Małgorzata Rószkiewicz

Kolegium Analiz Ekonomicznych
Szkoła Główna Handlowa w Warszawie

Metody predykcji w analitycznym *Consumer Relation Management* na potrzeby marketingu relacji

1. Wstęp

We współczesnym zarządzaniu przedsiębiorstwem kładzie się nacisk na kreowanie wartości przedsiębiorstwa dla jego akcjonariuszy. Oznacza to konieczność prowadzenia zarządzania zorientowanego na wzrost wartości rynkowej firmy oraz oceny rezultatów jej działalności. Zarządzanie to wymaga oceny kosztów i korzyści alternatywnych działań oraz rozpoznania optymalnej alokacji zasobów przeznaczonych na działania marketingowe według grup klientów[1]. Oznacza to wymóg prowadzenia analizy otoczenia rynkowego, identyfikację celów firmy i jej polityk dziedzinowych, a także potrzebę korzystania z metod szacowania niepewności, gromadzenia informacji i ich przetwarzania. Analizy te nie mogą obyć się bez modeli warunków działania firmy. Modele instrumentów jej oddziaływania na rynek pozwalają bowiem w warunkach zmiennego środowiska rynkowego przewidywać wyniki planowanych działań[2].

Utrzymanie wysokiej stopy zwrotu z kapitału oraz wysokiej stopy wzrostu jest możliwe jedynie przy ukształtowanym i przewidywalnym przychodzie ze sprzedaży. Cechami charakterystycznymi współczesnych rynków jest w większości przypadków ich nasycenie, relatywnie wysokie koszty utraty klienta oraz kształtowanie się kosztów utrzymania klientów lojalnych poniżej kosztów pozyskania klientów nowych. Dlatego też przewidywalność dobrych wyników

---

[1] P. Berger, R. Bolton, D. Bosman, E. Briggs, V. Kumar, A. Parasuraman, C. Terry, *Marketing Actions and Value of costumer Assets: A Framework for Costomer Asset Management*, „Journal of Service Research" 2002, no. 5, s. 39–54.

[2] E.W.T. Nagi, L. Xiu, D.C.K. Chau, *Application of Data Mining Techniques in Customer Relationship Management*, „A Literature Review and Classification. Expert System with Application" 2009, no. 36, s. 2592–2602.

ekonomicznych mogą zapewnić jedynie ci klienci, którzy są zadowoleni z oferty firmy i stali się klientami lojalnymi. Klienci tacy, w większym stopniu neutralni wobec zmiany cen i odporni na działania konkurencji, stają się stabilnym i przewidywalnym generatorem przychodów, co pozwala określać ich strumień z dużą pewnością. Jeśli zatem firma dąży do osiągnięcia przewagi konkurencyjnej na rynku, to musi dysponować zasobem, jaki stanowią zadowoleni i lojalni klienci[3]. Opcja ta, przez General Electric określona jako *Customer Focus*[4], zaś przez H. Nooriego i R. Radforda – jako *Customer Driven Strategy*[5], ogniskuje wszelkie działania firmy na tworzeniu wartości klienta. Celem prowadzącym do sukcesu rynkowego w długim okresie[6] staje się zatem orientacja na satysfakcję i zadowolenie klienta.

Dostosowanie wyrobów i usług do potrzeb oraz oczekiwań klientów staje się wyznacznikiem jakości prowadzonej działalności i – co podkreślają R.D. Buzzell i B.T. Gale[7] – tylko tak rozumiana jakość produktów oraz usług stymuluje stopę przychodów z inwestycji. Warto ponadto zauważyć, że ramy zakreślone dla gospodarki opartej na wiedzy obejmują, obok innowacji, jakości, zarządzania, aliansów, technologii, marki, pracowników i środowiska, również relacje z klientem[8].

Klient, będący wartością dla przedsiębiorstwa, staje się obiektem zarządzania. Kluczowe są więc działania skierowane na: identyfikację klientów z uwzględnieniem różnic w ich wartości, interakcje z klientem umożliwiające rozpoznanie jego celów konsumpcyjnych i wzorców zachowań rynkowych, a także modyfikowanie oferty firmy względem zmian oczekiwań klienta.

Działania dotyczące klienta zmieniają politykę marketingową przedsiębiorstwa, prowadząc do przejścia od marketingu transakcyjnego, który był skoncentrowany na wynikach sprzedaży, do marketingu relacyjnego, opartego na budowaniu więzi z klientem. Taki zakres działań wymaga poszukiwania wiedzy na podstawie danych o klientach i budowy modeli predykcyjnych, przewidujących efekty polityki sprzedażowej w indywidualnych przypadkach.

---

[3] J. Otto, *Marketing relacji. Koncepcja i stosowanie*, Wydawnictwo C.H. Beck, Warszawa 2004.

[4] http://www.ge.com/en/company/companyinfo/at_a_glance/key_growth_initiatives.htm.

[5] H. Noori, R. Radford, *Production and Operation Management. Total Quality and Responsiveness*, McGraw-Hill, New York 1995, s. 49–55.

[6] Ibidem, s. 65.

[7] R.D. Buzzell, B.T. Gale, *The PIMS Principles: Linking Strategy to performance*, Free Press, New York 1987.

[8] http://www.valuebasedmanagement.net/methods_valuecreationindex.html.

2. Istota modeli predykcyjnych w analitycznym CRM

Immanentną cechą marketingu relacji stała się analiza danych dotyczących klientów, co wymagało zintegrowania działań analitycznych z zarządczymi. Na potrzeby tego typu działań powstały systemy wspierające zarządzanie relacjami z klientem (*Consumer Relation Management* – CRM). Systemy te obejmują wielowątkowe działania, wśród których autorzy[9] wyróżniają takie zakresy, jak: operacyjny, koncentrujący się na opisie kontaktów między klientem a przedsiębiorstwem; współpracy, dotyczący problemów, które ujawniały się w ramach takich kontaktów; analityczny, odnoszący się do procesu gromadzenia i analizy danych. W ramach analitycznego zakresu CRM są budowane modele oceniające skuteczność działań rynkowych dotyczących klienta, zwane modelami predykcyjnymi.

W konstrukcji modeli predykcyjnych największy nacisk kładzie się na poprawność generowanych przewidywań, a ocena skuteczności predykcji staje się wręcz wykładnią oceny modelu. Można określić kilka podstawowych wyznaczników poprawności predykcji na podstawie modelu zbudowanego według wybranej metody. W pierwszej kolejności takim wyznacznikiem jest postać modelu, która jest zależna od założeń dotyczących relacji między zmiennymi. W drugiej kolejności model jest oceniany ze względu na stopień dopasowania do danych. W wielu przypadkach kryterium to rozstrzyga o wyborze ostatecznego rozwiązania spośród kilku wariantów modelu predykcyjnego. W dalszej kolejności ocenie podlega możliwość prowadzenia uogólnień, tak by własności opisane przez model można było uznać za reprezentację własności zachodzących w populacji. Warunek ten nie zawsze jest spełniony. Metody oparte na iteracyjnym dopasowaniu do danych często prowadzą do budowania modeli w znacznym stopniu dopasowanych, ale o własnościach, które nie zawsze są potwierdzane przez inne próby pobrane z tej samej populacji. Ostatnim składnikiem oceny modelu są jego walory eksplikacyjne. Chodzi tu o uzyskiwanie wiedzy o relacjach opisywanych przez model i ich interpretacji w języku dziedziny.

Z formalnego punktu widzenia przewidywanie wyników rynkowych rozważanych strategii biznesowych ma na celu optymalizację grup docelowych dla tych strategii. Wymaga to prognozowania zachowań klientów ze względu na przyjęte warunki. Zachowania jako takie sprowadzają się do wyboru jednego z kilku

[9] E. Gummesson, *Total relationship marketing*, Butterworth-Heinemann, Oxford 2008; M. Łapczyński, *Hybrydowe modele predykcyjne w marketingu relacji*, Wydawnictwa Uniwersytetu Ekonomicznego w Krakowie, Kraków 2015.

dostępnych wariantów decyzyjnych. Zgodnie z teorią racjonalnego wyboru jest dokonywana ocena krańcowych kosztów i krańcowych korzyści na podstawie oczekiwanych użyteczności wariantu decyzyjnego. Wynik tej kalkulacji jest operacjonalizowany w postaci zmiennej *Y*, interpretowanej jako miara użyteczności wariantu decyzyjnego[10]. O ile zmienna ta nie jest bezpośrednio obserwowana, o tyle są obserwowane zależne od jej poziomu wybory dokonywane przez klientów. Z tych powodów w modelowaniu jej wartości nie może być stosowane tradycyjne podejście, oparte na modelu regresji liniowej.

Ogólnie zmienne opisujące zachowania klientów mogą być zamiennymi binarnymi, wielomianowymi, licznikowymi lub ograniczonymi. W pierwszych trzech przypadkach są traktowane jako zmienne jakościowe[11]. Modele tego typu zmiennych noszą nazwę modeli wyboru dyskretnego[12]. Zmiennymi objaśniającymi mogą być wszystkie dostępne charakterystyki rynku, leżące zarówno po stronie klienta (jego cechy deskryptywne oraz behawioralne), jak i po stronie przedsiębiorstwa (charakter dotychczas odebranych działań marketingowych przez klienta), operacyjne i strategiczne.

W kontekście optymalizacji grup docelowych dla działań marketingowych kluczowa jest przynależność jednostki do grupy, rozumianej jako podzbiór zbioru wszystkich jednostek o charakterystycznej wartości zmiennej *Y*. W algorytmach identyfikujących tę przynależność można wyróżnić podejście parametryczne, semiparametryczne i nieparametryczne. Każde z nich przez swe własności wyznacza zarówno korzyści, jak i ograniczenia analityczne. Podejście parametryczne jest głęboko zakorzenione w metodzie indukcyjnej, w której falsyfikacja pozwala określić ramy poznania naukowego. Z kolei podejście nieparametryczne koncentruje się na eksploracji danych i pozwala poszukiwać rozwiązań o wysokich walorach predykcyjnych.

Podejście parametryczne polega na estymacji parametrów modelu regresyjnego, w którym zmienna zależna określa przynależność do grupy docelowej. Ogólnie model grupy docelowej można zapisać jako:

$$P(Y) = f(X_1, X_2, \ldots, X_k) + \xi, \tag{1}$$

[10] W.H. Greene, *Econometric Analysis*, Prentice Hall International Inc., Upper Saddle River 2000, s. 505–513.

[11] M. Gruszczyński, *Modele i prognozy zmiennych jakościowych w finansach i bankowości*, Oficyna Wydawnicza SGH, Warszawa 2002, s. 11–50.

[12] G. Maddala, *Intruduction to Economwetrics*, John Wiley & Sons Ltd, Chichester 2001, s. 317–338.

czyli jako model regresji zmiennych objaśniających $X_1, X_2,\ldots, X_k$. Co do funkcji f wymaga się, by była rosnąca. Parametry modeli nieliniowych są szacowane metodą największej wiarygodności. W modelach tego typu zmienna objaśniana Y może być różnie definiowana, co ma wpływ na szacowaną postać modelu. W zależności od przyjętych założeń model ten przyjmuje postać modelu dwumianowego, logitowego lub probitowego bądź też modelu tobitowego, bądź regresji uciętej. W modelach tych obowiązują ściśle określone założenia dotyczące rozkładu składnika losowego. Ponadto w modelu tobitowym jest wymagana ściśle określona struktura danych empirycznych. Z kolei modele logitowe i probitowe są dostosowane jedynie do klasy zjawisk liniowo separowanych, a w regresji uciętej zgubienie faktu ucięcia próby prowadzi do uzyskania wyników obciążonych i niezgodnych.

W podejściu semiparametrycznym dokonuje się poluzowania restrykcyjnych założeń podejścia parametrycznego, co upraszcza procedury estymacyjne i poprawia własności estymatorów dla parametrów szacowanych modeli. Utrzymywane jest założenie o postaci funkcyjnej modelu, lecz nie zakłada się konkretnej postaci rozkładu składnika losowego. Założenia mogą dotyczyć jedynie pewnych cech tego rozkładu, np. wartości niektórych parametrów pozycyjnych (wartość mediany = 0) lub kształtu rozkładu (rozkład symetryczny)[13]. Najbardziej elastyczne wydaje się podejście nieparametryczne, w którym przedmiotem oceny są jedynie wartości zmiennej zależnej, zaś budowany model zależności jest wolny od parametrów i wiążących założeń. Do grupy tego typu metod należą procedury dataminingowe, a wśród nich sieci neuronowe, programowanie genetyczne, uogólnione modele liniowe, a także metody klasyfikacji, a wśród nich metody rekurencyjnego podziału, zwane procedurami drzew klasyfikacyjnymi. W procedurach grupowania dokonuje się oceny podobieństwa obiektów ze względu na wielowymiarową informację o każdym z nich i utworzenia optymalnego podziału całej badanej zbiorowości na podgrupy o wysokim stopniu homogeniczności. W modelach tych dopuszcza się również brak rozróżnienia na zmienne zależne i niezależne, łącząc je w grupę deskryptorów procedury klasyfikacyjnej. W klasycznych podejściach taksonomicznych na ogół[14] wymaga się, by w procedurach oceny podobieństwa posługiwać się deskryptorami

[13] C.F. Manski, *Maximum score estimation of the stochastic utility model of choice*, „Journal of Economietrics" 1975, no. 3, s. 205–228; M. Owczarczuk, *Estymatory typu Maximum score dla wybranych modeli mikroekonometrycznych*, praca doktorska, Kolegium Analiz Ekonomicznych, SGH, Warszawa 2009, s. 31–35.

[14] M. Walesiak, *Metody analizy danych marketingowych*, Wydawnictwo Naukowe PWN, Warszawa 1996, s. 105–152.

jednakowego typu pod względem zasad skalowania. Jeśli w zbiorze deskryptorów występują zarówno zmienne ilościowe, jak i zmienne skalowane na skalach słabych (nominalne, porządkowe), to wówczas konieczna jest dyskretyzacja zmiennych ciągłych, co skutkuje utratą informacji. W nowoczesnych procedurach dataminingowych nie jest to wymagane. Dla przykładu, procedura dwustopniowego grupowania pozwala wykorzystywać zmienne różnie skalowane, jednakże rygorystycznie wymaga się względem nich spełnienia założeń o: niezależności wszystkich zmiennych, normalności rozkładu zmiennych ciągłych oraz porządku losowym bazy danych.

Najbardziej liberalne pod względem wymagań dotyczących danych i definiowanych za ich pomocą zmiennych wydają się modele predykcyjne budowane na podstawie metod rekurencyjnego podziału, tzw. procedur drzew klasyfikacyjnych. Podejście to ma charakter nieparametryczny, w tym sensie, że nie jest wymagana znajomość rozkładów analizowanych zmiennych. Można wskazać jeszcze inne jego zalety. Należy do nich w pierwszej kolejności brak wymagań co do znajomości klasy funkcji opisujących zależność między zmiennymi objaśniającymi i zmienną objaśnianą oraz brak specyfikacji zmiennych objaśniających (ich dobór jest dokonywany wraz z rozwojem drzewa). Ponadto nie ma żadnych ograniczeń co do typu tych zmiennych (mogą to być zmienne zarówno nominalne, porządkowe, jak i ilościowe), a także model jest odporny na transformacje monotoniczne predyktorów oraz na problemy wynikające z jakości materiału statystycznego, tj. obserwacje nietypowe oraz braki danych. Mankamentem tej procedury jest posługiwanie się mało precyzyjną funkcją schodkową (w przypadku oceny wpływ zmiennych mierzonych na skalach mocnych) na prawdopodobieństwo prognozowanego zdarzenia. Nie ma też możliwości oceny wpływu krańcowych zmian wartości zmiennej zależnej na modelowane prawdopodobieństwo.

Pokrótce scharakteryzowane główne podejścia analityczne w budowaniu modeli predykcyjne nie wyznaczają zatem uniwersalnego zestawu narzędzi do optymalizacji grup docelowych strategii biznesowych. Z każdym podejściem wiążą się albo ograniczenia walorów predykcyjnych, albo też ograniczenie walorów eksplikacyjnych, a nawet ich brak. Drogą do przełamania barier tkwiących w tych podejściach analitycznych wydaje się tzw. hybrydyzacja, łącząca metody parametryczne z nieparametrycznymi[15]. Podejście takie zastosowano przy próbie identyfikacji grupy docelowej, jaką była grupa zainteresowana udziałem

[15] M. Łapczyński, *Modele hybrydowe CART-LOGIT w analizie danych rynkowych*, „Prace Naukowe" Uniwersytetu Ekonomicznego we Wrocławiu, nr 51, *Projektowanie, ocena*

w projekcie badawczym „Uwarunkowania decyzji edukacyjnych" (UDE) wśród polskich gospodarstw domowych[16].

3. Hybrydyzacja modeli drzewa klasyfikacyjnego i regresji logistycznej w identyfikacji ogniska zainteresowania udziałem w projekcie UDE

W celu wyboru najlepszego modelu predykcyjnego w przypadku reakcji gospodarstwa domowego na propozycję udziału w projekcie UDE rozważono możliwość budowy modelu przy wykorzystaniu procedury rekurencyjnego podziału, dokonując wyboru między algorytmem CHAID oraz CART, a także według modelu regresji logistycznej oraz wzięto pod uwagę hybrydyzację tych podejść w celu zwiększenia efektywności uzyskanego ostatecznego rozwiązania. Za bazę empiryczną posłużyły wyniki badania zrealizowanego w ramach pierwszej rundy tego projektu w 2013 r. na próbie losowej 122 831 gospodarstw domowych. W próbie wylosowanej w 2013 r. niespełna 20% gospodarstw domowych wyraziło gotowość udziału w tym projekcie. Jako predyktory decyzji o udziale w projekcie przyjęto takie cechy wylosowanych gospodarstw, jak: makroregion, województwo, klasa miejscowości, dzień tygodnia i pory dnia nawiązania kontaktu przez ankietera.

W pierwszej kolejności rozważono możliwość budowy modelu predykcyjnego według obu algorytmów rekurencyjnego podziału. Widoczne niezrównoważenie próby skutkowało bardzo niską jakością rozwiązania początkowego, w którym nie nałożono żadnych kosztów błędnej klasyfikacji. Dlatego też rozważono nałożenie takich kosztów. Wyniki jakości kolejnych rozwiązań uzyskanych przy skokowej modyfikacji kosztów błędnej klasyfikacji, poczynając od 2:1 do 5:1, dla obu modeli zestawiono w tabeli 1.

Na podstawie przeglądu ocen jakości zbudowanych dziesięciu modeli zdecydowano się na wybór modelu utworzonego na podstawie algorytmu CHAID z nałożonymi kosztami błędnej klasyfikacji na poziomie 5:1. Decyzję tę podjęto, biorąc pod uwagę małe zróżnicowanie mierników jakości dla modyfikacji kosztów

i wykorzystanie danych rynkowych, red. J. Dziechciarz, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław 2009.

[16] Projekt Instytutu Badań Edukacyjnych „Uwarunkowania decyzji edukacyjnych", 2013–2015.

błędnej klasyfikacji w zakresie od 3:1 do 5:1 i wysokiego poziomu współczynnika czułości. Wybrany model powstał w rezultacie 11 rekurencji i wyznaczył 179 węzłów końcowych. Wartość przyrostu w trzecim decylu dla tego modelu wyniosła 2,1.

Tabela 1. Jakość modeli rekurencyjnego podziału według algorytmu CHAID oraz CART przy różnym poziomie kosztu błędnej klasyfikacji dla pozytywnej reakcji respondenta (w %)

Miara jakości modelu predykcyjnego	Algorytm	Koszty błędnej klasyfikacji				
		brak	2	3	4	5
Dokładność	CHAID	80,84	77,75	68,07	61,23	57,67
	CART	80,70	79,34	71,23	60,84	54,47
Specyficzność	CHAID	99,33	87,58	66,10	54,31	48,45
	CART	99,91	96,20	78,87	60,82	50,43
Czułość	CHAID	3,84	36,50	76,30	90,12	95,94
	CART	0,51	10,05	39,73	60,92	71,30
Precyzja – pozytywna	CHAID	57,85	41,15	34,98	32,08	30,96
	CART	57,48	39,14	31,33	27,10	25,69
Precyzja – negatywna	CHAID	18,87	14,72	7,90	4,18	1,98
	CART	19,26	18,54	15,64	13,31	12,03
Współczynnik błędu	CHAID	19,16	22,25	31,93	38,77	42,33
	CART	19,30	20,66	28,77	39,16	45,53
Średnia G	CHAID	3,82	31,96	50,43	48,94	46,48
	CART	0,51	9,67	31,34	37,05	35,96
Miara F	CHAID	7,21	38,69	47,97	47,31	46,81
	CART	1,02	15,99	35,04	37,51	37,77

Źródło: opracowanie własne.

Na podstawie tego samego zbioru zmiennych oszacowano model regresji logistycznej. Wszystkie predyktory okazały się zmiennymi istotnie wpływającymi na prawdopodobieństwo podjęcia współpracy, chociaż nie dotyczyło to wszystkich ich kategorii. Kategorię odniesienia wyznaczyły następujące kategorie rozważanych cech gospodarstw domowych:

- dla makroregionu był to makroregion małopolski;
- dla województwa było to województwo zachodniopomorskie;
- dla klasy miejscowości była to Warszawa;
- dla dnia tygodnia była to niedziela;
- dla pory dnia był to wieczór (po godzinie 17.00).

W tabeli 2 zestawiono wynik estymacji współczynników modelu logitowego.

Tabela 2. Oceny współczynników regresji logistycznej prawdopodobieństwa udziału w projekcie UDE

Predykator (kategoria odniesienia)	B	Błąd standardowy	Wald	df	Istotność	Exp(B)
Makroregion (małopolski)						
Centralny	,428	,132	10,454	1	,001	1,534
Wielkopolski	–,383	,179	4,566	1	,033	,682
Śląski	,141	,055	6,552	1	,010	1,151
Północno-wschodni	,499	,156	10,178	1	,001	1,646
Wschodni	,622	,135	21,182	1	,000	1,863
Województwo (zachodniopomorskie)						
Dolnośląskie	1,197	,125	91,116	1	,000	3,310
Kujawsko-pomorskie	,615	,110	31,353	1	,000	1,851
Lubuskie	,349	,123	8,050	1	,005	1,418
Małopolskie	,532	,178	8,922	1	,003	1,702
Opolskie	,505	,186	7,384	1	,007	1,658
Podkarpackie	,610	,179	11,677	1	,001	1,841
Podlaskie	,309	,110	7,868	1	,005	1,362
Pomorskie	,690	,048	205,010	1	,000	1,994
Śląskie	,569	,185	9,466	1	,002	1,767
Świętokrzyskie	,553	,178	9,662	1	,002	1,739
Warmińsko-mazurskie	,377	,094	15,890	1	,000	1,457
Wielkopolskie	,766	,106	51,973	1	,000	2,151
Klasa miejsca zamieszkania (Warszawa)			1335,619	8	,000	
Wieś	–,852	,059	207,946	1	,000	,426
Miasto do 10 tys.	–,732	,066	122,601	1	,000	,481
Miasto 10 tys.–19,9 tys.	–,498	,063	62,611	1	,000	,607
Miasto 20 tys.–49,9 tys.	–,479	,060	63,122	1	,000	,619
Miasto 50 tys.–99,9 tys.	–,413	,063	42,858	1	,000	,662
Miasto 100 tys.–199,9 tys.	–,237	,062	14,610	1	,000	,789
Miasto 200 tys.–499,9 tys.	–,198	,064	9,718	1	,002	,820
Dzień tygodnia (niedziela)						
Poniedziałek	–,086	,033	6,994	1	,008	,917
Wtorek	–,122	,032	14,721	1	,000	,885

Predykator (kategoria odniesienia)	B	Błąd standardowy	Wald	df	Istotność	Exp(B)
Środa	–,334	,031	117,974	1	,000	,716
Czwartek	–,087	,032	7,476	1	,006	,917
Pora dnia (wieczorem)			210,002	2	,000	
Rano	,304	,023	176,356	1	,000	1,355
Stała	1,334	,187	51,133	1	,000	3,796

Źródło: opracowanie własne.

Model regresji logistycznej, w przeciwieństwie do modelu drzewa klasyfikacyjnego, ma jedną niewątpliwą zaletę, jaką jest możliwość oszacowania prawdopodobieństwa przynależności do grupy docelowej dla każdej z rozważanych konfiguracji predyktorów, czyli *de facto* dla każdego przypadku empirycznego, co w modelu drzewa klasyfikacyjnego jest ograniczone jedynie do podklas wyodrębnionych według dokonanych rekurencji. Jednakże nie pozwala na porównanie siły wpływu poszczególnych predyktorów na prawdopodobieństwo przynależności do grupy docelowej oraz oceny efektów interakcji między predyktorami.

Zachowanie waloru eksplikacji przy podniesieniu jakości uzyskiwanego rozwiązania, czyli osiągnięcie w jednym rozwiązaniu korzyści dostępnych w każdym z tych odrębnych rozwiązań, jest możliwe przez zastosowanie podejścia hybrydowego. W tym celu zdecydowano się na dokonanie hybrydyzacji, łącząc model logitowy z modelem rekurencyjnego podziału poprzez uznanie przynależności do wyodrębnionych liści za dodatkowy predyktor modelu parametrycznego. By ograniczyć wymiar analizy, zdecydowano się przyciąć rozwiązanie otrzymane w procedurze rekurencyjnego podziału do trzech rekurencji i podwyższyć próg liczebności liści przed i po podziale, co wyznaczyło 18 węzłów końcowych. Charakterystykę wyznaczonych węzłów zestawiono w tabeli 3.

Tabela 3. Charakterystyki węzłów końcowych w modelu CHAID przyciętym do trzech rekurencji i przy podwyższeniu progu liczebności liści przed i po podziale

Węzeł	Charakterystyki węzła według kategorii predyktorów	Rozmiar węzła (w%)	Oszacowane prawdopodobieństwo udziału w projekcie
Węzeł (1)	miasto 10 tys.–19,9 tys.	7,1	0,207
Węzeł (2)	miasto do 10 tys.	4,7	0,241
Węzeł (3)	mazowieckie, dolnośląskie, miasta 500 tys.–999,9 tys.	7,1	0,094
Węzeł (4)	mazowieckie, lubelskie, podlaskie, pomorskie, 20 tys.–99,9 tys.	4,5	0,144

Węzeł	Charakterystyki węzła według kategorii predyktorów	Rozmiar węzła (w%)	Oszacowane prawdopodobieństwo udziału w projekcie
Węzeł (5)	małopolskie, śląskie, dolnośląskie, 20 tys.–99,9 tys.	5,7	0,174
Węzeł (6)	kujawsko-pomorskie, opolskie, lubuskie, wielkopolskie, 20 tys.–99,9 tys.	3,8	0,261
Węzeł (7)	gmina wiejska, rano (do godz. 12.00)	5,8	0,186
Węzeł (8)	mazowieckie, lubelskie, podlaskie, świętokrzyskie, pomorskie, miasto 200 tys.–499,9 tys.	8,6	0,105
Węzeł (9)	śląskie, kujawsko-pomorskie, zachodniopomorskie, 200 tys.–499,9 tys.	3,6	0,236
Węzeł (10)	mazowieckie, śląskie, dolnośląskie, miasto 100 tys.–199,9 tys.	4,1	0,115
Węzeł (11)	małopolskie, warmińsko-mazurskie, opolskie, zachodniopomorskie, podkarpackie, miasto 100 tys.–199,9 tys.	6,5	15,7
Węzeł (12)	kujawsko-pomorskie, lubuskie, wielkopolskie, miasto 100 tys.–199,9 tys.	4,3	0,227
Węzeł (13)	Warszawa, miasto 500 tys.–999,9 tys., małopolskie, łódzkie, wielkopolskie, rano (do 12.00) i po południu (12.00–17.00)	3,9	18,2
Węzeł (14)	Warszawa, miasto 500 tys.–999,9 tys., małopolskie, łódzkie, wielkopolskie, wieczorem (po 17.00)	3,3	0,154
Węzeł (15)	miasto 20 tys.–49,9 tys., warmińsko-mazurskie, świętokrzyskie, łódzkie, zachodniopomorskie, wtorek, środa, czwartek	3,7	0,211
Węzeł (16)	miasto 20 tys.–49,9 tys., warmińsko-mazurskie, świętokrzyskie, łódzkie, zachodniopomorskie, piątek, sobota, niedziela, poniedziałek	4,0	0,165
Węzeł (17)	gmina wiejska, mazowieckie, lubelskie, podlaskie, małopolskie, łódzkie, kujawsko-pomorskie, świętokrzyskie, lubuskie, podkarpackie, po południu (12.00–17.00) i wieczorem (po 17.00)	12,3	0,314
Węzeł (18)	gmina wiejska, warmińsko-mazurskie, opolskie, łódzkie, pomorskie, zachodniopomorskie, wielkopolskie, dolnośląskie, po południu (12.00–17.00) i wieczorem (po 17.00)	7,0	0,232

Źródło: opracowanie własne.

Zestawienie miar jakości przyciętego modelu CHAID, modelu logitowego oraz hybrydy obu podejść analitycznych przedstawia tabela 4.

Tabela 4. Poziomy miar jakości przyciętego modelu CHAID, oszacowanego modelu logitowego oraz oszacowanej hybrydy obu podejść analitycznych (w %)

Miara jakości modelu	Przycięty model CHAID	Model logitowy	Hybryda obu podejść analitycznych
Dokładność	57,76	56,13	59,00
Specyficzność	57,48	56,47	58,73
Czułość	58,89	54,68	60,16
Precyzja – pozytywna	24,94	23,16	25,91
Precyzja – negatywna	14,65	16,15	14,00
Współczynnik błędu	42,24	43,87	41,00
Średnia	33,85	30,88	35,33
Miara F	35,04	32,54	36,22

Źródło: opracowanie własne.

Jak pokazują dane zestawione w tabeli 5, w porównaniu z rozwiązaniem uzyskanym w podejściu nieparametrycznym model regresji logistycznej charakteryzował się niewiele niższą jakością pod względem większości wskaźników. Natomiast model powstały w wyniku hybrydyzacji obu podejść analitycznych charakteryzował się najlepszymi poziomami wskaźników jakości, a dodatkowo stworzył możliwość estymacji prawdopodobieństwa przynależności do grupy docelowej dla każdego z możliwych do rozważenia przypadków. Uwzględnienie przynależności do węzłów z rozwiązania nieparametrycznego pozwoliło ponadto na ocenę wpływu interakcji między predyktorami na prawdopodobieństwo przynależności do grupy docelowej (tabela 5).

Tabela 5. Istotna ocena współczynników regresji logistycznej w modelu hybrydowym

Predyktor (kategoria odniesienia)	B	Błąd standardowy	Wald	df	Istotność	Exp (B)
Makroregion (małopolski)						
Centralny	,392	,130	9,113	1	,003	1,479
Wielkopolski	–,566	,178	10,109	1	,001	,568
Zachodni	–,422	,194	4,760	1	,029	,655

Predyktor (kategoria odniesienia)	B	Błąd standardowy	Wald	df	Istotność	Exp (B)
Północno-wschodni	,423	,154	7,556	1	,006	1,526
Wschodni	,673	,132	25,863	1	,000	1,960
Województwo (zachodniopomorskie)						
Dolnośląskie	1,071	,126	71,958	1	,000	2,919
Kujawsko-pomorskie	,899	,111	66,130	1	,000	2,458
Lubuskie	,731	,126	33,873	1	,000	2,076
Małopolskie	,483	,177	7,443	1	,006	1,622
Opolskie	,382	,186	4,228	1	,040	1,465
Podkarpackie	,473	,177	7,145	1	,008	1,604
Pomorskie	,282	,057	24,935	1	,000	1,326
Śląskie	,476	,184	6,656	1	,010	1,609
Świętokrzyskie	,367	,176	4,338	1	,037	1,444
Warmińsko-mazurskie	,222	,094	5,536	1	,019	1,248
Wielkopolskie	,975	,107	82,475	1	,000	2,651
Miejsce zamieszkania (Warszawa)						
Miasto 20 tys.–49,9 tys.	–,126	,034	13,966	1	,000	,882
Miasto 500 tys.–999,9 tys.	1,064	,115	86,018	1	,000	2,897
Dzień tygodnia (niedziela)						
Poniedziałek	–,086	,033	6,899	1	,009	,918
Wtorek	–,122	,032	14,302	1	,000	,885
Środa	–,324	,031	107,900	1	,000	,723
Czwartek	–,080	,032	6,234	1	,013	,923
Pora dnia (wieczorem)						
Rano	,179	,028	42,104	1	,000	1,196
Po południu	,058	,017	11,534	1	,001	1,060
Węzeł (węzeł (18))						
Węzeł (1)	,136	,040	11,665	1	,001	1,145
Węzeł (2)	–,085	,042	4,027	1	,045	,919
Węzeł (3)	,751	,068	121,277	1	,000	2,119
Węzeł (4)	,642	,059	120,102	1	,000	1,900
Węzeł (5)	,285	,053	29,006	1	,000	1,329
Węzeł (6)	–,093	,048	3,696	1	,055	,911
Węzeł (7)	,133	,050	7,201	1	,007	1,142
Węzeł (8)	,932	,052	327,089	1	,000	2,539

Predyktor (kategoria odniesienia)	B	Błąd standardowy	Wald	df	Istotność	Exp (B)
Węzeł (9)	,014	,052	,069	1	,792	1,014
Węzeł (10)	,723	,060	143,017	1	,000	2,060
Węzeł (11)	,441	,044	99,438	1	,000	1,555
Węzeł (12)	,090	,049	3,439	1	,064	1,094
Węzeł (13)	–,839	,128	43,142	1	,000	,432
Węzeł (14)	–,538	,130	17,196	1	,000	,584
Węzeł (15)	,252	,052	23,167	1	,000	1,287
Węzeł (16)	,395	,055	52,476	1	,000	1,484
Węzeł (17)	–,427	,037	131,282	1	,000	,652
Stała	,824	,180	21,073	1	,000	2,280

Źródło: opracowanie własne.

Istotne interakcje ujawniły się w 14 na 18 rozważanych przypadków i dotyczyły interakcji zarówno drugiego stopnia, czyli koincydencji dwóch predyktorów, jak i trzeciego stopnia, czyli koincydencji trzech predyktorów. Spośród tych istotnych interakcji w dziesięciu przypadkach ich efekt stymulował prawdopodobieństwo przynależności do grupy docelowej.

4. Podsumowanie

Model hybrydowy pozwala porównać efekt oddziaływania na prawdopodobieństwo przynależności do grupy docelowej każdej kategorii z tych predyktorów, które okazały się istotne, względem właściwej dla każdego z nich kategorii referencyjnej przy kontrolowanym wpływie pozostałych. Stworzył również możliwość oceny efektu interakcji między predyktorami. Uzyskane rozwiązanie ujawniło, iż interakcje te mają ściśle określone, lokalne uwarunkowania. Hybrydyzacja dostarczyła zatem zarówno waloru eksplikacyjnego dla efektów głównych oraz efektów złożonych z konfiguracji predyktorów, jak i waloru predykcyjnego, osiągając satysfakcjonującą poprawność przewidywań.

Bibliografia

Berger P., Bolton R., Bosman D., Briggs E., Kumar V., Parasuraman A., Terry C., *Marketing Actions and the Value of Customer Base. A Framework for Customer Asset Management*, „Journal of Service Research" 2002, vol. 5, no. 1.

Buzzell R.D., Gale B.T., *The PIMS Principles: Linking Strategy to Performance*, Free Press, New York 1987.

Greene W.H., *Econometric Analysis*, Prentice Hall International Inc., Upper Saddle River 2000.

Gruszczyński M., *Modele i prognozy zmiennych jakościowych w finansach i bankowości*, Oficyna Wydawnicza SGH, Warszawa 2002.

Gummesson E., *Total Relationship Marketing*, Butterworth-Heinemann, Oxford 2008.

Łapczyński M., *Hybrydowe modele predykcyjne w marketingu relacji*, Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, Kraków 2015.

Łapczyński M., *Modele hybrydowe CART-LOGIT w analizie danych rynkowych*, „Prace Naukowe" Uniwersytetu Ekonomicznego we Wrocławiu, nr 51, *Projektowanie, ocena i wykorzystanie danych rynkowych*, red. J. Dziechciarz, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław 2009.

Maddala G., *Intruduction to Econom*wetrics, John Wiley & Sons Ltd, Chichester 2001.

Manski C.F., *Maximum score estimation of the stochastic utility model of choice*, „Journal of Economietrics" 1975, no. 3.

Nagi E.W.T., Xiu L., Chau D.C.K., *Application of Data Mining Techniques in Customer Relationship Management*, „A Literature Review and Classification. Expert System with Application" 2009, no. 36.

Noori H., Radford R., *Production and operation management. Total Quality and responsiveness*, McGraw-Hill, New York 1995.

Otto J., *Marketing relacji. Koncepcja i stosowanie*, Wydawnictwo C.H. Beck, Warszawa 2004.

Owczarczuk M., *Estymatory typu Maximum score dla wybranych modeli mikroekonometrycznych*, praca doktorska, Kolegium Analiz Ekonomicznych, SGH, Warszawa, 2009.

Walesiak M., *Metody analizy danych marketingowych*, Wydawnictwo Naukowe PWN, Warszawa 1996.

Źródła sieciowe

http://www.ge.com/en/company/companyinfo/at_a_glance/key_growth_initiatives.htm.

http://www.valuebasedmanagement.net/methods_valuecreationindex.html.

* * *

Prediction methods in analytical CRM for relationship marketing

Summary

Close contacts with the customer, analysis of customer service costs and revenues generated by them, databases and information technology have become the prerequisites for an effective marketing orientation. Analysis cannot do without models referring to the conditions of companies' activities. Models of the instruments of market influence allow to predict the results of the planned activities in the changing market environment. For these reasons, much attention is paid to the design and formal verification of empirical models optimising marketing communication addressed to the client in order to develop the client's value for the company. The paper reviews the formal models used to optimise the selection of target groups for marketing communications in analytical Customer Relation Management. The effectiveness of parametric, non-parametric and semi-parametric predictive tools is discussed. The paper also considers a quantitative approach to the optimisation of target groups among Polish households that was to participate in the scientific project entitled "Determiners of decisions concerning education". The hybrid approach was also described. In this case, this kind of approach consists in combining data mining tools (classification tree) with other analytical tools (logistic regression).

Keywords: models of business operation conditions, CRM, optimisation of the selection of target groups, hybrid approach