PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 27 — 2008
Ekonometria 22

Piotr Michalski

ZASTOSOWANIE WYBRANYCH ESTYMATORÓW MODELU REGRESJI LOGISTYCZNEJ W CREDIT SCORINGU

1. Wstęp

Zadanie oceny zdolności kredytowej rozważa się często w kontekście problemu klasyfikacji pod nadzorem, co oznacza w istocie uznanie tożsamości pojęć metody credit scoringu i metody klasyfikacji pod nadzorem (zob. np. [7; 8]). Umożliwia to zastosowanie w problemach credit scoringu popularnego liniowego modelu regresji logistycznej. W artykule zostanie empirycznie zweryfikowana przydatność w zastosowaniach credit scoringu bardziej zaawansowanych podejść do estymacji modelu logitowego: addytywnego modelu regresji logistycznej oraz boostingu drzew klasyfikacyjnych.

2. Credit scoring jako zadanie klasyfikacji pod nadzorem

W artykule przyjęto założenie, że doświadczenia banku ze współpracy z klientami pozwalają na wyróżnienie dwóch klas klientów — „dobrych" i „złych". Wtedy zadanie credit scoringu odpowiada problemowi binarnej klasyfikacji pod nadzorem. W ogólności, w klasyfikacji pod nadzorem przyjmuje się istnienie stochastycznej zależności między wielokategorialną cechą objaśnianą Y, której numery kategorii zakodowane są np. jako zbiór $G=\{1,2,\ldots,g\}$, oraz wektorem cech objaśniających $\mathbf{X}=\left[X_1 \;\; X_2 \;\ldots\; X_p\right]$. Wówczas modelowane są prawdopodobieństwa warunkowe $p(Y=k|\mathbf{X})=p_k(\mathbf{X}) \;\; k=1,2,\ldots,g$ lub ich monotoniczne transformacje. Klasyfikacja obserwacji z klasy i do klasy j powoduje stratę (koszt) $L(i,j)$, przy czym zwykle zakłada się, że $L(i,j)\leq 0$, natomiast $L(i,j)>0$

$\forall i, j \in G$. Funkcję strat L można przedstawić za pomocą macierzy strat $\left[L(i,j)\right]_{g\times g}$. Funkcją decyzyjną, minimalizującą oczekiwany błąd predykcji w próbie testowej, jest klasyfikator bayesowski

$$d^*(x) = \arg\min_k \sum_{i=1}^{g} L(i,k)p_i(x), \tag{1}$$

gdzie, zgodnie z twierdzeniem Bayesa, $p_i(x) = \pi_i p(x|i) \Big/ \sum_{r=1}^{g} \pi_r p(x|r)$ oraz $\pi_i = P(Y=i)$, $i = 1, 2, \ldots, g$. Zadaniem zatem jest oszacowanie na podstawie dostępnej próby uczącej $T = \{x_i, y_i\}_1^n$ prawdopodobieństw $p_i(x)$ i podstawienie ich do wzoru (1). W ten sposób uzyskuje się oszacowanie klasyfikatora bayesowskiego, które dalej będzie oznaczane jako $\hat{d}(x)$. Prawdopodobieństwa $p_i(x)$ mogą być bezpośrednio estymowane za pomocą prezentowanych w artykule metod regresji logistycznej.

W rozpatrywanym przypadku dwóch klas klientów banków zbiór G przyjmie postać zbioru {0; 1} dla liniowego i addytywnego modelu logitowego oraz {–1; 1} dla modelu boostingu drzew klasyfikacyjnych. Tak zakodowane kategorie nie tylko będą identyfikować klasy, lecz staną się również przedmiotem obliczeń jako wartości liczbowe.

3. Wybrane modele regresji logistycznej

Model regresji logistycznej można zapisać ogólnie jako funkcję logitową pewnej funkcji warunkowej wartości oczekiwanej cechy Y, wyrażonej jako funkcja wektora cech $\mathbf{X}$:

$$\operatorname{logit}\left[g\left(E\left(Y|\mathbf{X}\right)\right)\right] = h(\mathbf{X}), \tag{2}$$

gdzie: $\operatorname{logit}(x) = \ln\left[1/(1-x)\right]$,

$g(\bullet)$ – pewna monotoniczna funkcja,

$h(\mathbf{X})$ – predyktor.

W teorii uogólnionych modeli liniowych zakłada się, że obserwacje zmiennej objaśnianej pochodzą z rozkładu należącego do rodziny wykładniczych rozkładów prawdopodobieństwa. W przypadku rozpatrywanego problemu klasyfikacji pod nadzorem, w którym występują mikrodane, a kodowanie kategorii jest binarne, cecha Y posiada należący do rodziny rozkładów wykładniczych rozkład zero-jedynkowy, a logitowa funkcja wiążąca jest kanoniczna. W tab. 1 przedstawiono sposób zapisu liniowego modelu regresji logistycznej (GLM-Logit), addytywnego

Tabela 1. Charakterystyki trzech modeli regresji logistycznej

Model	Kodowanie kategorii cechy Y	$E(Y\|\mathbf{X})$	$g(E(Y\|\mathbf{X}))$	$h(\mathbf{X})$
GLM-Logit	{0, 1}	$P(Y=1\|\mathbf{X})$	$E(Y\|\mathbf{X})$	$\alpha_0+\sum_{i=1}^{p}\alpha_i X_i$
GAM-Logit	{0, 1}	$P(Y=1\|\mathbf{X})$	$E(Y\|\mathbf{X})$	$\alpha_0+\sum_{i=1}^{m}\alpha_i X_i+\sum_{j=m+1}^{p} f_j(X_j)$
Real AdaBoost	{−1, 1}	$2P(Y=1\|\mathbf{X})-1$	$[E(Y\|\mathbf{X})+1]/2$	$2\sum_{m=1}^{M} d_m(x)$

Źródło: opracowanie własne.

modelu regresji logistycznej (GAM-Logit) oraz modelu boostingu drzew klasyfikacyjnych Real AdaBoost w konwencji wzoru (2).

W tab. 1 indeksowane wielkości α to pewne parametry, $d(x)$ oznaczają klasyfikatory, natomiast $f_j(\bullet)$ to pewne nieparametryczne, gładkie funkcje. Rozwiązując równanie (2) względem $p_1(x)$, uzyskuje się prawdopodobieństwa *a posteriori* przynależności do klas, wykorzystywane we wzorze (1):

$$p_1(x)=\frac{\exp[h(\mathbf{X})]}{1+\exp[h(\mathbf{X})]}. \tag{3}$$

Model GLM-Logit stał się podstawowym narzędziem analizy regresji w przypadku jakościowej cechy objaśnianej. Z postaci podanej w tab. 1 wynika, że w modelu tym logit prawdopodobieństwa *a posteriori* przynależności do klasy 1 modelowany jest za pomocą tradycyjnego, liniowego predyktora. Uogólniony model addytywny jest rozszerzeniem tego powszechnie wykorzystywanego w credit scoringu narzędzia. W modelu GAM-Logit zakłada się *explicite*, że cechy X_1, X_2, ..., X_m posiadają formę liniowego, parametrycznego predyktora, co może być podyktowane jakościowym bądź dyskretnym charakterem cech X_1, X_2, ..., X_m, albo uprzednią wiedzą o liniowych efektach tych zmiennych. Z kolei ilościowe cechy X_{m+1}, ..., X_p dopasowywane są technikami nieparametrycznymi w nadziei odkrycia ich nieliniowych efektów. Najczęściej wykorzystywane metody nieparametrycznej estymacji jednowymiarowej funkcji regresji $f(\bullet)$ to technika wygładzonej funkcji sklejanej (*smoothing spline*) oraz technika lokalnej regresji, znana w anglojęzycznej literaturze jako *local regression* lub *loess* (zob. np. [2; 6]).

Uogólnione modele addytywne znajdują zastosowanie w wielu problemach klasyfikacji pod nadzorem. Wydaje się, że ocena zdolności kredytowej jest jednym z naturalnych pól zastosowań, szczególnie gdy zbiór danych obejmuje obserwacje cech zarówno ilościowych, jak i jakościowych. Modelowanie nieliniowe cech

ilościowych bowiem pozwala odkrywać w danych nowe wzorce. Możliwe jest przy tym sterowanie stopniem wygładzania, co zapobiega nadmiernemu dopasowaniu do danych, przez co zachowana zostaje korzystna relacja między obciążeniem a wariancją modelu. Poza zastosowaniem predyktywnym, uogólnione modele addytywne służą jako dobre narzędzie eksploracyjne. Uogólniony model addytywny wykorzystuje część narzędzi wnioskowania o uogólnionych modelach liniowych oraz posiada nie mniej użyteczne możliwości interpretacyjne. Problemem pozostaje dobór zmiennych do modelu. Nierozwiązaną kwestią jest też procedura specyfikacji postaci zmiennych w modelu. Problematyczny jest często wybór między parametryczną a nieparametryczną reprezentacją danej cechy. Warunkiem powodzenia zastosowania modelu GAM-Logit w credit scoringu jest zatem umiejętny dobór cech modelowanych nieparametrycznie, wybór nieparametrycznego estymatora oraz ustalenie stopnia wygładzania.

Model GLM-Logit estymowany jest na ogół metodą największej wiarogodności lub jej ekwiwalentem – algorytmem IRWLS (*iteratively re-weighted least squares algorithm*). Algorytm ten zaadaptowano do estymacji modelu GAM-Logit. Krok polegający na zastosowaniu ważonej metody najmniejszych kwadratów (*weighted least squares step*) zastępowany jest procedurą ważonego algorytmu wielokrotnego dopasowania (*weighted backfitting algorithm*) (zob. np. [2; 6]).

Uogólnione modele liniowe są powszechnie oprogramowane (np. pakiet *Statistica* lub funkcja *glm* pakietu *Stats* w środowisku *R*). Model można również oszacować w dowolnym arkuszu kalkulacyjnym, maksymalizując logarytm funkcji wiarogodności (np. za pomocą narzędzia optymalizacyjnego *Solver* w arkuszu kalkulacyjnym *Excel*). Uogólnione modele addytywne oprogramowane są np. pod postacią rozbudowanego modułu w programie *Statistica Data Miner*. W środowisku *R* do wyboru jest kilka pakietów, wśród których najpopularniejszy jest pakiet *gam*. Pakiet *gam* jest implementacją algorytmu IRWLS i umożliwia wybór lokalnej regresji lub wygładzanej funkcji sklejanej w charakterze nieparametrycznego estymatora funkcji regresji.

Przedstawiony tu zostanie algorytm boosting, który został uznany za jedną z najlepszych technik statystycznego uczenia się. Algorytm ten opiera się na pomyśle łączenia decyzji wielu niezależnych klasyfikatorów, przy czym zakłada się, że pojedyncze klasyfikatory są nieco lepsze od klasyfikatora losowego. Algorytm boosting spełnia w przybliżeniu dwa powyższe warunki, dopasowując sekwencyjnie modele klasyfikacyjne do danych ważonych za każdym razem innymi wagami. Obserwacja, która w poprzedzającym kroku została błędnie zaklasyfikowana, otrzymuje większą wagę w kolejnym kroku. Najczęściej wykorzystywanym klasyfikatorem bazowym jest drzewo klasyfikacyjne, a szczególnie jego dwuliściowa wersja. Każdorazowe ważenie obserwacji wypełnia (w sposób niedoskonały) postulat niezależności klasyfikatorów, przy czym najczęściej rozważa się dwie alternatywne koncepcje ważenia:

- wbudowanie wag w mechanizm klasyfikatora,
- szacowanie klasyfikatora na podstawie pseudopróby, powstałej w wyniku losowania ze zwracaniem z próby uczącej ustalonej liczby elementów zgodnie z rozkładem prawdopodobieństwa wyznaczonym przez wagi (konieczność, gdy klasyfikator nie obsługuje obserwacji ważonych).

Należy zauważyć, że każdorazowe ważenie obserwacji nie gwarantuje całkowitej niezależności prób uczących. Mimo to algorytm boostingu prowadzi w wielu przypadkach do znacznego zmniejszenia obciążenia oraz redukcji wariancji i jest obecnie uważany za jedną z najlepszych metod klasyfikacji pod nadzorem.

Najpopularniejszą odmianą algorytmu boosting jest algorytm rzeczywistego boostingu drzew klasyfikacyjnych Real AdaBoost. Algorytm ten opublikowali w 2000 r. w podstawowej wersji Hastie, Tibshirani i Friedman (zob. [5]). Algorytm Real AdaBoost przedstawiono w tab. 2. Przyjęto, że $d_m(x)$ oznacza klasyfikator generujący wartości rzeczywiste.

Tabela 2. Algorytm Real AdaBoost

Krok algorytmu	Procedura
1	Przyjmij wagi początkowe $w_i = 1/n$, $i = 1, 2, \ldots, n$.
2	Powtarzaj dla $m = 1, 2, \ldots, M$: a) oszacuj za pomocą mechanizmu klasyfikacji pod nadzorem prawdopodobieństwa $p_m(x) = \hat{P}_w(Y = 1\|x)$, stosując do próby uczącej wagi w_i; b) podstaw $d_m(x) \leftarrow 0{,}5\ln\left[p_m(x)/(1 - p_m(x))\right]$; c) podstaw $w_i \leftarrow w_i \exp\left[-y_i d_m(x)\right]$, $i = 1, 2, \ldots, n$, i dokonaj renormalizacji (żeby $\sum_i w_i = 1$).
3	Przyjmij za decyzję klasyfikacyjną wielkość $\operatorname{sign}\left[\sum_{m=1}^{M} d_m(x)\right]$.

Źródło: [5, s. 340].

Doniosłe odkrycie Hastiego, Tibshiraniego i Friedmana (zob. [5]) dotyczące statystycznych podwalin algorytmu AdaBoost wskazuje, że algorytm ten jest w istocie pewną procedurą estymacji modelu regresji logistycznej. Rozpatrywana jest funkcja rzeczywista $J(h) = E\left[\exp(-Yh(\mathbf{X}))\right]$. Populacyjną funkcją minimalizującą $J(h)$ w punkcie x jest

$$\tilde{h}(x) = \arg\min_{h(x)}\left\{E_{Y|\mathbf{X}}\left[\exp(-Yh(x))\right]\right\} = \frac{1}{2}\ln\frac{P(Y=1|x)}{P(Y=-1|x)}, \tag{4}$$

co można wykazać, minimalizując funkcję $J(h)$ pod warunkiem $X = x$:

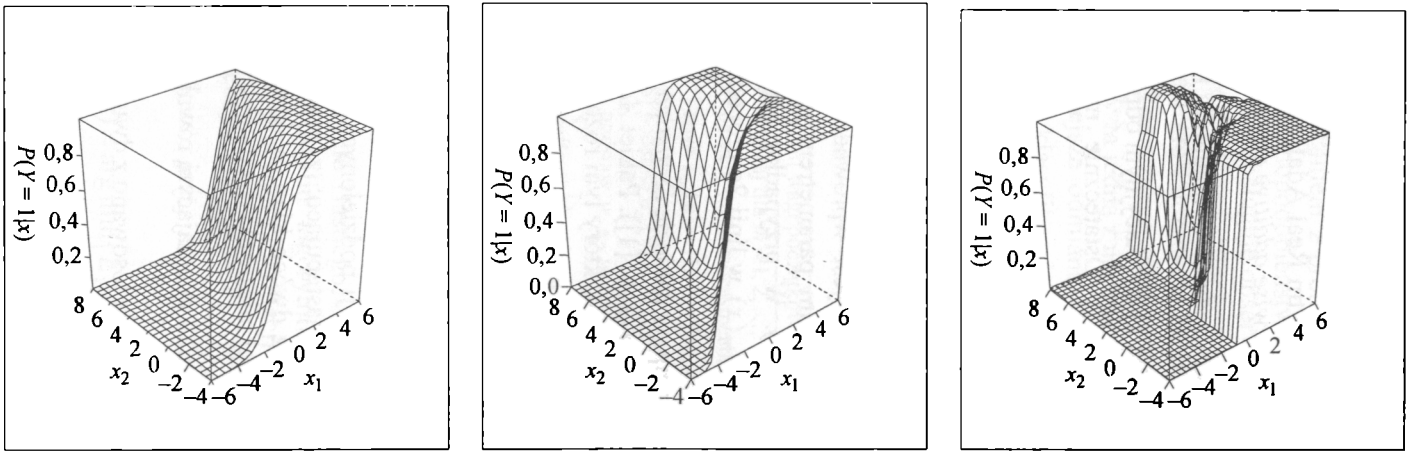
$$E\left[\exp(-Yh(x))|x\right] = P(Y=1|x)e^{-h(x)} + P(Y=-1|x)e^{h(x)}. \quad (5)$$

Obliczając pochodną funkcji (5) względem $h(x)$ i przyrównując ją do zera, otrzymuje się równanie (4). Okazuje się, że algorytm Real AdaBoost dopasowuje w modelu (4) za pomocą procedury *forward stagewise additive modelling* (zob. np. [6]) addytywną funkcję $\tilde{h}(x) = \sum_{m=1}^{M} d_m(x)$, wykorzystując wykładniczą funkcję straty $L(y, h(x)) = \exp(-yh(x))$. Wielkość $yh(x)$ jest klasyfikacyjnym odpowiednikiem reszty, gdy klasy są zakodowane jako $\{-1; 1\}$. Ostatecznie można zapisać $h(x) = 2\tilde{h}(x)$.

Metaparametrem algorytmów boostingu jest wielkość pojedynczego drzewa. Najczęściej stosowane podejście zakłada ustalenie jednakowej wielkości drzewa dla każdej iteracji. Z doświadczeń praktycznych wynika, że optymalna wielkość drzewa zawiera się zwykle w przedziale od 4 do 8 liści i rzadko przekracza 10. Ważną kwestią jest też regularyzacja (ustalenie optymalnej złożoności modelu). W boostingu jednym ze sposobów regularyzacji jest wprowadzenie parametru spowalniającego proces uczenia algorytmu. Takim parametrem jest tzw. stopa szybkości uczenia się $v \in (0; 1]$ (*learning rate*). W przypadku algorytmu Real AdaBoost przez v mnożone jest oszacowanie $d_m(x)$ w linii 2c procedury. Zwykle ustalana jest mała wartość v, np. 0,1.

Algorytmy boostingu oprogramowane są np. w pakiecie *Statistica Data Miner*. W środowisku *R* najpopularniejszy pakiet to *Ada* (zob. [1]). Pakiet *Ada* jest implementacją stochastycznego algorytmu Real AdaBoost, który tym różni się od opisanego, że trenowanie klasyfikatora (linia 2a) odbywa się na podstawie podpróby próby uczącej, powstałej w wyniku losowania z niej bez zwracania nieco mniejszej od n liczby obserwacji. Zamysłem tego zabiegu jest wprowadzenie do algorytmu elementu losowości i obniżenie oczekiwanego błędu predykcji w próbie testowej.

Do celów ilustracyjnych na rys. 1 zaprezentowano oszacowanie prawdopodobieństwa (3) uzyskane za pomocą trzech modeli dla sztucznie wygenerowanej próby uczącej. Próba ta składa się z tysiąca obserwacji podzielonych na dwie klasy. Mechanizm generujący dane przedstawiał się następująco: z prawdopodobieństwem 0,5 przydzielano obserwacje do jednej z dwóch klas. W klasie pierwszej ($Y = 1$) obserwacje pochodzą z dwuskładnikowej mieszanki rozkładów normalnych $N\left(\begin{bmatrix}2\\0\end{bmatrix}, \begin{bmatrix}1 & 1\\1 & 3\end{bmatrix}\right)$, $N\left(\begin{bmatrix}4\\4\end{bmatrix}, \begin{bmatrix}1 & -1\\-1 & 2\end{bmatrix}\right)$ o współczynniku mieszającym równym 0,5, natomiast w drugiej klasie ($Y = 0$ lub $Y = -1$) obserwacje losowano z dwuwymiarowego rozkładu normalnego $N\left(\begin{bmatrix}0\\2\end{bmatrix}, \begin{bmatrix}4 & 2\\2 & 2\end{bmatrix}\right)$.



Rys. 1. Hiperpowierzchnie logistyczne $\hat{p}_1(x)$ uzyskane modelami GLM-Logit, GAM-Logit oraz Real AdaBoost

Źródło: opracowanie własne wykonane w środowisku R.

Przykłady zastosowania opisanych metod do rzeczywistych zbiorów danych kredytowych zostaną przedstawione w kolejnym punkcie artykułu.

4. Zastosowanie modeli do danych rzeczywistych

Zbiór danych pod nazwą German Credit pochodzi z ogólnodostępnej internetowej bazy danych[1]. Zbiór ten obejmuje obserwacje 20 cech objaśniających pochodzących z 1000 wniosków kredytowych jednego z dużych banków południowych Niemiec. Z ogólnej liczby wniosków 300 to aplikacje klientów uznanych za „złych", a 700 to wnioski klientów „dobrych". Nie wiadomo, czy próba ucząca oddaje rzeczywisty stosunek liczebności dwóch klas klientów banku. Zwykle odsetek klientów mających problemy ze spłatą kredytu jest znacznie mniejszy niż 30%, przeto można wnioskować, że próba ucząca nie jest reprezentatywna. Wśród predyktorów znajduje się 7 cech mierzalnych i 13 cech jakościowych, o liczbach kategorii wynoszących od 2 do 11. Cechy oznaczono symbolami od X_1 do X_{20}. Przyjęto następujący sposób kodowania cechy objaśnianej: $Y=1$, gdy klient jest „dobry", $Y=0$ (lub $Y=-1$) w przeciwnym wypadku. Wśród cech mierzalnych zmienne X_8, X_{11}, X_{16} oraz X_{18} są dyskretne, natomiast cechy X_2 (okres spłaty), X_5 (kwota kredytu) oraz X_{13} (wiek) mają charakter ciągły i tym samym mogą być wykorzystane w nieliniowej części semiparametrycznego modelu GAM-Logit. Opis cech objaśniających zawiera tab. 3.

Zbiór German Credit jest również analizowany w publikacjach [3] i [4]. W pracy [3] zastosowano model regresji logistycznej, drzewo klasyfikacyjne typu CART i jedną z odmian sieci neuronowej. Z kolei w monografii [4] do budowy modelu scoringowego wykorzystano liniowy model prawdopodobieństwa oraz GLM-Logit. W niniejszym artykule zastosowano trzy omawiane modele regresji logistycznej.

Modele ekonometryczne GLM-Logit oraz GAM-Logit wymagają kodowania cech jakościowych i wstępnego doboru zmiennych. Zabiegi te nie są koniecznością w przypadku bazującego na modelu drzew klasyfikacyjnych algorytmu Real AdaBoost. Do wstępnego określenia istotności predyktorów posłużył wskaźnik relatywnej ważności cech (zob. [6]) uzyskany za pomocą algorytmu Real AdaBoost na podstawie oryginalnej próby uczącej (100 iteracji, $\nu = 0{,}1$). Algorytm wskazał cechę X_{20} (pracownik zagraniczny) jako najmniej istotną w zbiorze predyktorów. Cecha ta charakteryzuje się ponadto małą zmiennością, dlatego zdecydowano się na jej wyłączenie z analizy. Na potrzeby analizy ekonometrycznej zakodowano cechy jakościowe i w rezultacie otrzymano zbiór 19 zmiennych zero-jedynkowych (binarnych). Do zbadania zależności cechy objaśnianej Y i zmiennych binarnych zastosowano test niezależności chi-kwadrat. Z analizy wyłączono zmienne binarne nieistotnie skojarzone z cechą objaśnianą Y. Ostatecznego doboru

[1] http://www.niaad.liacc.up.pt/old/statlog (maj 2007).

Tabela 3. Opis predyktorów zbioru German Credit

Cecha	Opis	Rodzaj cechy	Liczba kategorii
X_1	Stan obecnego rachunku rozliczeniowego	Jakościowa	4
X_2	Okres spłaty w miesiącach	Ilościowa	–
X_3	Historia kredytowa	Jakościowa	5
X_4	Przeznaczenie kredytu	Jakościowa	11
X_5	Kwota kredytu	Ilościowa	–
X_6	Konto oszczędnościowe/obligacje	Jakościowa	5
X_7	Okres obecnego zatrudnienia	Jakościowa	5
X_8	Wielkość raty w procentach dochodu rozporządzalnego	Ilościowa	–
X_9	Płeć i stan cywilny	Jakościowa	5
X_{10}	Inni dłużnicy/żyranci	Jakościowa	3
X_{11}	Okres obecnego zamieszkania w latach	Ilościowa	–
X_{12}	Majątek	Jakościowa	4
X_{13}	Wiek kredytobiorcy w latach	Ilościowa	–
X_{14}	Inne zobowiązania ratalne	Jakościowa	3
X_{15}	Mieszkanie	Jakościowa	3
X_{16}	Liczba obecnie spłacanych kredytów w tym banku	Ilościowa	–
X_{17}	Zatrudnienie	Jakościowa	4
X_{18}	Liczba osób na utrzymaniu	Ilościowa	–
X_{19}	Telefon	Jakościowa	2
X_{20}	Pracownik zagraniczny	Jakościowa	2

Źródło: opracowanie własne na podstawie German Credit Data/Dataset Description.

zmiennych dokonano za pomocą procedury postępującej regresji krokowej (w ramach modelu GLM-Logit). Spośród trzech cech ciągłych cecha oznaczająca wiek okazała się nieistotna.

Kolejną kwestią jest ustalenie semiparametrycznej struktury modelu GAM-Logit. W tab. 4 zestawiono wybrane statystyczne charakterystyki konkurencyjnych modeli GLM-Logit oraz GAM-Logit z trzema kombinacjami nieparametrycznych form zmiennych X_2 oraz X_5: *deviance* – odległość między modelem nasyconym a danym; df – liczba stopni swobody; pseudo R^2 – mierzony w kategoriach procentu objaśnienia wariancji zmiennej objaśnianej (nie jest to poprawna miara w przypadku modeli GLM-Logit i GAM-Logit); *deviance explained* – liczba jeden pomniejszona o iloraz *deviance* danego modelu i *deviance* modelu zawierającego wyłącznie wyraz wolny; AIC – wartość kryterium informacyjnego Akaikego. Charakterystyki podano po uśrednieniu dziesięciu realizacji uzyskanych w próbach uczących, losowo pobranych z próby pierwotnej (70% obserwacji), natomiast oszacowania oczekiwanego błędu predykcji (Err) otrzymano po uśrednieniu oszacowań błędu w pozostałych częściach prób (próby testowe, 30% obserwacji). W tab. 4 zawarto również wartość Err wyznaczoną na podstawie całej próby uczącej (resub-

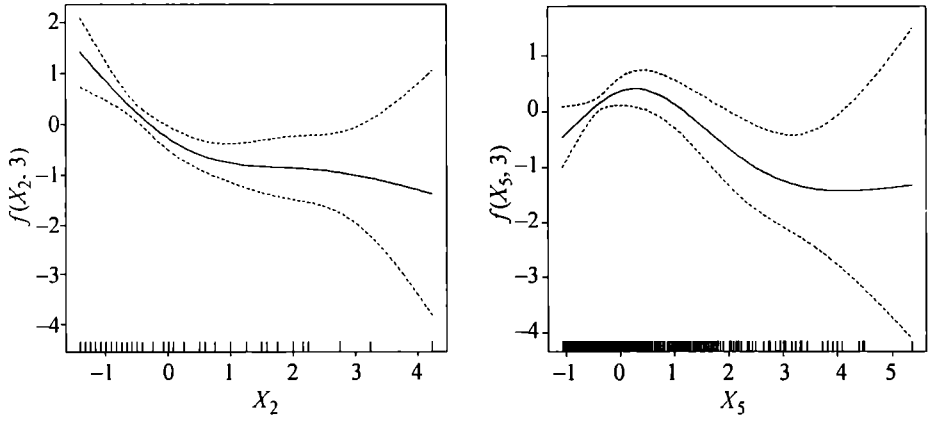
stytucja). Przyjęto zero-jedynkową postać funkcji straty oraz 3 stopnie swobody dla funkcji sklejanych modelu GAM-Logit.

Tabela 4. Zestawienie statystycznych charakterystyk modeli GLM-Logit oraz trzech semiparametrycznych wariantów modelu GAM-Logit (wielkości uśrednione)

Charakterystyki statystyczne modelu	GLM-Logit	GAM-Logit nieparametryczny w:		
		X_2	X_5	X_2, X_5
df	687	684	684	681
Deviance	712,4	708,5	703,3	697,6
AIC	738,4	740,5	735,3	735,6
Pseudo R^2	0,217	0,217	0,225	0,227
Deviance explained	19,1%	19,2%	20,1%	20,4%
p-value X_2	0,000	0,099	0,000	0,027
p-value X_5	0,070	0,079	0,024	0,004
Err Resubstytucja		0,233	0,237	0,235
Err Próba testowa	0,249	0,248	0,246	0,242

Źródło: opracowanie własne wykonane w środowisku R.

Przyjęcie semiparametrycznej postaci zmiennej X_5 lub zmiennych X_2 i X_5 poprawia nieznacznie parametry statystyczne oraz wyniki klasyfikacji. Ostrożnie należy traktować wielkości *p-value*, które w przypadku modelu GAM-Logit ze zmienną nieparametryczną X_2 lub X_5 odnoszą się do różnych testów. Na rys. 2 przedstawiono wykresy cząstkowej predykcji (tj. wykresy funkcji $\hat{f}_j(\bullet)$) modelu nieparametrycznego w zmiennych X_2 oraz X_5.



Rys. 2. Wykresy cząstkowej predykcji zmiennych X_2 oraz X_5 z przedziałem ufności 95%

Źródło: opracowanie własne wykonane w środowisku R.

Po analizie wykresów cząstkowej predykcji oraz charakterystyk konkurencyjnych modeli wskazano jako najlepszy model GAM-Logit nieparametryczny w zmiennych X_2 oraz X_5.

Z wykresu cząstkowej predykcji dla zmiennej X_2 wynika, że ryzyko niewypłacalności rośnie (coraz wolniej) wraz z wydłużaniem się okresu spłaty kredytu, *ceteris paribus*. Z kolei wykres cząstkowej predykcji zmiennej X_5 ujawnia interesującą zależność, mianowicie do pewnej kwoty kredytu ryzyko niewypłacalności maleje, a po jej przekroczeniu wzrasta, *ceteris paribus*. Można innymi słowy powiedzieć, że kredyty o średniej wartości są mniej ryzykowne od kredytów o niskiej i wysokiej kwocie.

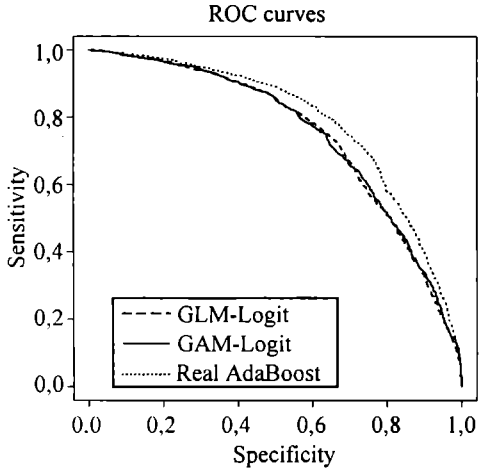
W tab. 5 porównano wyniki klasyfikacji z zastosowaniem modeli GLM-Logit, GAM-Logit nieparametrycznego w zmiennych X_2 oraz X_5 z trzema stopniami swobody oraz algorytmu Real AdaBoost o parametrze $v = 0{,}1$ i 100 iteracjach, opartego na oryginalnym zbiorze cech z wyłączeniem cechy X_{20}. Dodatkowo w tab. 5 uwzględniono drzewo klasyfikacyjne typu CART, będące klasyfikatorem bazowym algorytmu Real AdaBoost. Oszacowania oczekiwanego błędu predykcji podano po uśrednieniu (dziesięciokrotnie losowano z próby uczącej próbę treningową – 70% obserwacji i próbę testową – 30% obserwacji). Przyjęto zero-jedynkową funkcję straty. Obliczono ponadto oszacowania prawdopodobieństw błędnej klasyfikacji (I i II rodzaju).

Tabela 5. Wyniki klasyfikacji pod nadzorem (wielkości uśrednione)

Model	Resubstytucja	Próba testowa (30%)	Oszacowanie prawdopodobieństwa błędu I rodzaju	Oszacowanie prawdopodobieństwa błędu II rodzaju
GLM-Logit	0,235	0,256	0,592	0,112
GAM-Logit	0,233	0,249	0,588	0,103
CART	0,219	0,276	0,544	0,160
Real AdaBoost	0,106	0,237	0,602	0,081

Źródło: opracowanie własne wykonane w środowisku R.

Należy zauważyć, że dane German Credit stanowią dość trudny problem klasyfikacji pod nadzorem. Uzyskane wyniki są bowiem tylko nieco lepsze od metody polegającej na przydzielaniu wszystkich obserwacji do liczniej reprezentowanej klasy. Najlepsze ogólne wyniki klasyfikacyjne uzyskano za pomocą algorytmu Real AdaBoost oraz modelu GAM-Logit. Porównanie metod dla różnych punktów odcięcia (wykres krzywych operacyjno-charakterystycznych ROC) przedstawiono na rys. 3. Punkty (*specificity*, *sensitivity*), tj. dopełnienia do liczby jeden odpowiednio oszacowań prawdopodobieństw popełnienia błędów I i II rodzaju, zostały dziesięciokrotnie uśrednione. Im bardziej krzywa ROC ciąży ku prawemu górnemu rogowi wykresu, tym lepszymi właściwościami predykcyjnymi charakteryzuje się



Rys. 3. Krzywe operacyjno-charakterystyczne liniowego modelu regresji logistycznej (GLM-Logit), addytywnego modelu logistycznego (GAM-Logit) oraz algorytmu Real AdaBoost

Źródło: opracowanie własne wykonane w środowisku R.

rozpatrywany klasyfikator. Algorytm Real AdaBoost dał najlepsze rezultaty dla niemal każdego punktu odcięcia.

Drugi zbiór danych – Australian Credit – pochodzi z tego samego repozytorium co wcześniej rozpatrywany zbiór German Credit. Zbiór składa się z 690 obserwacji 14 cech objaśniających oraz etykiety klasy. Próba ucząca obejmuje 307 obserwacji jednej klasy oraz 383 klasy drugiej. Wśród predyktorów ryzyka znajduje się 6 cech mierzalnych i 8 cech jakościowych, o liczbach kategorii od 2 do 14. Dane związane są ze sferą usług kart kredytowych, jednak ze względów poufności utajnione zostały nazwy cech objaśniających, opisy kategorii cech jakościowych, a do tego nieznana jest interpretacja etykiety klas. Nie jest zatem możliwe nadanie interpretacji ekonomicznej procesowi budowy klasyfikatora. W związku z tym przedstawione zostaną tylko najważniejsze wyniki.

Zmienną – kandydatką do nieparametrycznej reprezentacji w modelu GAM-Logit – była zmienna X33. Efekt wprowadzenia zmiennej X33 do modelu GLM-Logit i GAM-Logit podsumowuje tab. 6. Wyniki uzasadniają nieparametryczne wprowadzenie zmiennej X_{33}.

Ostateczne wyniki klasyfikacji zaprezentowano w tab. 7. Wyniki te uzyskano zgodnie z procedurą opisaną przy analizie zbioru German Credit. Ponownie zwracają uwagę najlepsze rezultaty algorytmu Real AdaBoost oraz dobre wyniki modelu GAM-Logit z nieparametrycznie modelowaną zmienną X_{33}.

Tabela 6. Zestawienie statystycznych charakterystyk modeli GLM-Logit oraz GAM-Logit nieparametrycznego w zmiennej X_{33} (wielkości uśrednienione)

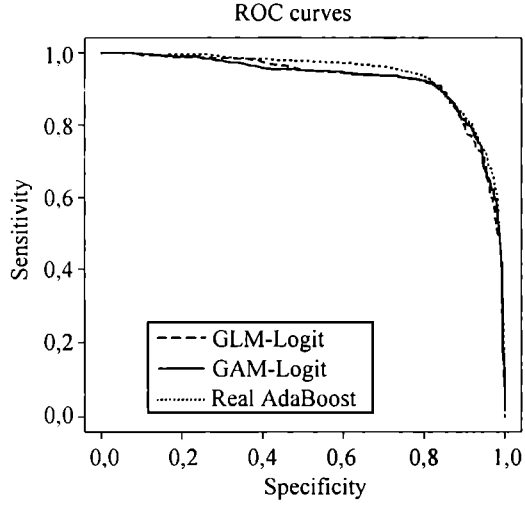
Charakterystyki statystyczne modelu	GLM-Logit bez X_{33}	GLM-Logit z X_{33}	GAM-Logit nieparametryczny w X_{33}
df	467	466	463
Deviance	311,20	306,24	298,89
AIC	329,19	326,24	324,89
Pseudo R^2	0,590	0,601	0,604
Deviance explained	52,4%	53,1%	53,8%
p-value X_{33}	–	0,0725	0,0275

Źródło: opracowanie własne wykonane w środowisku R.

Tabela 7. Wyniki klasyfikacji pod nadzorem (wielkości uśrednione)

Model	Resubstytucja	Próba testowa (30%)	Oszacowanie prawdopodobieństwa błędu I rodzaju	Oszacowanie prawdopodobieństwa błędu II rodzaju
GLM-Logit	0,130	0,137	0,165	0,103
GAM-Logit	0,131	0,133	0,158	0,101
CART	0,101	0,152	0,116	0,198
Real AdaBoost	0,044	0,128	0,126	0,130

Źródło: opracowanie własne wykonane w środowisku R.



Rys. 4. Krzywe operacyjno-charakterystyczne liniowego modelu regresji logistycznej (GLM-Logit), addytywnego modelu logistycznego (GAM-Logit) oraz algorytmu Real AdaBoost

Źródło: opracowanie własne wykonane w środowisku R.

Na rys. 4 zamieszczono wykres krzywych operacyjno-charakterystycznych dla danych Australian Credit.

Widoczna jest ogólna przewaga algorytmu boosting oraz addytywnego modelu regresji logistycznej nad modelem liniowym dla punktów odcięcia, przy których specyficzność i czułość przyjmują jednocześnie wysokie wartości.

Przedstawione przykłady zastosowań pokazują, że uogólnione modele addytywne oraz modele boostingu są interesującym uzupełnieniem/rozszerzeniem liniowej analizy regresji logitowej w problemach credit scoringu. Odpowiednio regularyzowany model GAM-Logit jest ciekawym narzędziem pogłębiającym możliwości interpretacyjne modelu klasyfikacyjnego. Nieparametryczne modelowanie wybranych cech ilościowych pozwala często zastąpić dyskretyzację cechy (w celu uwzględnienia nieliniowości) w ramach liniowego modelu GLM-Logit. Dobrą metodą dokładnej estymacji prawdopodobieństw na potrzeby indywidualnej wyceny kredytu wydaje się z kolei algorytm Real AdaBoost. Boosting redukuje obciążenie i wariancję klasyfikatora bazowego, osiągając zarówno w przedstawionych przykładach, jak i w większości problemów klasyfikacyjnych znakomite wyniki (zob. np. [5; 6]).

Literatura

[1] Culp M., Johnson K., Michailidis G., *Ada: An R Package for Stochastic Boosting*, „Journal of Statistical Software", vol. 17, issue 2, October 2006.

[2] Faraway J.J., *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Chapman & Hall/CRC Press, London 2006.

[3] Giudici P., *Applied Data Mining: Statistical Methods for Business and Industry*, John Wiley & Sons, New York 2003.

[4] Gruszczyński M., *Modele i prognozy zmiennych jakościowych w finansach i bankowości*, SGH, Warszawa 2002.

[5] Hastie T., Tibshirani R., Friedman J., *Additive Logistic Regression: a Statistical View of Boosting*. „The Annals of Statistics" 2000, vol. 28, no. 2, s. 337-407.

[6] Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, New York 2003.

[7] Mester L.J., *What's the Point of Credit Scoring*, „Business Review", September/October 1997, Federal Reserve Bank of Philadelphia.

[8] Migut G., Wątroba J., *Scoring kredytowy a modele data mining*, „Ryzyko w Instytucji Finansowej" 2005 nr 1.

APPLICATION OF SELECTED ESTIMATORS OF LOGISTIC REGRESSION MODELS IN CREDIT SCORING

Summary

The article examines the application opportunities of different logistic regression models in a credit scoring supervised classification problem. The paper covers linear and generalized additive logistic regression model, as well as a classification trees boosting method – Real AdaBoost. The empirical study of two real credit datasets is given.

Piotr Michalski – mgr, doktorant w Katedrze Ekonometrii Uniwersytetu Ekonomicznego we Wrocławiu.