne z sieciami neuronowymi albo drzewa klasyfikacyjne z regresją logistyczną.

Model hybrydowy CART-logit wykorzystany w niniejszych badaniach łączy algorytm drzew klasyfikacyjnych i regresyjnych CART z dwumianowymi modelami logitowymi. CART jest uznawany za najbardziej zaawansowany algorytm do budowy modeli drzew[4]. Zmienna zależna i zmienne niezależne mogą znajdować się na dowolnym poziomie pomiaru, a sama analiza nie wymaga spełnienia właściwie jakichkolwiek założeń[5]. Dwumianowe modele logitowe są natomiast popularnym narzędziem stosowanym w sytuacji, gdy badacz dysponuje dwuwariantową zmienną zależną i zbiorem dowolnych zmiennych niezależnych.

Cechy algorytmu CART wyróżniające go od modeli logitowych to:

- automatyczny wybór najlepszych predyktorów (budowa rankingu ważności zmiennych niezależnych);
- brak konieczności transformacji zmiennych (np. logarytmowania, pierwiastkowania);
- automatyczne odkrywanie efektów interakcji;
- odporność na obserwacje nietypowe;
- niewymagane zastępowanie braków danych;
- w trakcie budowy modelu wymagany jedynie umiarkowany nadzór ze strony badacza.

Algorytm CART rozpoznaje strukturę danych, jednak przy bardzo rozbudowanym modelu drzewa nie zapewnia czytelnej prezentacji modelu. Zdarza się ponadto, że drzewo o dużej liczbie liści przedstawia bardzo prostą zależność między zmiennymi.

Budowa modeli logitowych wymaga z kolei obecności doświadczonego badacza i najczęściej trwa znacznie dłużej niż budowa drzewa klasyfikacyjnego. Modele logitowe są wrażliwe na obserwacje nietypowe i wymagają imputacji braków danych (przypadki z brakami danych są usuwane z analizy). Dużą zaletą tego podejścia jest możliwość obliczenia unikatowego prawdopodobieństwa

---

[4] L. Breiman et al.: Classification and Regression Trees. Wadsworth International Group, Belmont, CA, 1984. O drzewach klasyfikacyjnych i regresyjnych w języku polskim pisali m.in.: E. Gatnar: Nieparametryczna metoda dyskryminacji i regresji. Wydawnictwo Naukowe PWN, Warszawa 2001 oraz M. Łapczyński: Drzewa klasyfikacyjne w badaniach rynkowych i marketingowych. UE w Krakowie, Kraków 2010.

[5] Czasami pojawiają się problemy natury technicznej. Zdarza się, że do analizy nie można włączyć jakościowych zmiennych niezależnych z dużą liczbą kategorii albo zbyt dużej liczby zmiennych niezależnych. Dotyczy to jednak zwykle analizowania dużych zbiorów obserwacji, gdzie liczba predyktorów przekracza 1000, a liczba kategorii zmiennych jakościowych jest wyższa od 100.

przynależności do klasy (kategorii zmiennej zależnej) dla każdego przypadku. Drzewa klasyfikacyjne dostarczają tylko tyle prawdopodobieństw, ile mają liści i odnoszą się one do wszystkich przypadków, które w nich się znajdują.

Za jedną z pierwszych prób łączenia drzew klasyfikacyjnych z modelami logitowymi można uznać podejście CHAID-logit[6]. Hybrydyzacja polegała na sekwencyjnym użyciu tych narzędzi analitycznych. Po wstępnej eksploracji zbioru obserwacji za pomocą algorytmu CHAID podzielono przypadki na rozłączne podzbiory, wykorzystując do tego celu węzły końcowe drzewa. W drugim etapie procedury dla każdego podzbioru budowano niezależne dwumianowe modele logitowe.

Inną koncepcję hybrydyzacji zaproponowano kilka lat później[7], łącząc algorytm CART z modelami logitowymi. Tym razem procedura również była dwuetapowa, jednak do zbioru zmiennych niezależnych w modelu logitowym wprowadzono dodatkową zmienną sztuczną, której kategorie informowały o przynależności przypadku do węzła końcowego drzewa klasyfikacyjnego CART. Drzewo powstało w oparciu o ten sam zbiór zmiennych objaśniających, a każdy liść uwzględniał interakcję między predyktorami. Autorzy uznali, że taki sposób hybrydyzacji jest bardziej skuteczny, gdyż podział zbioru obserwacji na podzbiory według pierwszej koncepcji wiąże się z nadmierną redukcją liczebności (przypadki są rozdzielane na liście) oraz utratą informacji (zdarza się, że zbiory zmiennych niezależnych są różne w różnych podzbiorach). Do zalet podejścia CART-logit zaliczyli ponadto wyższą trafność predykcji modelu hybrydowego, szybsze odkrywanie interakcji przez algorytm CART oraz najczęściej – brak konieczności zastępowania braków danych[8].

---

[6] W.E. Lindahl, C. Winship: A Logit Model with Interactions for Predicting Major Gift Donors. „Research in Higher Education" 1994, Vol. 35, No. 6, s. 729-743.

[7] D. Steinberg, N.S. Cardell: The hybrid CART-logit model in classification and data mining, www.salford-systems.com, 1998 [10.01.2008]. Nazwa hybrydy CART-logit została zastąpiona nazwą C&RT-logit, ponieważ nazwa CART jest znakiem towarowym firmy Salford Systems – producenta oprogramowania CART®. Większość obliczeń w tym artykule została przeprowadzona za pomocą programu STATISTICA 10.0 z wykorzystaniem algorytmu *classification and regression trees*, który w programach statystycznych innych producentów jest oznaczany skrótem C&RT lub CRT.

[8] Wyróżniono kilka sposobów postępowania z brakami danych w modelach hybrydowych CART-logit: 1) ignorowanie braków danych – uzasadniane małą wrażliwością algorytmu na ten problem; 2) przypisanie przypadkom z brakami danych prawdopodobieństw otrzymanych za pomocą algorytmu CART, a pozostałym przypadkom – prawdopodobieństw uzyskanych za pomocą modelu hybrydowego; 3) imputacja braków danych wszelkimi znanymi badaczowi sposobami; 4) wprowadzenie zmiennych sztucznych informujących o braku danych.

2. Budowa modelu hybrydowego C&RT-logit w analizie migracji klientów

2.1. Charakterystyka zbioru obserwacji

W badaniach wykorzystano zbiór obserwacji dotyczący migracji klientów sieci Orange, który był wykorzystany podczas zawodów analitycznych KDD Cup w 2009 roku[9]. Jego niewątpliwą zaletą jest zgodność danych ze stanem faktycznym (rozkłady zmiennych, braki danych, obserwacje nietypowe), natomiast wadą – sposób kodowania zmiennych w sposób uniemożliwiający merytoryczną interpretację modelu. Zbiór zawierał 50 tys. przypadków, dwuwariantową zmienną zależną (rezygnacja z usługi/brak rezygnacji z usługi) i 229 zmiennych niezależnych. Procent klientów migrujących wynosił 7,34%, co świadczy o dosyć silnie niezbilansowanej próbie. W pierwszym kroku zbiór został podzielony na próbę uczącą (80% całego zbioru obserwacji) i próbę testową (20%). Odsetek klientów migrujących w obu próbach przedstawiono w tabeli 1. Ze względu na problem niezrównoważonych klas w próbie uczącej zredukowano w sposób losowy liczbę klientów lojalnych w taki sposób, że procent klientów migrujących wynosił 20%. Struktura próby testowej pozostała niezmieniona. Na potrzeby modelu regresji logistycznej zmienne jakościowe zostały zastąpione zestawem zmiennych sztucznych. Zmienne niezależne, dla których odsetek braków danych przekraczał 10%, zostały usunięte z analizy. W pozostałych przypadkach braki danych zastąpiono, w zależności od poziomu pomiaru, wartością średniej arytmetycznej lub modalnej dla danej zmiennej. Osiem niezależnych zmiennych jakościowych miało bardzo dużą liczbę kategorii – przekraczającą nawet 4 tys. Ze względu na problemy techniczne (algorytm C&RT nie może analizować wielowariantowych zmiennych nominalnych z tak dużą liczbą wariantów) oraz problemy dotyczące organizacji zbioru obserwacji (tworzenie kilku tysięcy zmiennych sztucznych dla modelu regresji logistycznej) postanowiono wykorzystać te zmienne w analizie skupisk. Za pomocą uogólnionej metody analizy skupisk (EM) pogrupowano wszystkie przypadki na 12 skupień i do modelu wprowadzono dodatkową zmienną informującą o przynależności przypadku do skupienia. Predyktory wielowariantowe zostały wyłączone z modelu.

[9] Dane pobrano ze strony http://www.kddcup-orange.com/data.php.

Tabela 1

Liczebność i rozkład procentowy zmiennej zależnej w próbie uczącej i testowej

Zbiór obserwacji	Liczebność	Liczba i udział klientów migrujących w zmiennej zależnej	
		liczba	procent
Cały zbiór obserwacji	50 000	3672	7,34
Próba ucząca	39 913	2932	7,35
Zmodyfikowana próba ucząca (20%-80%)	14 666	2932	20,00
Próba testowa	10 087	740	7,34

2.2. Model C&RT

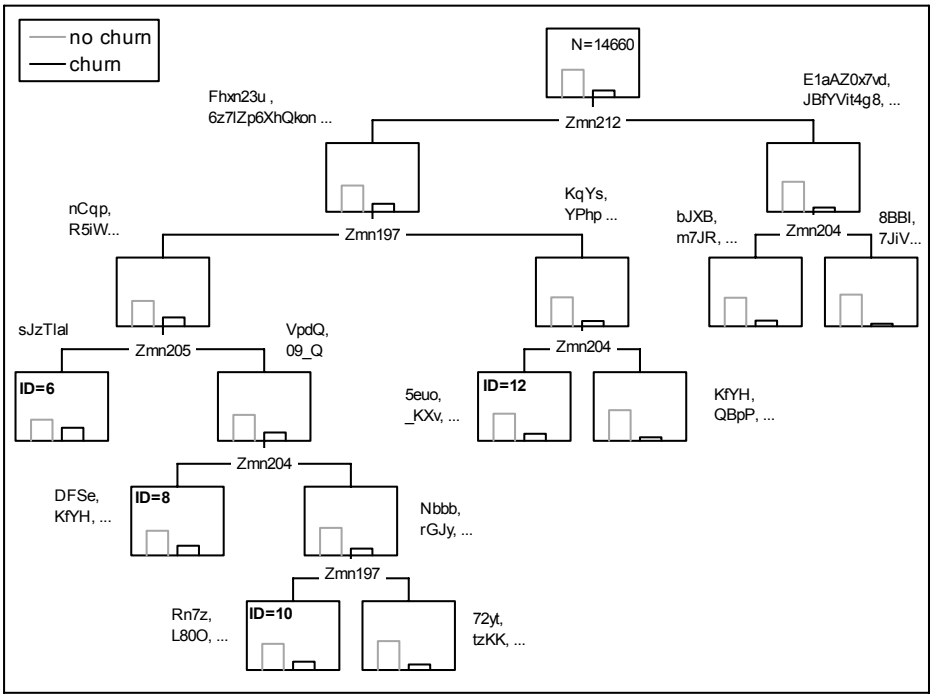
Ze względu na problem niezrównoważonych klas (tylko 20% obserwacji w próbie uczącej to klienci migrujący) oraz standardowe ustawienia algorytmu C&RT (równe koszty błędnych klasyfikacji i równe prawdopodobieństwa *a priori*[10]) w żadnym liściu drzewa klasyfikacyjnego (rys. 1) nie zaobserwowano przewagi kategorii „churn". Węzłami końcowymi, w których współczynnik przyrostu (*lift*) dla klientów nielojalnych przekraczał jedność, były te o numerach:

- liść ID 6 – udział klientów migrujących = 38,1%, współczynnik *lift* = 1,91;
- liść ID 8 – udział klientów migrujących = 28,1%, współczynnik *lift* = 1,40;
- liść ID 10 – udział klientów migrujących = 24,1%, współczynnik *lift* = 1,20;
- liść ID 12 – udział klientów migrujących = 20,1%, współczynnik *lift* = 1,03.

Z rankingu ważności predyktorów wynika ponadto, że zmiennymi niezależnymi o największej mocy dyskryminacyjnej są „z204" (100 punktów w rankingu), „z197" (84), „z73" (69) i „z205" (64)[11].

[10] Minimalną liczebność liścia ustalono na poziomie 5% zbioru uczącego (733 przypadki).

[11] Nazwy zmiennych uniemożliwiają identyfikację cech, jednak sporządzenie rankingu jest konieczne w celu porównania modeli prezentowanych w niniejszej pracy.



Rys. 1. Model drzewa klasyfikacyjnego C&RT

2.3. Model regresji logistycznej

Do modelu regresji logistycznej[12] trafiły 22 zmienne niezależne (19 jakościowych oznaczonych literą (j) i trzy ilościowe oznaczone literą (i)) istotnie powiązane ze zmienną zależną (szczegóły w tabeli 2). Wśród zmiennych niezależnych zwiększających prawdopodobieństwo migracji klientów znajdują się:

– z-195 taul (j) – prawdopodobieństwo migracji wśród klientów, dla których zmienna „z-195 taul" przyjmuje wartość 1, jest o 30% wyższe niż prawdopodobieństwo migracji wśród klientów, dla których jedynka występuje w kategorii odniesienia;
– z-197 ssAy (j) – prawdopodobieństwo odejścia wśród nabywców, dla których zmienna „z-197 ssAy" przyjmuje wartość 1, jest o 29% wyższe niż prawdopodobieństwo odejścia wśród nabywców, dla których jedynka występuje w kategorii bazowej;
– z-197 TyGl (j) – prawdopodobieństwo migracji w grupie nabywców, dla których zmienna „z-197 TyGl" przyjmuje wartość 1, jest o 16% wyższe niż

[12] Model regresji logistycznej będzie dalej nazywany modelem podstawowym lub modelem bazowym.

prawdopodobieństwo migracji w grupie nabywców, dla których jedynka występuje w kategorii odniesienia;
- z-211 L84s (j) – prawdopodobieństwo odejścia w grupie klientów, dla których zmienna „z-211 L84s" przyjmuje wartość 1, jest o 49% wyższe niż prawdopodobieństwo odejścia w grupie klientów, dla których jedynka występuje w kategorii bazowej;
- z-212 NhsEn4L (j) – prawdopodobieństwo migracji w grupie osób, dla których zmienna „z-212 NhsEn4L" przyjmuje wartość 1, jest o 44% wyższe niż prawdopodobieństwo migracji w grupie osób, dla których jedynka występuje w kategorii odniesienia.

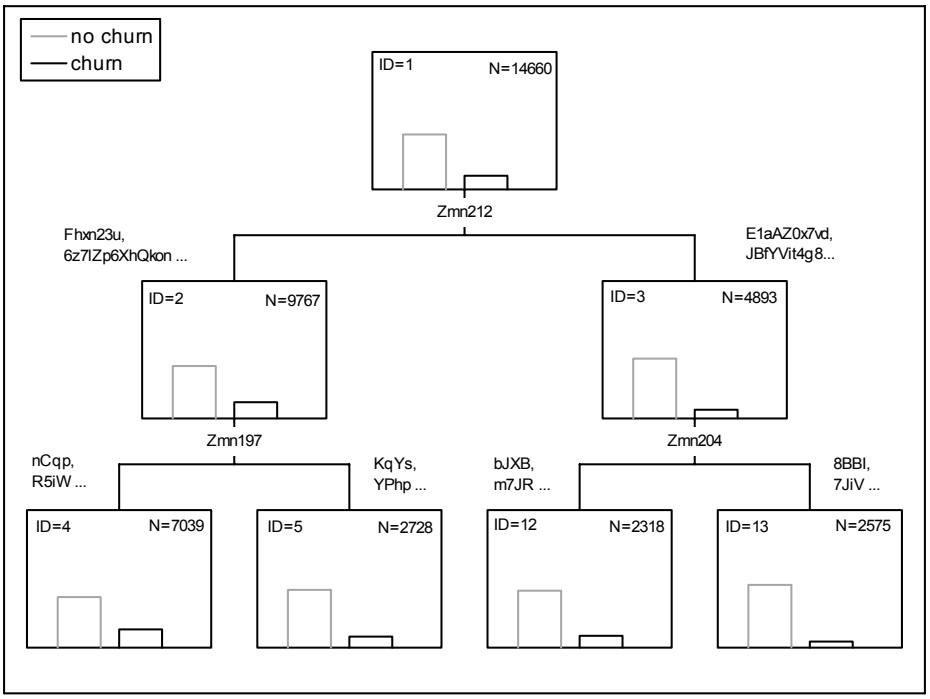
Tabela 2

Wyniki modelu regresji logistycznej

Zmienna	Ocena	Błąd standardowy	Poziom p	Iloraz szans
Wyraz wolny	0,140	0,311	0,653	1,15
z-57 (i)	-0,021	0,010	0,044	0,98
z-73 (i)	-0,006	0,001	0,000	0,99
z-113 (i)	0,000	0,000	0,000	1,00
z-195 taul (j)	0,262	0,128	0,040	1,30
z-197 ssAy (j)	0,255	0,103	0,013	1,29
z-197 TyGl (j)	0,145	0,074	0,050	1,16
z-205 VpdQ (j)	-0,553	0,063	0,000	0,58
z-205 09_Q (j)	-0,634	0,073	0,000	0,53
z-207 me75fM6ugJ (j)	-0,744	0,256	0,004	0,47
z-207 7M47J5GA0pTYIFxg5uy (j)	-0,541	0,260	0,037	0,58
z-210 uKAI (j)	-0,571	0,086	0,000	0,56
z-211 L84s (j)	0,398	0,061	0,000	1,49
z-212 NhsEn4L (j)	0,367	0,082	0,000	1,44
z-212 CrNX (j)	-0,908	0,235	0,000	0,40
z-218 cJvF (j)	-0,211	0,045	0,000	0,81
z-226 Qu4f (j)	-0,202	0,077	0,009	0,82
z-226 WqMG (j)	-0,162	0,078	0,037	0,85
z-226 szEZ (j)	-0,249	0,099	0,012	0,78
z-226 7P5s (j)	-0,343	0,110	0,002	0,71
z-227 02N6s8f (j)	-0,565	0,236	0,017	0,57
z-227 6fzt (j)	-0,525	0,263	0,046	0,59
SKUPISKO clus7 (j)	-0,181	0,082	0,028	0,83

2.4. Model hybrydowy C&RT-logit

Przed przystąpieniem do budowy modelu hybrydowego zredukowano wielkość drzewa klasyfikacyjnego C&RT do czterech węzłów końcowych (rys. 2). Jedynym liściem, dla którego współczynnik przyrostu jest wyższy od jedności, jest ten o numerze 4. Udział klientów migrujących jest w nim równy 26,6%, co daje wartość *lift* równą 1,33.



Rys. 2. Zredukowany model drzewa klasyfikacyjnego C&RT

Do modelu hybrydowego C&RT-logit wprowadzono dodatkową zmienną jakościową – „przynależność klienta do liścia drzewa". Zmienną zastąpiono zmiennymi sztucznymi z kategorią odniesienia – węzeł nr 13. W tabeli 3 zamieszczono wyniki dla modelu hybrydowego.

Tabela 3

Wyniki modelu hybrydowego C&RT-logit

Zmienna	Ocena	Błąd standardowy	Poziom p	Iloraz szans
Wyraz wolny	-0,762	0,158	0,000	0,47
z-57 (i)	-0,022	0,011	0,035	0,98
z-73 (i)	-0,006	0,001	0,000	0,99
z-113 (i)	0,000	0,000	0,000	1,00
z-197 lK27 (j)	-0,170	0,075	0,024	0,84
z-205 VpdQ (j)	-0,552	0,064	0,000	0,58
z-205 09_Q (j)	-0,635	0,073	0,000	0,53
z-210 uKAI (j)	-0,578	0,086	0,000	0,56
z-211 L84s (j)	0,407	0,061	0,000	1,50
z-212 NhsEn4L (j)	-0,455	0,139	0,001	0,63
z-212 CrNX (j)	-0,274	0,123	0,026	0,76
z-212 Ie_5MZs (j)	0,767	0,172	0,000	2,15
z-218 cJvF (j)	-0,178	0,047	0,000	0,84
z-226 FSa2 (j)	0,128	0,058	0,028	1,14
z-227 02N6s8f (j)	-0,316	0,132	0,016	0,73
z-227 6fzt (j)	-0,557	0,147	0,000	0,57
liść 4 (j)	1,350	0,142	0,000	3,86
liść 5 (j)	0,692	0,148	0,000	2,00
liść 12 (j)	0,740	0,090	0,000	2,10

W modelu hybrydowym C&RT-logit zmiennymi niezależnymi, które wyraźnie zwiększają prawdopodobieństwo migracji klientów, są:

– z-211 L84s (j) – prawdopodobieństwo odejścia w grupie klientów, dla których zmienna „z-211 L84s" przyjmuje wartość 1, jest o 50% wyższe niż prawdopodobieństwo odejścia w grupie klientów, dla których jedynka występuje w kategorii bazowej (w modelu podstawowym regresji logistycznej iloraz szans był równy 1,49);
– z-212 Ie_5MZs (j) – prawdopodobieństwo migracji w grupie nabywców, dla których zmienna „z-212 Ie_5MZs" przyjmuje wartość 1, jest o 16% wyższe niż prawdopodobieństwo migracji w grupie nabywców, dla których jedynka występuje w kategorii odniesienia;

- z-226 FSa2 (j) – prawdopodobieństwo odejścia wśród nabywców, dla których zmienna „z-226 FSa2" przyjmuje wartość 1, jest o 29% wyższe niż prawdopodobieństwo odejścia wśród nabywców, dla których jedynka występuje w kategorii bazowej;
- liść 4 (j) – prawdopodobieństwo migracji nabywców z węzła końcowego nr 4 jest 3,86 razy wyższe niż prawdopodobieństwo migracji nabywców z węzła 13[13];
- liść 5 (j) – prawdopodobieństwo odejścia klientów z węzła końcowego nr 5 jest 2-krotnie wyższe niż prawdopodobieństwo odejścia klientów z węzła 13;
- liść 12 (j) – prawdopodobieństwo migracji klientów z węzła końcowego nr 12 jest 2,1 razy wyższe niż prawdopodobieństwo migracji klientów z węzła 13.

Warto zwrócić uwagę, że w modelu hybrydowym:

1) pojawiły się wszystkie zmienne odnoszące się do węzłów końcowych drzewa C&RT;
2) pojawiły się dwie nowe zmienne niezależne zwiększające prawdopodobieństwo migracji (z-212 Ie_5MZs i z-226 FSa2);
3) nie ma zmiennych objaśniających, które w podstawowym modelu regresji logistycznej zwiększały prawdopodobieństwo odejścia (z-195 taul, z-197 ssAy, z-197 TyGl, z-212 NhsEn4L);
4) występują zmienne niezależne z wyraźnie wyższymi wartościami ilorazów szans w porównaniu do modelu bazowego regresji logistycznej.

Brak zmiennych niezależnych z podstawowego modelu regresji logistycznej jest spowodowany tym, że węzły końcowe drzewa są opisane za pomocą zmiennych o numerach 197, 204 i 212. Uwzględniają zatem efekty z modelu bazowego.

2.5. Porównanie trafności klasyfikacji zbudowanych modeli

Do oceny jakości modeli wykorzystano popularne miary, jak dokładność, czułość (określana w literaturze *data mining* terminem *recall*), precyzja, specyficzność, średnia G (*G-mean*) i miara F (*F-measure*). Trzy kolejne tabele (4-6) przedstawiają macierze błędnych klasyfikacji dla trzech modeli wdrożonych na próbie testowej.

Tabela 4

Macierz błędnych klasyfikacji dla modelu C&RT

Wartości obserwowane	Wartości przewidywane		Razem
	brak migracji (0)	migracja (1)	
Brak migracji (0)	4975	4372	9347
Migracja (1)	308	432	740
Razem	5283	4804	10 087

[13] Udział klientów migrujących w liściu ID13 (kategorii odniesienia) był równy ok. 8%.

Tabela 5

Macierz błędnych klasyfikacji dla modelu regresji logistycznej

Wartości obserwowane	Wartości przewidywane		Razem
	brak migracji (0)	migracja (1)	
Brak migracji (0)	342	9005	9347
Migracja (1)	8	732	740
Razem	350	9737	10 087

Tabela 6

Macierz błędnych klasyfikacji dla modelu hybrydowego C&RT-logit

Wartości obserwowane	Wartości przewidywane		Razem
	brak migracji (0)	migracja (1)	
Brak migracji (0)	729	8618	9347
Migracja (1)	23	717	740
Razem	752	9335	10 087

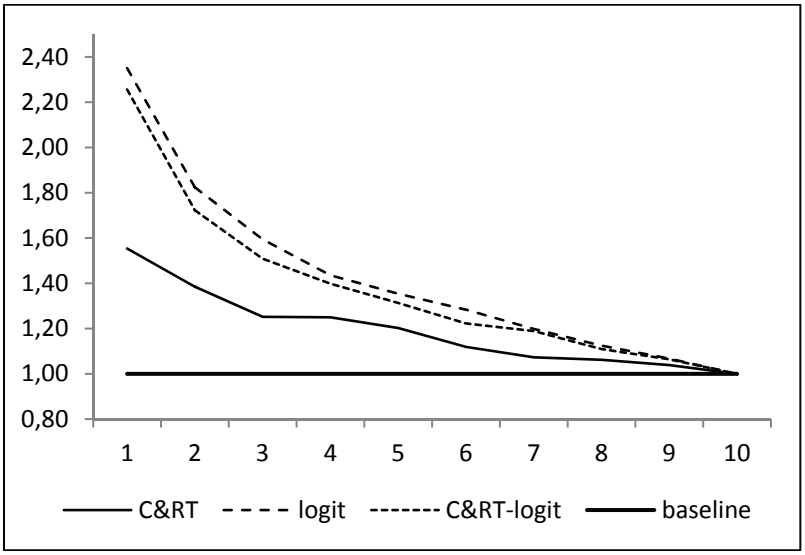
Na podstawie danych z macierzy błędnych klasyfikacji obliczono sześć miar jakości modeli predykcyjnych, które przedstawiono w tabeli 7. Najlepsze wyniki wyróżniono odcieniem szarości. Jak łatwo zauważyć, model C&RT charakteryzuje się, poza czułością, najwyższymi wartościami wykorzystanych współczynników. Model hybrydowy C&RT-logit jest, ogólnie rzecz ujmując, lepszy od podstawowej wersji modelu regresji logistycznej, ale gorszy od modelu drzewa klasyfikacyjnego. Jedyną miarą, która wskazuje na wyższość podstawowego modelu regresji logistycznej, jest czułość.

Tabela 7

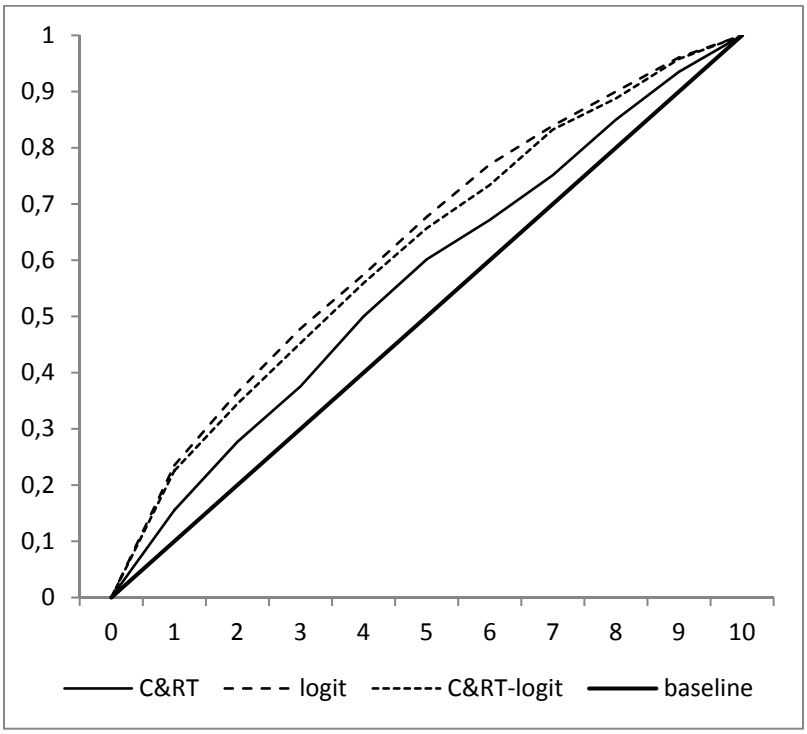
Miary jakości dla trzech modeli

Miara jakości	Model C&RT	Model regresji logistycznej	Model hybrydowy C&RT-logit
Dokładność	0,536	0,106	0,143
Czułość	0,584	0,989	0,969
Precyzja	0,090	0,075	0,077
Specyficzność	0,532	0,037	0,078
Średnia G	0,557	0,190	0,275
Miara F	0,156	0,140	0,142

Wartości miar jakości nie znajdują potwierdzenia na wykresie współczynnika *lift* (rys. 3) i na wykresie korzyści (rys. 4). Widoczna jest przewaga prostego modelu regresji logistycznej nad modelem drzewa klasyfikacyjnego i modelem hybrydowym. Należy jednak zwrócić uwagę, że z punktu widzenia przyrostu różnica pomiędzy modelami opartymi na regresji logistycznej nie jest duża. W trzecim, czwartym i piątym decylu skumulowany współczynnik *lift* dla modelu logistycznego wynosi odpowiednio: 1,59; 1,44 i 1,35, podczas gdy dla modelu hybrydowego wartości te wynoszą odpowiednio: 1,51; 1,40 i 1,31.



Rys. 3. Wykres dla skumulowanej wartości współczynnika *lift* (*lift chart*)



Rys. 4. Wykres korzyści (*gain chart*)

Podsumowanie

Wykorzystany w niniejszych badaniach model hybrydowy C&RT-logit pozwolił na połączenie cech algorytmu C&RT oraz cech regresji logistycznej. W tym konkretnym zbiorze obserwacji model hybrydowy dostarczył wyższych miar jakości dla tabel czteropolowych niż bazowy model logistyczny oraz wyższych miar przyrostu (*lift*) i korzyści (*gain*) niż model drzewa klasyfikacyjnego. O przewadze modelu hybrydowego nad modelem bazowym świadczą również wyższe miary współczynników pseudo R^2 (Coxa i Snella oraz Negelkerke'a). Efekty interakcji, które w automatyczny sposób zostały wskazane przez drzewo C&RT, trafiły do modelu hybrydowego, wzbogacając jego interpretację i zwiększając ilorazy szans dla zmiennych niezależnych.

Literatura

Breiman L., Friedman J.H., Olshen R.A., Stone C.J.: Classification and Regression Trees. Wadsworth International Group, Belmont, CA, 1984.

Gatnar E.: Nieparametryczna metoda dyskryminacji i regresji. Wydawnictwo Naukowe PWN, Warszawa 2001.

Gummesson E.: Total relationship marketing. Third Edition. Butterworth-Heinemann, Oxford 2008.

Łapczyński M.: Drzewa klasyfikacyjne w badaniach rynkowych i marketingowych. UE w Krakowie, Kraków 2010.

Lindahl W.E., Winship C.: A Logit Model with Interactions for Predicting Major Gift Donors. „Research in Higher Education" 1994, Vol. 35, No. 6.

Sheth J.N., Parvatiyar A.: Relationship Marketing in Consumer Markets. Antecedents and Consequences. W: Handbook of Relationship Marketing. Eds. J.N. Sheth, A. Parvatiyar. Sage Publications, California 2000.

Steinberg D., Cardell N.S.: The hybrid CART-logit model in classification and data mining, www.salford-systems.com, 1998 [10.01.2008].

THE USE OF HYBRID C&RT-LOGIT MODEL IN CHURN ANALYSIS

Summary

The purpose of this article is to use the hybrid C&RT-logit model in churn analysis in Orange telecom company. The combination of decision tree (C&RT algorithm) with logistic model improved the accuracy of prediction and enriched the interpretation of model. The dataset used during the study comes from KKD Cup in 2009.