Nr 51

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

Projektowanie, ocena i wykorzystanie danych rynkowych

Redaktor naukowy
Józef Dziechciarz

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2009

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 51 2009
Projektowanie, ocena i wykorzystanie danych rynkowych

Mariusz Łapczyński
Uniwersytet Ekonomiczny w Krakowie

MODELE HYBRYDOWE CART-LOGIT W ANALIZIE DANYCH RYNKOWYCH

1. Wstęp

Celem artykułu jest omówienie hybrydowego podejścia CART-LOGIT do budowy modeli predykcyjnych w badaniach marketingowych. Modele zbudowane za pomocą narzędzi *data mining* – CART i RandomForests, model regresji logistycznej oraz dwa powstałe w podejściu dwufazowym CART/RandomForests-LOGIT zostaną porównane z modelami hybrydowymi – łączącymi drzewa klasyfikacyjne CART z regresją logistyczną. Przykład opiera się na zbiorze danych zebranych w trakcie badań ankietowych zrealizowanych w czerwcu i lipcu 2005 r. Celem tych badań była próba modelowania preferencji konsumentów na rynku samochodów osobowych[1]. Rozważania zawężono do tzw. deklarowanych preferencji (*stated preferences*), które w odróżnieniu od preferencji ujawnionych (*revealed preferences*) nie są rzeczywistymi aktami zakupu, a jedynie wyrażeniem gotowości do nabycia przez respondentów danej kategorii/marki produktu. Zestaw zmiennych objaśniających obejmował pozycje ze skali wartości LOV (*list of values*), preferowane atrybuty samochodu ze skali struktury korzyści (BSA) oraz zmienne demograficzne respondentów.

2. Model zbudowany przy użyciu drzew klasyfikacyjnych CART

Drzewa klasyfikacyjne CART[2] to narzędzie analityczne *data mining*, które jest uznawane za najbardziej zaawansowaną metodę podziału rekurencyjnego. Mimo że metoda ta powstała na początku lat osiemdziesiątych ubiegłego wieku, to do dziś

[1] Opisane w dalszej części pracy wyniki badań dotyczą nowych i używanych samochodów osobowych, których cena wynosi w przybliżeniu 40 tys. zł.

[2] Pierwsza praca poświęcona algorytmowi CART to pozycja [Breiman i in. 1984]. W literaturze polskiej najpopularniejsza praca to [Gatnar 2001].

doczekała się tylko nieznacznych modyfikacji. Próbowano wprawdzie stworzyć bayesowski CART [Chapman, George, McCulloch 1998, s. 935-960], dokonywano jego modyfikacji w NASA (pakiet IND) [Buntine 1992]; usiłowano także udoskonalić podział drzew (FACT) [Loh, Vanichsetakul 1988, s. 715-729] poprzez połączenie właściwości CART i liniowej analizy dyskryminacyjnej; podejmowano próby zastąpienia wielokrotnej walidacji krzyżowej metodą Monte Carlo [Crawford 1989, s. 197-217]; jednak rdzeń metody z jego nowatorskimi rozwiązaniami do dziś pozostał niezmieniony.

Model predykcyjny zbudowany za pomocą drzew klasyfikacyjnych CART miał 17 liści i był głęboki na 8 poziomów. Jakość mierzona odsetkiem poprawnie sklasyfikowanych przypadków została przedstawiona w tab. 1.

Tabela 1. Macierz błędnych klasyfikacji dla modelu CART

CART / 10CV / *a priori* szacowane – procent poprawnych klasyfikacji 70,43%		
	Przewidywana UŻYWANY	Przewidywana NOWY
Obserwowana UŻYWANY	232 (88,21%)	31 (11,79%)
Obserwowana NOWY	105 (53,30%)	92 (46,70%)

Źródło: opracowanie własne z wykorzystaniem programu STATISTICA Data Miner.

Ranking ważności predyktorów obejmował fazę 1, 4 i 5 cyklu życia rodziny, pozycje ze skali LOV (poczucie spełnienia i poczucie przynależności) oraz pozycje ze skali struktury korzyści (funkcjonalność, komfort, osiągi techniczne i bezpieczeństwo).

3. Model zbudowany za pomocą narzędzia RandomForests

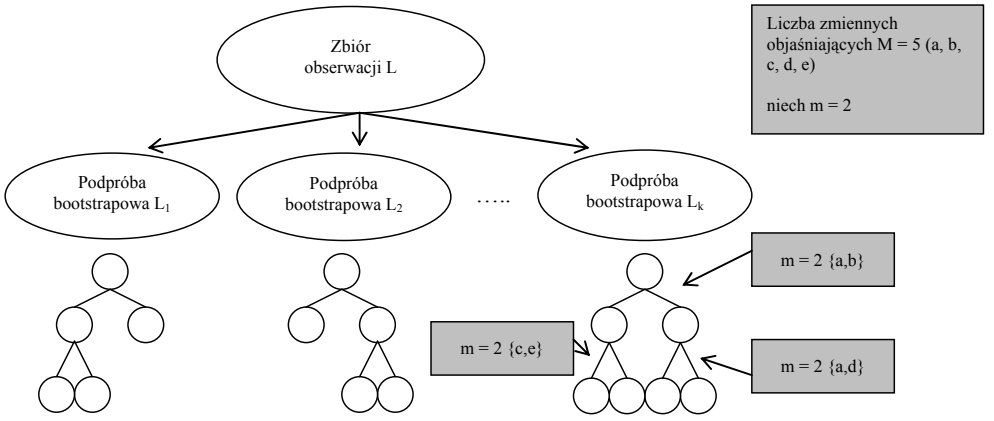
Losowy las (RandomForests) to narzędzie analityczne *data mining*, które buduje wiele modeli o strukturze drzewa w trakcie jednej analizy. Spośród dużej liczby drzew (tzw. lasu) wybierane jest to, które zapewnia najwyższą trafność predykcji mierzoną odsetkiem poprawnie sklasyfikowanych przypadków. Procedura analityczna składa się z trzech etapów [Breiman, Cutler, s. 3]:

1. Ze zbioru danych L o liczebności n losuje się w sposób prosty niezależny (ze zwracaniem) próbę liczącą n przypadków; powstaje w ten sposób podpróba bootstrapowa (L_1, L_2, …, L_k), wykorzystana do zbudowania pojedynczego modelu.

2. Z całego zbioru M zmiennych niezależnych losuje się zestaw m zmiennych (gdzie $m < M$); losowanie to przeprowadza się na każdym etapie podziału pojedynczych modeli, przy czym ustalona na początku wartość m pozostaje niezmieniona przez cały czas trwania analizy.

3. Każdy model budowany jest do maksymalnie dużych rozmiarów (mimo że używa się tu algorytmu CART, to jednak nie korzysta się z opcji przycinania).

Schemat powstawania losowego lasu przedstawiono na rys. 1.



Rys. 1. Schemat powstawania losowego lasu

Źródło: opracowanie własne.

Predykcja przynależności nowych obiektów do klas (wariantów zmiennej zależnej) odbywa się na podstawie klasyfikacji pojedynczych drzew z losowego lasu. Przy ocenie poszczególnych rozwiązań stosuje się zasadę majoryzacji zbiorów, która polega na tym, że liść drzewa otrzymuje etykietę od klasy występującej w nim najliczniej. W literaturze anglojęzycznej proces ten nosi nazwę „głosowania" (*voting*), co należałoby tu wyjaśnić jako oddawanie głosów przez poszczególne modele na przypisanie obiektu do danej klasy.

Tabela 2. Macierz błędnych klasyfikacji dla modelu RandomForests

RF/ 5 zmiennych / 100 drzew – procent poprawnych klasyfikacji 76,43%		
	Przewidywana UŻYWANY	Przewidywana NOWY
Obserwowana UŻYWANY	151 (87,28%)	22 (12,72%)
Obserwowana NOWY	52 (36,88%)	89 (63,12%)

Źródło: opracowanie własne z wykorzystaniem programu STATISTICA Data Miner.

W tym wypadku model cechuje się dosyć wysokim odsetkiem poprawnych klasyfikacji (tab. 2). Na liście zmiennych objaśniających o największej liczbie punktów w rankingu ważności predyktorów znajdują się: płeć (kobieta), wykształcenie (średnie), faza 1 i 5 cyklu życia rodziny, pozycje ze skali LOV (bezpieczeństwo, szacunek dla samego siebie, serdeczne stosunki z innymi, poczucie przynależności) oraz pozycje ze skali struktury korzyści (komfort, koszty eksploatacji).

4. Model zbudowany z zastosowaniem regresji logistycznej

Regresja logistyczna jest matematycznym podejściem modelowym opisującym zależność między zestawem zmiennych niezależnych a jakościową – dychotomiczną zmienną zależną. Ze względu na prostotę tego popularnego narzędzia pominięto jego opis, koncentrując się wyłącznie na wynikach analizy. Rozwiązanie jest istotne statystycznie ($p = 0{,}000$), a wśród zmiennych objaśniających włączonych do modelu znalazły się[3]: funkcjonalność ze skali struktury korzyści (0,65), płeć − kobieta (3,82), faza 1 cyklu życia rodziny (0,21), faza 2 cyklu życia rodziny (0,24) i faza 3 cyklu życia rodziny (0,29). Ogólna jakość modelu jest najniższa z trzech dotychczas zbudowanych (por. tab. 3).

Tabela 3. Macierz błędnych klasyfikacji dla regresji logistycznej

Regresja logistyczna – procent poprawnych klasyfikacji 66,74%, iloraz szans = 3,77		
	Przewidywana UŻYWANY	Przewidywana NOWY
Obserwowana UŻYWANY	199 (75,66%)	64 (24,34%)
Obserwowana NOWY	89 (45,18%)	108 (54,82%)

Źródło: opracowanie własne z wykorzystaniem programu STATISTICA Data Miner.

5. Podejście dwufazowe CART-LOGIT

Pierwszy etap analizy w podejściu dwufazowym polega na wykorzystaniu algorytmu CART do wstępnej eksploracji danych. Już w latach siedemdziesiątych ubiegłego stulecia pisano bowiem o drzewach klasyfikacyjnych jako narzędziach służących do przetrząsania danych. W etapie drugim użyto prostej regresji logistycznej, jednak do modelu włączono tylko te zmienne, które zostały „uznane" za potencjalnie wartościowe przez algorytm CART. Były to *de facto* zmienne znajdujące się wysoko w rankingu ważności predyktorów. Ostatecznie do modelu trafiły: faza 4 cyklu życia rodziny z jednostkowym ilorazem szans równym 1,90 oraz faza 5 z ilorazem równym 4,34. Jakość rozwiązania przedstawiono w tab. 4.

Tabela 4. Macierz błędnych klasyfikacji dla modelu dwufazowego CART-LOGIT

CART → LOGIT – procent poprawnych klasyfikacji 63,26%, iloraz szans = 2,81		
	Przewidywana UŻYWANY	Przewidywana NOWY
Obserwowana UŻYWANY	215 (81,75%)	48 (18,25%)
Obserwowana NOWY	121 (61,43%)	76 (38,57%)

Źródło: opracowanie własne z wykorzystaniem programu STATISTICA Data Miner.

[3] W nawiasie za każdą zmienną objaśniająca podano jednostkowe ilorazy szans.

6. Podejście dwufazowe RandomForests-LOGIT

W drugim wariancie podejścia dwufazowego zamiast algorytmu CART użyto techniki RandomForests (w skrócie RF). Za takim rozwiązaniem przemawiają dwa argumenty. Po pierwsze, metoda RF jest oparta na występującym w algorytmie CART indeksie Giniego, a po drugie, jednym z wyników analizy RF jest ranking ważności predyktorów, który może być podstawą do dalszych rozważań nad strukturą danych.

Jakość rozwiązania jest pod każdym względem najniższa spośród wszystkich opisanych dotychczas modeli (por. tab. 5). W modelu znalazły się trzy statystycznie istotne zmienne objaśniające: płeć − kobieta (jedn. il.szans = 3,62), faza 1 cyklu życia rodziny (jedn. il.szans = 0,55) i faza 5 cyklu życia rodziny (jedn. il.szans = 3,95).

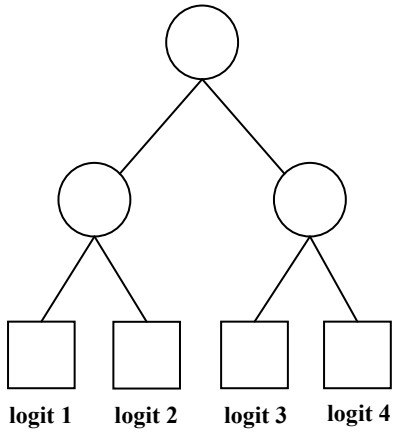
Tabela 5. Macierz błędnych klasyfikacji dla modelu dwufazowego RF-LOGIT

RF → LOGIT – procent poprawnych klasyfikacji 68,91%, iloraz szans = 4,66		
	Przewidywana UŻYWANY	Przewidywana NOWY
Obserwowana UŻYWANY	206 (78,33%)	57 (21,67%)
Obserwowana NOWY	86 (43,66%)	111 (56,34%)

Źródło: opracowanie własne z wykorzystaniem programu STATISTICA Data Miner.

7. Model hybrydowy CART-LOGIT (odmiana I)

Tworzenie hybrydowych narzędzi analitycznych polega na łączeniu różnych technik analitycznych. W tym wypadku będą to, wywodzące się z *data mining*, drzewa klasyfikacyjne CART oraz regresja logistyczna. Modele logitowe są budowane w



Rys. 2. Sposób budowania modelu hybrydowego CART-LOGIT (odmiana I)

Źródło: opracowanie własne.

podzbiorach całego zbioru obserwacji, jakie trafiły do poszczególnych liści drzewa. Liczba liści wyznacza zatem liczbę modeli logitowych, co z kolei wymusza budowę drzewa klasyfikacyjnego o względnie małej wielkości (por. rys. 2). Można tego dokonać albo poprzez ograniczenie *a priori* głębokości drzewa, ograniczenie wielkości drzewa, albo też ręczne przycięcie modelu.

W niniejszym rozwiązaniu punktem wyjścia był model drzewa z dwoma liśćmi i „płcią respondenta" jako zmienną wykorzystaną na jedynym etapie podziału. Owa dychotomia determinowała dwukrotne wykorzystanie regresji logistycznej – raz w grupie kobiet i drugi raz w grupie mężczyzn. Dwa modele to dwie macierze błędnych klasyfikacji (tab. 6 i 7) oraz dwa zestawy zmiennych objaśniających. W grupie kobiet były to: bezpieczeństwo samochodu (jedn. il.szans = 4,75) i funkcjonalność samochodu (jedn. il.szans = 0,26), natomiast w grupie mężczyzn: faza 1 cyklu życia rodziny (jedn. il.szans = 0,29) i faza 3 cyklu życia rodziny (jedn. il.szans = 0,36).

Tabela 6. Macierz błędnych klasyfikacji dla modelu hybrydowego CART-LOGIT (odmiana I – podzbiór: kobiety)

CART-LOGIT 1 – procent poprawnych klasyfikacji 72,03%, iloraz szans = 7,50		
	Przewidywana UŻYWANY	Przewidywana NOWY
Obserwowana UŻYWANY	15 (34,88%)	28 (65,12%)
Obserwowana NOWY	5 (6,67%)	70 (93,33%)

Źródło: opracowanie własne z wykorzystaniem programu STATISTICA Data Miner.

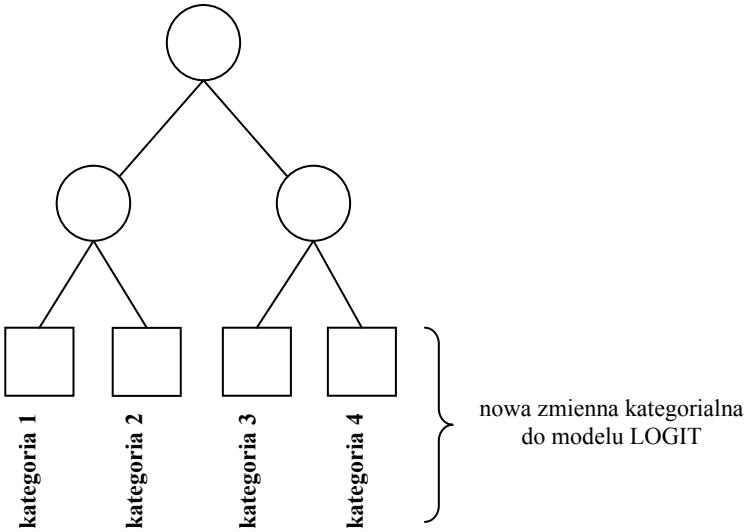
Tabela 7. Macierz błędnych klasyfikacji dla modelu hybrydowego CART-LOGIT (odmiana I – podzbiór: mężczyźni)

CART-LOGIT 1 – procent poprawnych klasyfikacji 65,20%, iloraz szans = 3,00		
	Przewidywana UŻYWANY	Przewidywana NOWY
Obserwowana UŻYWANY	155 (70,45%)	65 (29,55%)
Obserwowana NOWY	54 (44,26%)	68 (55,74%)

Źródło: opracowanie własne z wykorzystaniem programu STATISTICA Data Miner.

8. Model hybrydowy CART-LOGIT (odmiana II)

Nieco inne podejście do stworzenia hybrydy CART-LOGIT polega na budowie modelu logitowego dla całego zbioru obserwacji z wykorzystaniem informacji o zawartości liści. W pierwszym kroku należy zatem zbudować model CART o niezbyt dużej głębokości. Następnie należy utworzyć zmienną kategorialną, gdzie każda z kategorii oznacza przynależność przypadku do liścia. Następnie zamienia się tę zmienną na zestaw zmiennych sztucznych (*dummies*) i dołącza do pozostałych zmiennych objaśniających w modelu logitowym. Zmienna ta reprezentuje regułę „jeżeli…, to…" i strukturę interakcji odkrytą przez CART (rys. 3).



Rys. 3. Sposób budowania modelu hybrydowego CART-LOGIT (odmiana II)

Źródło: opracowanie własne.

Tabela 8. Macierz błędnych klasyfikacji dla modelu hybrydowego CART-LOGIT (odmiana II)

CART-LOGIT 2 – procent poprawnych klasyfikacji 69,35%, iloraz szans = 5,25		
	Przewidywana UŻYWANY	Przewidywana NOWY
Obserwowana UŻYWANY	222 (84,41%)	41 (15,59%)
Obserwowana NOWY	100 (50,77%)	97 (49,23%)

Źródło: opracowanie własne z wykorzystaniem programu STATISTICA Data Miner.

Jakość rozwiązania hybrydowego przedstawiono w tab. 8. Zmienne objaśniające, które trafiły do ostatecznej postaci modelu, to: faza 4 cyklu życia rodziny (jedn. il.szans = 2,22), liść 1 (jedn. il.szans = 5,75), liść 2 (jedn. il.szans = 1,19) i liść 3 (jedn. il.szans = 1,22). Konieczne w tym miejscu jest wyjaśnienie, w jaki sposób interpretować liście drzewa:

- liść 1 to kobiety z każdej – poza pierwszą – fazy cyklu życia rodziny;
- liść 2 to kobiety z pierwszej fazy cyklu życia rodziny;
- liść 3 to mężczyźni z piątej fazy cyklu życia rodziny.

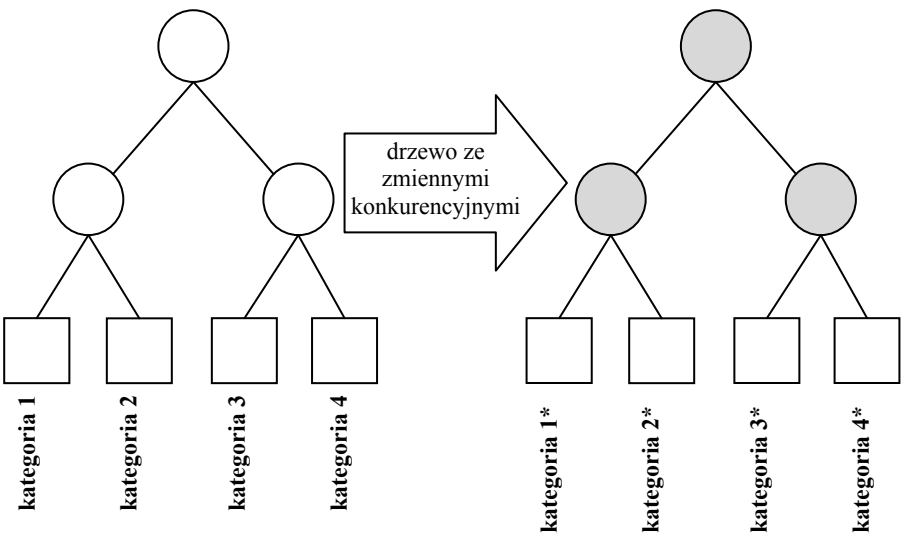
9. Model hybrydowy CART-LOGIT (odmiana III)

Trzecie podejście w budowie modelu hybrydowego CART-LOGIT wykorzystuje charakterystyczne dla CART istnienie zmiennych zastępczych i zmiennych konkurencyjnych. Na każdym etapie podziału drzewa zestawiany jest ranking zmien-

nych niezależnych, które zapewniają najlepszy podział danego węzła. Pozycja w tym rankingu zależy od trafności predykcji zmiennej zależnej w wydzielanych węzłach potomnych. Najlepszy z predyktorów wykorzystywany jest do budowy modelu, a pozostałe pełnią funkcję bądź to zmiennych konkurencyjnych, bądź zmiennych zastępczych (albo obie te funkcje jednocześnie). Kolejność w tych lokalnych rankingach zależy od wartości tzw. wskaźnika poprawy (*improvement*). Zmienna niezależna, dla której wartość ta jest najwyższa, zostaje uznana za najlepszy predyktor pierwotny, który dzieli dany węzeł. Kolejność pozostałych predyktorów jest determinowana posortowanymi malejąco wartościami wskaźnika poprawy.

Różnica między zmiennymi konkurencyjnymi a zmiennymi zastępczymi sprowadza się do tego, że te pierwsze zapewniają zbliżoną redukcję heterogeniczności węzła, te drugie zaś, oprócz redukcji nieczystości węzła, „naśladują" najlepszy predyktor, rozdzielając konkretne przypadki ze zbioru obserwacji w sposób jak najbardziej zbliżony do podziału pierwotnego (*case-by-case*).

Zmienna zastępcza to zazwyczaj predyktor, który w danym miejscu drzewa zajmuje drugą pozycję i jednocześnie dzieli dany węzeł w sposób zbliżony do tego, jaki daje zmienna niezależna z pierwszej pozycji. Wprowadzenie zmiennych zastępczych niesie ze sobą kilka korzyści, z których najbardziej tu przydatną jest lepsze zrozumienie badanego zjawiska. Badacz może się przyjrzeć dokładnie strukturze drzewa i sprawdzić, jakie inne zmienne mogłyby być potencjalnymi predyktorami na poszczególnych etapach podziału. Niektóre programy (np. *STATISTICA* Data Miner, CART® wersja 6.0) pozwalają na ingerencję analityka w strukturę modelu. Możliwe jest usuwanie (dodawanie) dowolnych gałęzi i wprowadzanie własnych podziałów.



Rys. 4. Sposób budowania modelu hybrydowego CART-LOGIT (odmiana III)

Źródło: opracowanie własne.

W trzeciej odmianie modeli hybrydowych wykorzystano taką właśnie zmodyfikowaną przez badacza postać drzewa klasyfikacyjnego CART (por. rys. 4).

Zmienne objaśniające, które znalazły się w ostatecznym rozwiązaniu, to: faza 1 cyklu życia rodziny (jedn. il.szans = 0,27), faza 2 cyklu życia rodziny (jedn. il.szans = 0,32), faza 3 cyklu życia rodziny (jedn. il.szans = 0,40), liść 1* (jedn. il.szans = 3,75) i liść 3* (jedn. il.szans = 4,37). Przypadki, które znajdują się w liściu 1*, to osoby z piątej fazy cyklu życia rodziny, które nie przykładają wagi do wyglądu samochodu, zaś w liściu 3* znajdują się kobiety z każdej – poza piątą – fazy cyklu życia rodziny. Procent poprawnych klasyfikacji i iloraz szans zamieszczono w tab. 9.

Tabela 9. Macierz błędnych klasyfikacji dla modelu hybrydowego CART-LOGIT (odmiana III)

CART-LOGIT 3 – procent poprawnych klasyfikacji 68,48%, iloraz szans = 4,62		
	Przewidywana UŻYWANY	Przewidywana NOWY
Obserwowana UŻYWANY	215 (81,75%)	48 (18,25%)
Obserwowana NOWY	97 (49,24%)	100 (50,76%)

Źródło: opracowanie własne z wykorzystaniem programu STATISTICA Data Miner.

10. Podsumowanie

W poniższej tabeli (tab. 10) zamieszczono porównanie wszystkich modeli z uwzględnieniem pięciu kryteriów: odsetka poprawnych klasyfikacji dla całego modelu, odsetka poprawnych klasyfikacji klientów preferujących nowy samochód, odsetka poprawnych klasyfikacji klientów preferujących samochód używany, ilorazu szans oraz – dla modeli logitowych – współczynnika R^2 McFaddena.

Tabela 10. Podsumowanie wyników dla ośmiu modeli predykcyjnych

Model	Procent poprawnych klasyfikacji	Procent dla klasy „NOWE”	Procent dla klasy „UŻYWANE”	Iloraz szans	R^2 McFaddena
CART	70,43	46,70	88,21	6,55	XXX
RandomForests	76,43	63,12	87,28	11,75	XXX
LOGIT	66,74	54,82	75,66	3,77	0,113
CART → LOGIT	63,26	38,57	81,75	2,81	0,045
RF → LOGIT	68,91	56,34	78,33	4,66	0,098
CART-LOGIT 1 (kobiety)	72,03	93,33	34,88	7,50	0,082
CART-LOGIT 1 (mężczyźni)	65,20	55,74	70,45	3,00	0,052
CART-LOGIT 2	69,35	49,23	84,41	5,25	0,115
CART-LOGIT 3	68,48	50,76	81,75	4,62	0,127

Źródło: opracowanie własne.

Najwyższą ogólną trafność predykcji zapewnił model RandomForests, pierwsza odmiana hybrydowego CART-LOGIT dla kobiet oraz klasyczny CART. Jeśli chodzi o trafność predykcji dla klasy „samochód nowy", to najlepsze wyniki otrzymano z zastosowaniem: RandomForests, podejścia dwufazowego RF → LOGIT oraz pierwszej odmiany modelu hybrydowego CART-LOGIT dla kobiet. W odniesieniu do odsetka poprawnych klasyfikacji respondentów preferujących samochód używany najlepszy rezultat dostarczyły dwa narzędzia *data mining* oraz druga odmiana modelu hybrydowego CART-LOGIT. Jeśli spojrzeć na iloraz szans, to najlepiej wypadły CART, RF i pierwsza odmiana hybrydowego CART-LOGIT dla kobiet, natomiast jeśli brać pod uwagę R^2 McFaddena, to najlepsze są trzecia i druga odmiana modelu hybrydowego oraz prosta regresja logistyczna. Wydaje się zatem, że nie ma tutaj narzędzia/podejścia, które byłoby zdecydowanie lepsze od pozostałych, co z kolei zdaje się potwierdzać opinię o konieczności budowy modeli predykcyjnych za pomocą wielu narzędzi analitycznych.

Literatura

Breiman L. i in., *Classification and regression trees*, Chapman and Hall, 1984.

Breiman L., Cutler A., *Random Forests*, stat-www.berkeley.edu (15.10.2007).

Buntine W., *Tree classification software*, ''Technology 2002", Baltimore, December 1992.

Chapman H.A., George E.I., McCulloch R.E., *Bayesian CART model search*, "Journal of the American Statistical Association", September 1998, vol. 93, no. 443, s. 935-960.

Crawford S.L., *Extension to the CART algorithm*, "International Journal Man-Machine Studies" 1989, vol. 31, s. 197-217.

Gatnar E., *Nieparametryczna metoda dyskryminacji i regresji*, PWN, Warszawa 2001.

Loh W-Y., Vanichsetakul N., *Tree-structured classification via generalized discriminant analysis*, "Journal of the American Statistical Association", September 1988, vol. 83, no. 403, s. 715-729.

Łapczyński M., *Predykcja zjawisk rynkowych za pomocą narzędzia RandomForests*, opracowanie cząstkowe w ramach badań statutowych pod kier. prof. dr. hab. Adama Sagana, Uniwersytet Ekonomiczny, Kraków 2007.

Łapczyński M., *Zmienne zastępcze i konkurencyjne w interpretacji drzew klasyfikacyjnych CART*, Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie (w druku).

Steinberg D., Cardell N.S., *The hybrid CART-logit model in classification and data mining*, www.salford-systems.com (10.01.2008).

Steinberg D., Colla P., *CART. Interface and documentation*, Salford Systems, http://www.salford-systems.com/, 1997.

THE HYBRID CART-LOGIT MODELS IN ANALYSING MARKET DATA

Summary

The goal of this article is to compare classic approaches which are used in building predictive models with hybrid methods that combine CART algorithm with logistic regression. The predictive model concerning the stated preferences of consumers on the motor vehicles market was developed on the basis of data collected during the survey research conducted in 2005. The list of independent variables consisted of demographic variables, LOV scale and items from the scale known as benefit structure analysis. Classic data mining and statistical tools (CART, RandomForests and logistic regression) were compared with three types of hybrid models and 2 two-stage approaches. The performance of predictive models was measured with percent of properly classified cases, odds ratio and R^2 measure proposed by McFadden.