

Michał Steuden

Katolicki Uniwersytet Lubelski

SELEKCJA KREDYTOBIORCÓW PRZEPROWADZONA ZA POMOCĄ MODELU PROBITOWEGO JAKO FUNKCJI KLASYFIKACJI W DRZEWIE DECYZYJNYM

1. Wstęp

Problem selekcji kredytobiorców pod względem prawdopodobieństwa spłacenia kredytu ma zasadnicze znaczenie w polityce udzielania kredytów. Bank musi mieć narzędzia, dzięki którym na podstawie cech kredytobiorcy może zdecydować o przyznaniu (bądź nie) kredytu. Spośród wielu dostępnych metod klasyfikacji (por. [4]), model probitowy umożliwia generowanie prawdopodobieństw spłaty za pomocą symulacji Monte Carlo i, co jest równie ważne, łatwą kalibrację modelu do zmieniających się warunków kredytowych. Jednak sposób tworzenia modelu praktycznie uniemożliwia wykorzystanie zmiennych niemierzalnych (następuje utrata ważnych informacji), co prowadzi do obniżenia jakości prognostycznej modelu. Prezentowane podejście ma na celu wykorzystanie funkcji probitowej jako elementu decyzyjnego drzewa klasyfikacyjnego, umożliwiając poprawienie jakości modelu.

2. Model probitowy

Do najczęściej spotykanych modeli należą modele różnicujące grupy za pomocą funkcji, której argumentami są cechy kredytobiorców. W taki sposób zbudowany jest model probitowy [2, s. 665; 5; 6], który ponadto daje jednoznaczną odpowiedź, do jakiej grupy zaliczyć pewien wektor wartości cech klienta banku.

Załóżmy, że decyzję o przyznaniu (lub nie) kredytu można opisać za pomocą następującego modelu:

$$D_j' = \mathbf{x}_j \mathbf{a} + \xi_j, \quad (1)$$

gdzie D' oznacza zmienną decyzyjną (decyzja o przyznaniu (bądź nie) kredytu opiera się na wielkości zmiennej binarnej $D_j = I_{\{D_j \geq 0\}}$), $\mathbf{x}_j$ oznacza realizację wektora k zmiennych opisowych kredytobiorcy j ($j = 1, \dots, m$), $\boldsymbol{\alpha}$ – wektor parametrów równania, ξ jest błędem modelu – zakładamy, że ma on rozkład $N(0,1)$.

Estymatory parametrów równania można wyznaczyć za pomocą metody największej wiarygodności. Funkcję prawdopodobieństwa zapisujemy w następujący sposób:

$$l = \prod_{D_j \geq 0} pr(D_j' \geq 0) \cdot \prod_{D_j < 0} pr(D_j' < 0)^1, \quad (2)$$

$$l = \prod_{j=1}^m pr(D_j' \geq 0)^{D_j} \cdot \prod_{j=1}^m pr(D_j' < 0)^{1-D_j}, \quad (3)$$

po zlogarytmowaniu zaś otrzymujemy następującą postać:

$$\ln l = \sum_{j=1}^m D_j \cdot \ln(pr(\xi_j \geq -\mathbf{x}_j \boldsymbol{\alpha})) + \sum_{j=1}^m (1 - D_j) \cdot \ln(pr(\xi_j < -\mathbf{x}_j \boldsymbol{\alpha})). \quad (4)$$

Biorąc pod uwagę to, że ξ ma rozkład $N(0,1)$, powyższe równanie możemy zapisać w następujący sposób:

$$\ln l = \sum_{j=1}^m D_j \cdot \ln(\Phi(\mathbf{x}_j \boldsymbol{\alpha})) + \sum_{j=1}^m (1 - D_j) \ln(1 - \Phi(\mathbf{x}_j \boldsymbol{\alpha})). \quad (5)$$

W celu zmaksymalizowania powyższej funkcji można skorzystać z algorytmu wyszukiwania ekstremum Newtona-Raphsona i jego różnych modyfikacji, np.: Davidona-Fletcher-Powella [3].

Gradienty powyższej funkcji można zapisać w następujący sposób [2, s. 672]:

$$\frac{\partial \ln l}{\partial \boldsymbol{\alpha}} = \sum_{j=1}^m \left[\frac{q_j \cdot \phi(q_j \mathbf{x}_j' \boldsymbol{\alpha})}{\Phi(q_j \mathbf{x}_j' \boldsymbol{\alpha})} \right] \cdot \mathbf{x}_j = \sum_{j=1}^m \lambda_j \mathbf{x}_j, \quad (6)$$

gdzie $q_j = 2D_j - 1$.

Z kolei macierz wariancji-kowariancji może być wyliczona jako estymator White'a:

$$Var(\boldsymbol{\alpha}) = \mathbf{H}^{-1} \mathbf{B} \mathbf{H}^{-1}, \quad (7)$$

gdzie $\mathbf{H}$ oznacza Hessian:

$$\mathbf{H} = \frac{\partial^2 \ln l}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}'} = \sum_{j=1}^m -\lambda_j (\lambda_j + \mathbf{x}_j' \boldsymbol{\alpha}) \mathbf{x}_j \mathbf{x}_j', \quad (8)$$

$\mathbf{B}$ oznacza estymator Berndta, Halla i Hausmana:

$$\mathbf{B} = \sum_{j=1}^m \lambda_j^2 \mathbf{x}_j \mathbf{x}_j'. \quad (9)$$

¹ Inna wersja tego równania, która uwzględnia wagi prawdopodobieństw, została zastosowana w [6].

3. Symulacja* i polityka banku

Mając tak skonstruowany model probitowy, można przystąpić do symulacji Monte Carlo. Pierwszym etapem jest zbudowanie zasady udzielania kredytu:

$$\left. \begin{array}{l} \text{udzielam} \\ \text{nie} - \text{udzielam} \end{array} \right\} \text{jeżeli} \left\{ \begin{array}{l} pr(D_j = 0) < \delta' \\ pr(D_j = 0) \geq \delta' \end{array} \right. \quad (10)$$

przy czym „delta” jest pewnym parametrem.

W każdym punkcie rozgałęzienia postępujemy według następujących kroków:

1. Dla danego δ' z przedziału (0,1) wykonaj czynności z punktów 2-5.
2. Wyznacz wartość $\tilde{D}_j'$ z rozkładu $N(\mathbf{x}_j\hat{\alpha}, \sigma_{\mathbf{x}_j\hat{\alpha}})$ dla wszystkich $j=1,2,\dots,m$, gdzie

$$\sigma_{\mathbf{x}_j\hat{\alpha}}^2 = \mathbf{x}_j \cdot \mathbf{Var}(\alpha) \cdot \mathbf{x}_j^T. \quad (11)$$

3. Aby zweryfikować scoring kredytowy pod względem reguły przyznawania kredytu, oblicz wartości oczekiwane prawdopodobieństw defaultu ze wzoru:

$$E[\tilde{p}_j] = 1 - \Phi \left(\frac{\tilde{D}_j'}{\sqrt{1 + \sigma_{\mathbf{x}_j\hat{\alpha}}^2}} \right), \quad (12)$$

następnie zastosuj regułę (10). Oznacz kredytobiorców, którym został udzielony kredyt kolejno jako 1, ..., m_A .

4. Dla kredytobiorców, którym został udzielony kredyt oblicz wartość spodziewanego zysku²:

$$Cost = \sum_{j=1}^{m_A} E[p_j] \cdot LGD_j, \quad (13)$$

$$Interest = \sum_{j=1}^{m_A} (1 - E[p_j]) \cdot RAROC_j, \quad (14)$$

$$Profit = Interest - Cost, \quad (15)$$

gdzie LGD_j oznacza procentową wartość utraty pożyczki w przypadku defaultu kredytobiorcy o numerze j , $RAROC$ zaś średni zysk z tytułu udzielenia kredytu.

5. Powtarzaj kroki 1-4 dotąd, aż uzyskasz rozkład o zadanej stabilności.

6. Określ δ' maksymalizujące zysk.

* W zdecydowanej większości opracowano na podstawie [5].

² Zawsze można zastosować tutaj inne kryterium.

4. Drzewo klasyfikacyjne

Kolejną czynnością jest budowa modelu klasyfikacyjnego³ opartego na funkcji probitowej, w której można wyróżnić następujące etapy:

1. Jesteś w gałęzi o numerze $n > 1$, przejdź do punktu 2 (jeżeli jesteś w korzeniu $n = 1$, rozpocznij budowę drzewa, przejdź do 2).
2. Dobierz zmienne do modelu tak, aby można było obliczyć macierz wariancji-kowariancji. Można tutaj zastosować techniki eliminacji zmiennych. Ważne, aby nie dopuścić do wystąpienia błędu przy odwracaniu macierzy.
3. Zbuduj równanie probitowe i przeprowadź estymację parametrów.
4. Przeprowadź symulację, która wygeneruje prawdopodobieństwa defaultu. Wybierz najlepszy zestaw kredytobiorców, opierając się na prawdopodobieństwie defaultu i zadany kryterium, i ustaw ich w gałęzi o numerze $2n$. Pozostałych kredytobiorców wprowadź do gałęzi $2n+1$.
5. Użyj wybranej funkcji do obliczenia korzyści płynących z podziału danej gałęzi. Obierz kryterium zatrzymania procesu rozgałęziania. W razie zatrzymania podziału przejdź do punktu 7.
6. Zwiększ n o jeden, przejdź do punktu 2.
7. Za pomocą obranego kryterium zbadaj stabilność drzewa (np. *cross-validation*). Obetnij zbędne rozgałęzienia (*pruning*).

5. Przykład

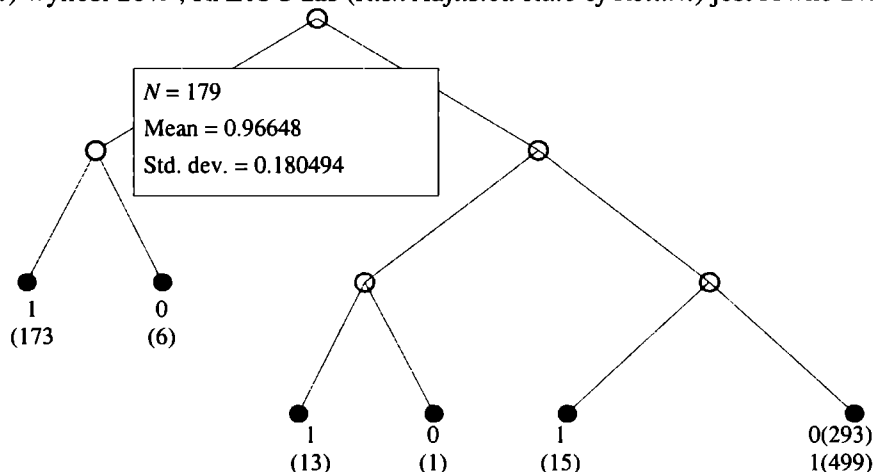
Poniżej przedstawiono wyniki działania zaproponowanego modelu drzewa decyzyjnego dla zbioru 1000 kredytobiorców (700 spłaciło kredyt, a 300 nie), zbiór ten zamieszczony był na stronie <http://www.stat.uni-muenchen.de/service/datenarchiv/>, kredytobiorcy opisani byli za pomocą 21 zmiennych, m.in. takich jak: wysokość kredytu, czas spłaty, udział raty w dochodach, przeznaczenie kredytu, historia kredytowa, informacja o innych kredytach, informacja o posiadanym koncie bankowym, wiek, płeć, stan cywilny, długość zatrudnienia w obecnym miejscu pracy, zawód itd.

Spośród podanych zmiennych wybrano takie, które są zmiennymi ilościowymi lub dadzą się zapisać w formie zmiennych ilościowych⁴. Poszczególni kredytobiorcy zostali zakwalifikowani do różnych grup (przyznaj kredyt lub odmów) w zależności od tego, jak zostali sklasyfikowani przez drzewo.

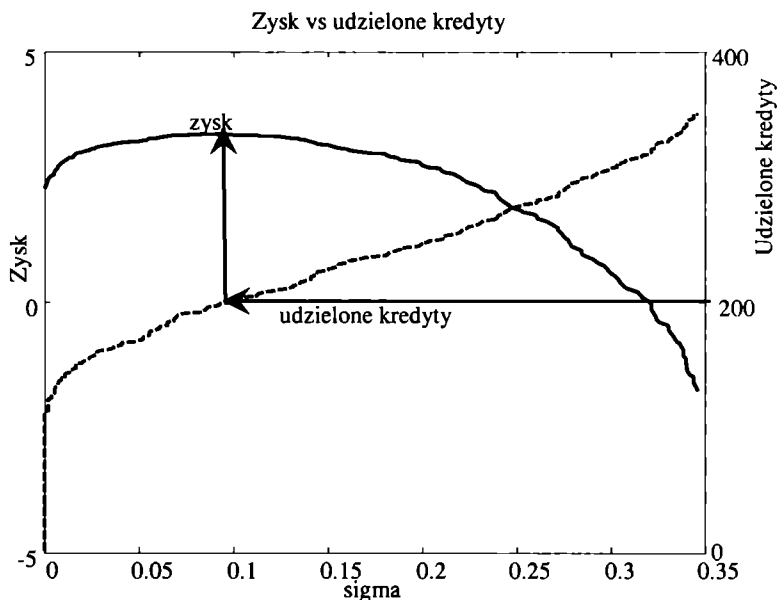
³ Najczęściej cytowaną pozycją na temat drzew klasyfikacyjnych jest [1].

⁴ Model powinien być oparty na zmiennych ilościowych. Jeżeli jednak w próbie bazy kredytowej znajdują się zmienne opisujące kredytobiorców za pomocą jakościowych miar (nie dających się bezpośrednio przełożyć na ilościowe), to można się pokusić o wykorzystanie techniki sigma restricted lub też zamieniać każdą możliwą realizację takiej zmiennej za pomocą identyfikatora (Statistica 6.0 Help: GLM Introductory Overview – Sigma-Restricted and Overparameterized Model). Możliwe jest też wpisanie modelu probitowego w drzewo klasyfikacyjne lub regresyjne.

Rysunek 1 przedstawia wyestymowany model, w którym do selekcji użyto 5 równań probitowych. Optymalizacja działania drzewa następowała w każdym z tych punktów, opierając się na następujących założeniach: LGD (*Loss Given Default*) wynosi 20% , RAROC zaś (*Risk Adjusted Rate of Return*) jest równe 2%.



Rys. 1. Drzewo klasyfikacyjne modelu probitowego przy założeniu, że LGD=20% oraz RAROC = 2%, 10x1000 symulacji
Źródło: opracowanie własne.



Rys. 2. Wygenerowany zysk z udzielonych kredytów w zależności od przyjętego sigma jako granicznego prawdopodobieństwa niespłacenia kredytu, 200x1000 symulacji
Źródło: opracowanie własne.

Zaproponowany model prawidłowo rozpoznaje próbkę 201 kredytobiorców⁵, wśród których nie było żadnego, który nie spłaciłby kredytu. Jednak każdemu z kredytobiorców (mimo że wszyscy zostali prawidłowo rozpoznani) przypisywane jest ryzyko niespłacenia długu (każdy kredytobiorca generuje stratę oraz zysk).

Rysunek 2 prezentuje możliwe do wygenerowania zyski w zależności od przyjętego progu prawdopodobieństwa defaultu jako granicznej dla udzielenia kredytu. Potwierdza się spostrzeżenie, że 200 odpowiednio dobranych osób powinno otrzymać kredyt.

Przy założeniu, że każdy kredytobiorca posiada kredyt w wysokości 1 jednostki, przynosi on:

- zysk = prawdopodobieństwo spłaty · *RAROC*,
- strata = prawdopodobieństwo defaultu · *LGD*.

Sumarycznie zysk z całej akcji kredytowej może wynieść maksymalnie 3.34 jednostki i taki wynik został osiągnięty.

6. Podsumowanie

Modele probitowe są w dużym stopniu pomijane przy budowie systemów klasyfikacji kredytobiorców. Ich wadą jest nieumiejętność wykorzystania zmiennych niemierzalnych, co może prowadzić do utraty wielu cennych informacji. Problem ten można rozwiązać na dwa sposoby – zamienić zmienne niemierzalne za pomocą odpowiednich przekształceń na zmienne mierzalne (powoduje to bardzo duży wzrost ich liczby) lub też wykorzystać taki model jako funkcję drzewa decyzyjnego i zbudować system oparty na znacznie mniejszej liczbie równań. To ostatnie podejście daje (jak się okazuje) bardzo duże możliwości:

- zachowuje główne zalety modelu: możliwość generowania prawdopodobieństw oraz łatwą kalibrację,
- pozwala wykorzystać zmienne niemierzalne poprzez ich zamianę na mierzalne (równania mogą następować po sobie i nie muszą być tak rozbudowane, przez co łatwiej pogrupować je na klasy),
- poprawia jakość klasyfikacji,
- pozwala wykorzystać metody typowe dla drzew klasyfikacyjnych do optymalizacji wielkości drzewa,
- umożliwia generowanie prawdopodobieństw, które są przydatne do zarządzania portfelami (np. CreditRisk+).

⁵ Porównaj z [4] oraz modelem dostępnym na stronie http://www.ryerson.ca/~dgrimsha/courses/cps820/Resources/rules_g.txt, którego parametry estymowane były za pomocą programu WEKA [7].

Literatura

- [1] Breiman L., Friedman J., Olshen R., *Classification and Regression Trees*, Chapman and Hall, Boca Raton 1993.
- [2] Green W., *Econometric Analysis*, Prentice Hall, Upper Saddle River, New Jersey 2003.
- [3] Hamilton J.D., *Time Series Analysis*, Princeton University Press, New Jersey 1994.
- [4] Migut G., *Modele ryzyka kredytowego*, StatSoft, Polska 2003.
- [5] Roszbach C., *Bank Lending Policy, Credit Scoring and Value at Risk*, 1998.
- [6] Steuden M., *Credit Scoring and Its Costs on the Basis of Probit Model*, [w:] Proceedings of the 26th International Scientific School: Information Systems Architecture and Technology, Wrocław 2005.
- [7] Witten I., Eibe F., *WEKA. Machine Learning Algorithms in Java*, Morgan Kaufmann Publishers, 2000.

PROBIT MODEL AS THE CART FUNCTION IN THE PROCESS OF BORROWERS' SELECTION

Summary

This article shows how to incorporate probit model into CART system in an aim to enable proper classification of potential borrowers. Probit systems are widely known but the problems they face is the level of accuracy and difficulties with treating categorical variables. Although there are techniques which incorporate such variables but at the same time the number of variables increases significantly and a covariance matrix can be difficult to estimate. Incorporation of a probit model into CART system enables to build system of probit equations with fewer explaining variables (at every split node) and with higher rate of accuracy at the same time. What is more important is that it inherits ability to predict default instances.