STUDIA INFORMATICA 2011
Volume 32 Number 2A (96)

Rafał BAJEK
Politechnika Śląska, Instytut Informatyki

WYKORZYSTANIE METOD EKSPLORACJI DANYCH DO BUDOWY MODELI SCORINGOWYCH

Streszczenie. Każda decyzja o udzieleniu kredytu obarczona jest ryzykiem. Im ryzyko jest większe, tym straty spowodowane błędną decyzją mogą być większe. Istotnie ważnym elementem jest zbadanie, czy osoba starająca się o kredyt daje szansę jego spłaty. W związku z tym, zbierane są pewne dane charakteryzujące danego kredytobiorcę, a następnie dany wniosek jest oceniany przez system scoringowy oraz ekspertów ryzyka kredytowego. Aby system scoringowy spełniał należycie swoje cele, nie może opierać się na jakiejś sztywno przyjętej teorii definicji „złych" klientów. Wykorzystując metody eksploracji danych poszukujemy pewnych wzorców w zebranych wcześniej danych, na podstawie innych wniosków kredytowych.

Słowa kluczowe: data mining, eksploracja danych, drążenie danych, scoring kredytowy

USE OF DATA MINING ALGORITHMS TO BUILD SCORING MODELS

Summary. Any decision to grant loan, it is fraught with risk. If the risk is higher, than the losses caused by an incorrect decision could be higher. Indeed, an important element is whether the person applying for a loan gives you the chance of its repayment. Consequently, collected some data which characterize the borrower and then the application is assessed by the scoring system and experts of the credit risk. To meet the scoring system due to its golas, can not be based on an accepted theory of rigid definition of "bad" clients. Using data mining methods we are looking for certain patterns in the data already collected under other loan applications.

Keywords: data mining, drilling data, credit scoring

1. Wstęp

W literaturze można spotkać wiele różnych definicji scoringu kredytowego. Na potrzeby niniejszego artykułu została przyjęta następująca: *scoring kredytowy jest metodą służącą do oceny ryzyka kredytowego, związanego z przyznaniem kredytu danemu aplikantowi* [1]. Artykuł dotyczy scoringu aplikacyjnego, czyli scoringu stosowanego dla nowych klientów. Podstawowym elementem scoringu kredytowego jest zastosowany model. Przez pojęcie model należy rozumieć narzędzie oceny i zarządzania ryzykiem związanym z indywidualnym klientem i całym portfelem kredytowym [1]. W najprostszej postaci, podczas budowy modelu scoringowego (tzw. karta scoringowa) ustala się zakres danych wejściowych, którym będzie przypisywana określona wartość liczbowa. Lista potencjalnych danych wejściowych obejmuje różnego rodzaju informacje zarówno na temat samego klienta, jak i jego otoczenia. Uzyskana wartość liczbowa, będąca sumą bądź średnią z uzyskanych przez klienta punktów (zależności od konkretnej realizacji metody) jest podstawą (wraz z wewnętrznymi przepisami banku) do zakwalifikowania danego kredytobiorcy do określonej decyzji. Wartość miernika wykorzystywana jest do podziału klientów na „dobrych" oraz „złych" i dlatego nazywana jest wartością (punktem) odcięcia. Obecnie do wyznaczania modeli scoringowych korzysta się z metod statystycznych oraz metod sztucznej inteligencji.

W artykule zostały przedstawione wybrane metody eksploracji danych, obecnie wykorzystywane do tworzenia modeli scoringowych. Zaprezentowano również wyniki testów efektywności wybranych metod klasyfikacji, przeprowadzonych na danych pochodzących z jednego z banków niemieckich.

2. Populacja bazowa zbioru danych (through-the-door population)

Dane uczące, wykorzystywane podczas budowy modeli scoringowych, powinny zawierać kredytobiorców o podobnych cechach do potencjalnie przyszłych klientów (tak zwana populacja bazowa *through-the-door population*) [2]. Dane wykorzystane do uczenia klasyfikatorów, przedstawionych w niniejszym artykule dotyczą klientów, którzy zaciągnęli kredyty konsumenckie w jednym z banków niemieckich. Zestaw danych składa się z 1000 historycznych obserwacji. W praktyce może okazać się, że taka liczba jest niewystarczająca do zbudowania poprawnie działającego modelu. Przyjmuje się, że aby stworzyć skuteczny system, przy jednoczesnym zachowaniu kosztów finansowych i czasowych, związanych ze zbieraniem i analizą danych na odpowiednim poziomie, liczba obserwacji powinna zawierać 3000 przypadków [2]. Oczywiście próba powinna składać się z dostatecznej liczby przypadków

„dobrych" i „złych" kredytobiorców. Optymalnym rozwiązaniem jest 50-procentowy udział każdej z grup w całej populacji bazowej.

W danych wyodrębniono binarną zmienną zależną, określającą czy kredytobiorca spłacił kredyt czy też nie. Zmienna ta pełni rolę etykiety klasy. Ponadto, dane zawierają 20 atrybutów opisujących daną obserwację, pełniących rolę zmiennych niezależnych. Zbiór uczący wraz z opisem jest dostępny na stronie Uniwersytetu w Monachium (http://www.stat.uni-muenchen.de/service/datenarchiv/kredit/kreditvar_e.html).

3. Metody budowy modeli scoringowych

Obecnie istnieje dość duża liczba metod wykorzystywanych przy modelowaniu kart scoringowych. Główną metodą, jaką stosuje się przy modelowaniu kart scorignowych (w kontekście metod eksploracji danych), jest metoda klasyfikacji nadzorowanej. Metody stosowane w credit scoringu można podzielić na dwie główne grupy:

- statystyczne,
- niestatystyczne (zaliczane do grupy metod, określane terminem *machine learning*).

Tabela 1

Podział metod scoringowych

Statystyczne	Niestatystyczne
Analiza dyskryminacyjna	Programowanie matematyczne
Regresja liniowa	(liniowe i całkowitoliczbowe)
Regresja logistyczna	Sieci neuronowe
Drzewa klasyfikacyjne	Algorytmy genetyczne
k-NN	Systemy eksperckie
SVM	

Oprócz wyżej wymienionych metod, w literaturze są dostępne również opisy innych metod, np. genetycznych klasyfikatorów rozmytych oraz neurorozmytych, wykorzystywanych do budowy modeli scoringowych [3].

4. Przebieg eksperymentu

W punkcie tym zostanie przeprowadzony eksperyment, w którym zostaną wykorzystane wybrane metody stosowane obecnie w credit scoringu. Do przeprowadzenia eksperymentu zostały wybrane następujące metody uczące:

- regresja logistyczna,
- drzewo decyzyjne (C4.5),

- k-NN,
- sieci neuronowe:
  - jednokierunkowa, perceptron wielowarstwowy (MPL)
  - o radialnych funkcjach bazowych (RBF),
- support vector machines:
  - radialna funkcja jądra,
  - liniowa funkcja jądra.

Nauczone klasyfikatory zostaną następnie poddane ocenie. Wykorzystana tutaj zostanie metoda eksperymentalna o nazwie *walidacja krzyżowa* (*ang. k-fold cross validation*). W metodzie tej zbiór przykładów U jest losowo podzielony na k równolicznych podzbiorów E(i), dla i=1,...,k. W i-tej iteracji zbiór E = U-E(i) jest stosowany jako zbiór uczący, a sam zbiór E(i), jako zbiór przykładów testowych. Trafność klasyfikowania jest wyliczana jako wartość średnia z trafności estymowanych w każdej iteracji $\eta(E_i)$ [5]:

$$\overline{\eta}_{ov} = \frac{1}{k}\sum_{i=1}^{k}\eta(E_i) \tag{1}$$

Wartość parametru k, powinna być dobrana w zależności od rozmiarów analizowanego zbioru danych. Zalecaną wartością jest wartość k=10. Taka wartość jest ustawiona w eksperymencie.

Przebieg eksperymentu ma na celu zbudowanie modeli scoringowych za pomocą wymienionych wyżej metod, a następnie ocena jakości predykcyjnych zbudowanych modeli. W rzeczywistych warunkach, wybór optymalnego modelu powinno przeprowadzać się na podstawie procedury opisanej w [4] (s. 246):

1. Wybór architektury klasyfikatora i parametrów do optymalizacji (np. dla klasyfikatora SVM z jądrem Gaussa: parametry jądra, wartość parametru regulacyjnego C).
2. Uczenie klasyfikatora na zbiorze treningowym.
3. Ocena tymczasowego klasyfikatora na zbiorze walidacyjnym.
4. Powtarzanie kroków 1-3 dla różnych architektur i wartości parametrów.
5. Wybór najlepszego modelu (odpowiadającego najmniejszemu błędowi klasyfikacji), jego uczenie końcowe na zbiorze będącym sumą zbiorów treningowego i walidacyjnego.
6. Ocena nauczonego klasyfikatora na zbiorze testowym.

Ocenę końcową wybranego modelu należy przeprowadzić na osobnym zbiorze testowym, niemającym części wspólnej ze zbiorami: treningowym i walidacyjnym. Jeżeli, dla uzyskania zbiorów treningowego i walidacyjnego, została zastosowana metoda walidacji krzyżowej, punkty 1-3 należy wykonać dla każdej z k par: uczącej i testowej.

4.1. Analiza danych populacji bazowej

Przed przystąpieniem do konstrukcji modelu scoringowego niezbędne jest zadbanie o jakość danych użytych do jego budowy. Należy wziąć pod uwagę między innymi:

- **ilość zebranych danych**, na podstawie których model będzie wyznaczał zależności. Zbyt mała ilość może spowodować, że model nie będzie spełniał należycie swojego przeznaczenia,
- **równomierna obserwacja wszystkich grup ryzyka**. Model uzyskany tylko na podstawie obserwacji, z których zdecydowana większość opisuje sytuacje prawidłowej spłaty kredytu, będzie miał tendencje do zbyt optymistycznego uznawania kredytobiorców za wiarygodnych,
- **analiza poprawności jednorodności danych**. Szczególną uwagę należy zwrócić na to, aby dane zawierały informacje tylko o kredytobiorcach należących do jednorodnej grupy (np. klienci, którzy zaciągnęli kredyty konsumenkcie),
- **braki w danych**. Zgromadzone dane nie mogą również zawierać braków. Warto rozważyć nieuwzględnienie ich w analizie, ponieważ mogą mieć negatywny wpływ na jakość modelu,
- **warunki ekonomiczne**. Dane dla modelu powinny być zbierane dla okresu o porównywalnych warunkach ekonomicznych i rynkowych (hossa, bessa, kryzys).

W zależności od wybranej metody, należy odpowiednio przygotować dane (transformacja danych), aby nadawały się one do wykorzystania przez algorytmy eksploracji danych. Na przykład, stosując metodę indukcji drzew decyzyjnych należałoby przeprowadzić dyskretyzację wartości ciągłych[1]. Stosując dyskretyzację, należy zwrócić uwagę, aby liczba przypadków w powstałych grupach była odpowiednio duża. Przykładowo, jeżeli dla 300 przykładów w wyniku dyskretyzacji danego atrybutu, powstały 4 grupy, a w jednej z nich znalazło się np. tylko kilka kredytobiorców, to jest to sytuacja nieprawidłowa. W modelach scoringowych, w takich przypadkach, stosuje się operację o nazwie *grupowanie*. W grupowaniu wykorzystuje się tabulogramy poprzeczne o dużym stopniu szczegółowości [2]. Zawiera on takie informacje, jak: nazwa dyskretyzowanego atrybutu, liczność przypadków w grupie klientów „dobrych" oraz „złych", udział procentowy klientów w poszczególnych grupach, według określonego atrybutu oraz tak zwana „szansa bycia dobrym" (ang. *odds to be good*), którą obliczamy w następujący sposób:

$$Szansa\ bycia\ dobrym = \frac{\%\ klientów\ zaliczanych\ do\ grupy\ „dobrych" \ wg\ danego\ atrybutu}{\%\ klientów\ zaliczanych\ do\ grupy\ „złych" \ wg\ danego\ atrybutu} \tag{2}$$

[1] Istnieją metody indukcji drzew decyzyjnych, które nie wymagają dyskretyzacji atrybutów ciągłych, np. algorytm SPRINT. W niniejszym eksperymencie został wykorzystany algorytm C4.5, który jest dość często stosowany w modelach scoringowych i to on został wybrany do eksperymentu.

Znając udział procentowy oraz „szansę bycia dobrym" wg poszczególnych atrybutów, dokonuje się grupowania atrybutów.

Innym, ważnym elementem jest operacja kodowania wartości atrybutów. Podczas kodowania wartościom danego atrybutu przypisuje się pewien ustalony kod.

Zbiór danych uczących zawiera 1000 obserwacji, z których 700 przypadków należy do klasy „dobrych" kredytobiorców, a 300 do „złych". Taki rozkład obserwacji może mieć istotny wpływ na jakość zbudowanego klasyfikatora (modelu scoringowego). Przy takim rozkładzie obserwacji, klasyfikator przypuszczalnie będzie się mylił oceniając złe wnioski.

4.2. Wyniki

Tabela 2

Macierz pomyłek, jaka została wykorzystana przy ocenie modeli

	Oryginalne klasy		
Przewidywane klasy	Good	Bad	
Good	TP	FP	**Pozytywna wartość predykcyjna** TP/(TP+FP)
Bad	FN	TN	**Negatywna wartość predykcyjna** TN/(TN+FN)
	Wrażliwość TP/(TP+FN)	**Specyficzność** TN/(TN+FP)	**Dokładność** TP+TN/TP+TN+FP+FN

Uzyskane modele, zbudowane za pomocą wybranych metod, przedstawionych na początku niniejszego rozdziału, zostały poddane ocenie. Istnieje wiele metod oceny dokładności dyskryminacyjnej zbudowanych modeli scoringowych: Test Kolmogorowa-Smironowa, wskaźnik Gini, krzywe ROC itp. [14]. W pracach poświęconych badaniom porównawczym metod budowy modeli scoringowych autorzy bardzo często stosują krzywą ROC. Ta metoda również i tutaj została wykorzystana. Dodatkowo w celu głębszej analizy skuteczności klasyfikowania wykorzystano również *macierz pomyłek* (ang. *confusion matrix*). Ponieważ budowane modele dotyczą problemów klasyfikacji binarnej, dlatego dodatkowo wyznaczone zostały dwie miary: *wrażliwość* (ang. *sensitivity*) oraz *specyficzność* (ang. *specificity*).

Tabela 3

Macierz pomyłek dla drzewa klasyfikacyjnego C4.5

	Oryginalne klasy		
Przewidywane klasy	Good	Bad	
Good	596	164	78,42%
Bad	104	136	56,67%
	85,14%	45,33%	73,20%

Wyniki dla wszystkich modeli, jakie zostały zbudowane w ramach eksperymentu, zostały zebrane w poniższej tabeli. W wynikach końcowych pominięto miary: *pozytywna wartość predykcyjna* oraz *negatywna wartość predykcyjna*.

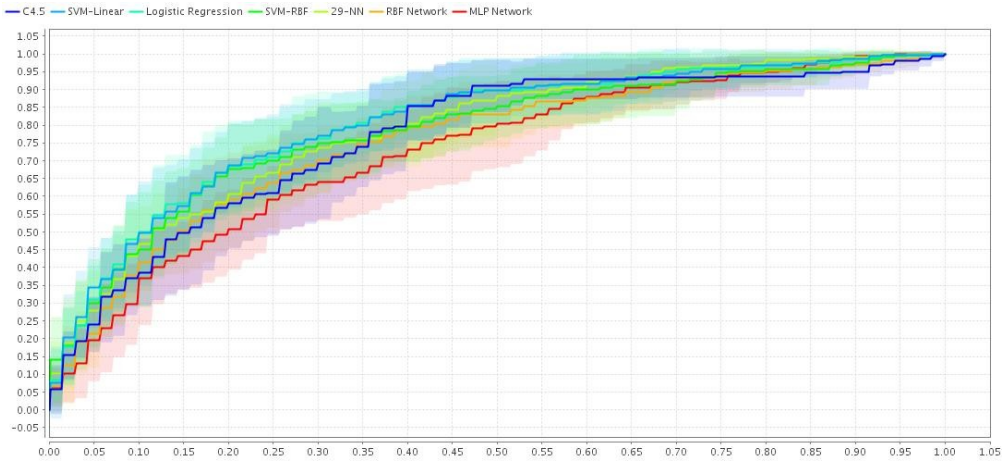
Tabela 4

Porównanie wyników zbudowanych modeli scoringowych

Metoda	Dokładność [%]	Wrażliwość [%]	Specyficzność [%]
Regresja logistyczna	76,7	88,86	48,33
C4.5	73,2	85,14	45,33
29-NN	75	88,86	42,67
MLP	71,9	82,29	47,67
Sieci – RBF	74,9	85,43	50,33
SVM – RBF	76,2	88,43	47,67
SVM – Liniowy	76,7	88,71	48,67

Analizując wyniki, można od razu potwierdzić przypuszczenia, jakie zostały zasugerowane podczas analizy danych (punkt 4.1), że prawdopodobnie model będzie się mylił oceniając złe wnioski. Jak widać w zebranych wynikach w tabeli 4, miara *specyficzność* oscyluje na poziomie 47,24%.

Najlepszą dokładność udało się osiągnąć dla modeli zbudowanych za pomocą algorytmu SVM oraz regresji logistycznej. Osiągnięta dokładność, dla modeli zbudowanych za pomocą sieci neuronowej RBF oraz algorytmu k-NN, była na zbliżonym poziomie i wyniosła 75%. Jeżeli chodzi o metodę k-NN, to najlepsze rezultaty zostały osiągnięte dla k=29. Sieci neuronowe MLP z kolei trochę gorzej radziły sobie z przykładami należącymi do klasy „dobrzy" kredytobiorcy (*wrażliwość* równa 71,9%), ale już dla przykładów należących do klasy „źli" kredytobiorcy charakteryzowały się takimi samymi wynikami, jak w algorytmie SVM-RBF (*specyficzność* równa 47,67%).



Rys. 1. Krzywe ROC dla oszacowanych modeli

Fig. 1. The ROC curves for estimated models

Krzywa ROC (ang. *Receiver Operating Characteristic*) jest narzędziem do oceny poprawności klasyfikatora, zapewnia ona łączny opis jego wrażliwości i specyficzności. Na osi odcię-

tych jest „*1-specyficzność*”, a na osi rzędnych „*wrażliwość*”. Wysoka „*wrażliwość*” (bliska wartości 1) oznacza, że system prawidłowo klasyfikuje kredytobiorców uważanych jako „dobrzy”. Mała wartość parametru „*1-specyficzność*” oznacza, że system niewiele kredytobiorców „złych” klasyfikuje jako „dobrych”. W związku z tym, pożądane są: wysoka wartość „*wrażliwość*” i niska wartość parametru „*1-specyficzność*”. Krzywa ROC bywa często wykorzystywana do oceny i porównywania między sobą modeli klasyfikacyjnych. Bardzo popularnym podejściem jest wyliczanie pola pod wykresem krzywej ROC, zwanego jako AUC (ang. *Area Under Curve*) i traktowanie go jako miarę dobroci i trafności danego modelu [10]. Wartość wskaźnika AUC przyjmuje wartości z przedziału [0,1]. Im większa wartość, tym lepszy model.

Tabela 5

Wartość wskaźnika AUC dla zbudowanych modeli

Metoda	**Wartość AUC**
Regresja logistyczna	0,806
C4.5	0,718
29-NN	0,766
MLP	0,726
Sieci – RBF	0,755
SVM – RBF	0,783
SVM – Liniowy	0,806

Na podstawie otrzymanych wyników można zaobserwować, że dla użytego zbioru danych najlepsze modele udało się uzyskać za pomocą regresji logistycznej oraz algorytmu SVM. W przypadku klasyfikacji kredytobiorców do grupy „złych” najlepiej radziły sobie sieci neuronowe RBF.

Tabela 6

Porównanie uzyskanych wartości AUC

Metoda	**A**	**B**
Regresja logistyczna	0,806	0,777
C4.5	0,718	0,747
29-NN	0,766	0,70(k=10), 0,761(k=100)
MLP	0,726	0,787
Sieci – RBF	0,755	-
SVM – RBF	0,783	0,772
SVM – Liniowy	0,806	0,766

W literaturze można znaleźć przykłady porównywania różnych metod stosowanych do budowy modeli scoringowych. Przykładami takich prac mogą być: [11, 12] lub chociażby [13]. Poprawność uzyskanych wyników w powyższym eksperymencie można zweryfikować na przykład w odniesieniu do pracy [15]. Autorzy porównali kilkanaście metod, na kilku różnych zbiorach uczących. Jednym ze zbiorów był zbiór wykorzystany w niniejszym artykule. Dla każdego uzyskanego modelu została wyznaczona wartość AUC. W odniesieniu do metod

wykorzystanych w niniejszym artykule, wyniki zostały zaprezentowane w poniższej tabeli. W celu lepszej przejrzystości, wyniki przeprowadzonego eksperymentu są oznaczone przez A, natomiast wyniki uzyskane w pracy [15] przez B.

5. Podsumowanie

Niniejszy artykuł dotyczy przedstawienia metod eksploracji danych, a dokładniej mówiąc klasyfikacji z nadzorem, obecnie stosowanych do budowy modeli scoringowych. Zastosowany zbiór danych bardzo dobrze uwidacznia jak ważną rolę stanowi jakość danych użytych do budowy modeli. Zastosowany zbiór danych miał 1000 obserwacji, z czego 700 dotyczyła „dobrych" klientów, a 300 „złych". Podczas klasyfikacji złych przypadków zbudowane modele często się myliły.

W monitoringu modeli scoringowych stosuje się trzy rodzaje raportów [1] – raporty typu: *front-end*, *back-end* oraz raporty uzupełniające. Magazyny danych należy sukcesywnie zasilać świeżymi danymi, z których korzysta model, aby algorytmy eksploracji danych prawidłowo wyznaczały zależności. Należy również co jakiś czas sprawdzać jakość generowanych wyników przez model scoringowy. Jeżeli zostaną wykryte nieprawidłowości w działaniu modelu, należy ponownie, na podstawie świeżych danych zbudować model.

BIBLIOGRAFIA

1. Matuszyk A.: Credit Scoring. CeDeWu, Warszawa 2008.
2. Janc A., Kraska M.: Credit-scoring, nowoczesna metoda oceny kredytowej. Biblioteka Menadżera, Warszawa 2001.
3. Hoffman F.: Comparing a genetic fuzzy and a neurofuzzy classifier for credit scoring, Computational Inteligent Systems do applied Researcg. Proceedings of the 5th International FLINS Conference, 2002, s. 355.
4. Stąpor K.: Automatyczna klasyfikacja obiektów, Exit, Warszawa 2005.
5. Krawiec K., Stefanowski J.: Uczenie maszynowe i sieci neuronowe. Wydawnictwo Politechniki Poznańskiej, 2004.
6. Cichosz P.: Systemy uczące się. Wydawnictwo Naukowo-Techniczne, Warszawa 2000.
7. Anderson R.: The Credit Scoring Toolkit. Oxford University Press, New York 2007.
8. Thomas L. Edelman D., Crook J.: Credit Scoring and Its Applications. Society for Industrial and Applied Mathematics, Philadelphia 2002.
9. Written I., Frank E.: Data Mining: Practical Machine Learning Tools and Techniques. Second Edition, Elsevier, Francisco 2005.

10. Bradley A. P.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 1997.
11. Zuccaro C.: Classification and Prediction in Customer Scoring. Presentation at the Global Trends Conference, Academy of Business Administration, Cancum 2008.
12. Baesens B., Setiono R., Mues C., Vanthienen J.: Using Neural Network Rule Extraction and Decision Tables for Credit Risk Evaluation. Computer Journal of Management Science, vol. 49, no. 3, 2003, s. 312÷329.
13. Gestel T., Baesens B., Garcia J., Dijcke P.: A support Vector Machine Approach to Credit Scoring. Computer Journal of Bank en Financiewezen, vol. 2, no. 4, 2006, s. 73÷82.
14. Ming-Yi Sun, Szu-Fang Wang: Validation of Credit Rating Models - A Preliminary Look at Methodology and Literature Review. JCIC Risk Research Team Column, 2007.
15. Baesens B., Gestel T., Viane S., Stepanova M., Suykens J., Vanthienen J.: Benchmarking State of the Art Classification Algorithms for Credit Scoring. Computer Journal of the Operational Research Society, vol. 54, no. 3, 2003, s. 627÷635.

Recenzenci: Dr inż. Katarzyna Harężlak
Prof. dr hab. inż. Tadeusz Morzy

Wpłynęło do Redakcji 16 stycznia 2011 r.

Abstract

Any decision to grant loan, it is fraught with risk. If the risk is higher, than the losses caused by an incorrect decision could be higher. Indeed, an important element is whether the person applying for a loan gives you the chance of its repayment. Consequently, collected some data which characterize the borrower and then the application is assessed by the scoring system and experts of the credit risk. To meet the scoring system due to its goals, can not be based on an accepted theory of rigid definition of "bad" clients. Using data mining methods we are looking for certain patterns in the data already collected under other loan applications.

Adres

Rafał BAJEK: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16, 44-100 Gliwice, Polska, rbajek@poczta.onet.pl.