ESTIMATION OF LOAN APPLICANTS DEFAULT PROBABILITY APPLYING DISCRIMINANT ANALYSIS AND SIMPLE BAYESIAN CLASSIFIER

Ricardas Mileris
Kaunas University of Technology, Lithuania, ricardas.mileris@ktu.lt

Abstract

In commercial banks risk management the credit risk measurement of each client is very important for the ability to discriminate reliable clients from not reliable. The need for models that predict defaults accurately is imperative, because bank crediting clients can take either preventive or corrective action. One of possible quantitative methods for solving credit risk estimation problems is discriminant analysis. In this paper 27 discriminant analysis models of various researchers for classification of companies were analyzed. The average classification accuracy of these models was evaluated. Often discriminant analysis is used as method to classify bank’s clients into two classes: default and not default. So the discriminant analysis model was developed to classify Lithuanian companies. The best classification accuracy rates were reached by model analyzing data about companies of 2 years. In this research according to 4 financial ratios, 8 ratings scale was created. The simple Bayesian classifier was applied for calculation of posterior probabilities to default of each rating. Created rating scale increased the correct classification rate of model from 84% to 98%.

Keywords: credit risk, discriminant analysis, posterior probability, probability of default, simple Bayesian classifier.

Introduction

Credit risk management in banks essentially focuses on determining likelihood of client’s default or credit deterioration and how costly it will turn out to be if it does occur (George, Sinha, Murali, 2007). Decisions concerning credits granting are one of the most important in commercial banks’ policy. Well-allocated credits may become one of the biggest sources of profits for banks. On the other hand, this kind of bank’s activity is connected with high risk as big amount of bad decisions may even cause bankruptcy. The key problem consists of distinguishing good (that surely repay) and bad (that likely default) credit applicants. The main investigations, in this area, are based on building credit risk evaluation models, allowing for automating or at least supporting credit granting decisions (Zakrzewska, 2007). According to Lai, Yu, Wang, Zhou (2006) many different approaches including individual models, such as linear discriminant analysis, logit analysis, probit analysis, linear programming, integer programming, k-nearest neighbour, classification tree, artificial neural networks, genetic algorithm, support vector machine and some hybrid models are widely applied to credit risk analysis tasks.

The purpose of the research – ascertain possibilities to measure banks clients default probability using discriminant analysis and simple Bayesian classifier credit risk estimation model.

The tasks of the research:
1. To describe the essence of discriminant analysis method applying it for estimation of credit risk.
2. To define posterior probabilities of banks clients default calculation process.
3. To develop the credit risk estimation model and measure bank’s clients probability of default.

Methods of the research:
1. Analysis of scientific literature.
2. Developing of credit risk estimation model and measuring bank’s clients probability of default.

Classification of companies by discriminant analysis method

The linear discriminant analysis (LDA), is one of the earliest formal modeling techniques based on the sampling paradigm (Nargundkar, Priestley, 2006). In 1936, Fisher introduced the idea of discriminating between groups in a population. In 1941, Durand realised that Fisher's discriminant analysis could be used to differentiate between good and bad loans. For many years, the decision to grant a loan had been done applying this method by credit analysts (Chwee, 2004).

The general objective of multivariate discriminant analysis within a credit assessment procedure is to distinguish solvent and insolvent borrowers as accurately as possible using a function which contains several factors.
independent creditworthiness criteria (e.g. figures from annual financial statements). Multivariate discriminant analysis is explained here on the basis of a linear discriminant function, which is the approach predominantly used in practice. In linear multivariate discriminant analysis, a weighted linear combination of indicators is created in order to enable good and bad cases to be classified with as much discriminatory power as possible on the basis of the calculated result, i.e. the discriminant score \( Z \) (Oesterreichische Nationalbank, 2004):

\[
Z = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_n x_n
\]

(1)

where \( x_i \) - the specific indicator value;

\( a_i \) - each indicator's coefficient within the scoring function.

The aim is to weight predictor scores so that a single composite variable, the discriminant score, is produced (Fielding, 2007). Linear discriminant analysis (LDA) provides the linear discriminant functions. The results of LDA are very easy to implement and interpret, but there are some limitations:

1. The LDA needs strong assumptions that the distribution of predictors within each class is a multivariate normal distribution and the classes have a common covariance matrix. If the classes cannot share a common covariance matrix, quadratic discriminant analysis should be used as a non-linear model.
2. Unordered categorical predictors cannot be used as the input variables.
3. The LDA is very sensitive to outliers compared to logistic regression (Karwowski, 2006).

Discriminant analysis models for enterprises classification of Abdou (2009), Zhou, Lai, Yu (2009), Ciampi, Gordini (2008), Marinakis and other (2008), Altman, Sabato (2007), Yun, Jianying, Lin (2007), Kim (2007), Bandyopadhyay (2006), Falavigna (2006), Kumar, Bhattacharya (2006), Lai, Yu, Wang, Zhou (2006), Nagundkar, Priestley (2006), Mileris (2006), Witkowska (2006), Yim, Mitchel (2005), Solojentsev (2003) etc. were analyzed in order to estimate average classification accuracy of these models. The average correct classification rate (CCR) of 27 models is 77.38%. This means that on the average 77.38% of clients are being classified correctly. Standard deviation, maximum and minimum values of this rate are shown in Table 1.

**Table 1. Descriptive statistics of discriminant analysis models CCR rate (%)**

<table>
<thead>
<tr>
<th>Number of models</th>
<th>Average</th>
<th>Standard dev.</th>
<th>Max</th>
<th>Min</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
<td>77.38</td>
<td>9.57</td>
<td>95</td>
<td>59.74</td>
</tr>
</tbody>
</table>

Measuring credit risk often it is not sufficient only classify bank’s clients into two groups (default or not default). Banks for their clients attribute credit ratings. One of the most important credit risk indicators is the client’s probability of default. Hamerle, Liebig, Rosch (2003) maintained that in probabilistic terms the default event is random. Any attempt to quantifying credit risk has to determine the probability of the default event (PD) within a given future time period.

**Simple Bayesian classifier and posterior probabilities**

The posterior probability of an event is the probability of an event after collecting some empirical data (Antonakis, Sfakianakis, 2009). It is obtained by integrating information from the prior probability with additional data related to the event in question (Rosner, 2006). Often analysis begins with initial or prior probability estimates for specific events of interest. Then from sources such as a sample we obtain additional information about the events. Given this new information the prior probability values can be updated by calculating revised probabilities, referred to as posterior probabilities. Bayesian theorem provides a means for making these probability calculations (Anderson, Sweeney, Freeman, 2007). Bayesian theorem is:

\[
P(C_i | X) = \frac{P(X | C_i)P(C_i)}{P(X)}
\]

(2)

where \( P(C_i | X) \) is the posterior probability of \( C_i \) conditioned on \( X \); \( P(C_i) \) is the prior probability of \( C_i \); \( P(X | C_i) \) is the probability of \( X \) conditioned on \( C_i \); \( P(X) \) is the prior probability of \( X \) (Han, Kamber, 2006).

According to Carruthers, Laurence, Stich (2008) \( P(H) \) is the probability of a hypothesis in the absence of any observed data. \( P(X|H) \) is named as likelihood and means the probability of data given the hypothesis. This includes assumptions about how likely the data are to be observed if we make some educated guesses about the sampling condition. \( P(X) \) is independent of \( H \). Combining priors and likelihood we can derive
posterior probabilities that give a quantitative measure of how likely it is that a particular hypothesis is true, given the observed data (Carruthers, Laurence, Stich, 2008). The steps in this probability revision process are shown in Figure 1 (Anderson, Sweeney, Freeman, 2007).

**Figure 1.** Probability revision using Bayes theorem (Anderson, Sweeney, Freeman, 2007)

Analysis of scientific literature (Goldstein, Wooff, 2007; Han, Kamber, 2006; Kennedy, 2003) has shown that the simple Bayesian classifier works as follows. Let $D$ be a training set of bank clients. Each client is represented by an $n$-dimensional attribute vector $X = (x_1, x_2, ..., x_n)$ of attributes $A_1, A_2, ..., A_n$. There are $m$ classes $C_1, C_2, ..., C_m$. The classifier will predict that client $X$ belongs to the class having the highest posterior probability, conditioned on $X$. Client $X$ belongs to the class $C_i$ if:

$$P(C_i|X) > P(C_j|X), \text{ for } 1 \leq j \leq m, j \neq i \quad (3)$$

The probability $P(C_i|X)$ must be maximized. By Bayesian theorem:

$$P(C_i | X) = \frac{P(X | C_i)P(C_i)}{P(X)} \quad (4)$$

$P(X)$ is constant for all classes, only $P(X|C_i)P(C_i)$ needs to be maximized. If prior probabilities of classes are not known it is commonly assumed that classes are equally likely:

$$P(C_1) = P(C_2) = ... = P(C_m) \quad (5)$$

Therefore $P(X|C_i)$ would be maximized. The class prior probabilities can be estimated by:

$$P(C_i) = C_{i,D} / D \quad (6)$$

where $C_{i,D}$ is the number of training cases (clients) of class $C_i$ (Han, Kamber, 2006).

Having data sets with many attributes probability $P(X|C_i)$ can be calculated:

$$P(X | C_i) = \prod_{k=1}^{n} P(x_k | C_i) = P(x_1 | C_i) \times P(x_2 | C_i) \times ... \times P(x_n | C_i) \quad (7)$$

The probabilities $P(x_1|C_i), P(x_2|C_i), ..., P(x_n|C_i)$ can be calculated from training set. The probabilities $P(X|C_i)$ for categorical and continuous-valued attributes are calculated differently. If $A_k$ is categorical, then $P(x_k|C_i)$ is the number of bank’s clients of class $C_i$ having the value $x_k$ for $A_k$, divided by $C_{i,D}$, the number of clients of class $C_i$ in $D$. If $A_k$ is continuous-valued then such attribute is typically assumed to have a Gaussian distribution with mean $\mu$ and standard deviation $\sigma$:

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (8)$$

Then:

$$P(x_k | C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}) \quad (9)$$

It is necessary to compute $\mu_{C_i}$ and $\sigma_{C_i}$, i.e. the mean and standard deviation of the values of attribute $A_k$ for training cases of class $C_i$. In order to estimate probability $P(x_k|C_i)$ these 2 quantities plugged into equation (8) together with $x_k$.

In order to predict the class of bank’s client $X$, probabilities $P(X|C_i)P(C_i)$ are evaluated for each class $C_i$. The classifier predicts that client $X$ belongs to class $C_i$ if:

$$P(X|C_i)P(C_i) > P(X|C_j)P(C_j), \text{ for } 1 \leq j \leq m, j \neq i \quad (10)$$

In other words the predicted class of banks client is class $C_i$ for which probability $P(X|C_i)P(C_i)$ is the maximum (Han, Kamber, 2006).
Credit risk estimation model

Using financial data about 100 Lithuanian companies (50 – defaulted, 50 – not defaulted) two discriminant functions were constructed. This model analyzes 11 financial ratios of 2 years. Correct classification rate indicated that overall 84% of companies were classified correctly by discriminant functions. Type I error rate was 30% (it is the proportion of not reliable companies classified as reliable) and type II error rate was 2% (it is the proportion of reliable companies classified as not reliable).

Further only clients were analyzed, which by discriminant analysis were classified as reliable. According to discriminant analysis results and values of 4 financial ratios, 1 of 8 ratings were attributed for each bank’s client. Credit risk estimation process is illustrated in Figure 2.

\[ DR = (D_j \cdot 100%) / R_j = (35 \cdot 100%) / 36 = 97.22\% \]

where \( D_j \) – number of companies defaulted in group of not reliable clients; \( R_j \) – number of companies in group of not reliable clients.

It means that 97.22% of companies, classified by discriminant analysis as not reliable, actually defaulted. Credit ratings were attributed according 4 financial ratios:

- Liquidity ratios:
  - Quick Ratio (QR) = (Current Assets – Inventory) / Current Liabilities.

- Profitability ratio:
  - Net Profit Margin (NPM) = Net Profit / Sales.

- Financial structure ratio:
  - Debt Ratio (DR) = Total Debt / Total Assets.

So liquidity of company determines 50%, profitability – 25%, financial structure – 25% of credit rating. In set of actually reliable clients minimum and maximum values of financial ratios were found (Table 2).

<table>
<thead>
<tr>
<th></th>
<th>CR</th>
<th>QR</th>
<th>NPM</th>
<th>DR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>0.197399</td>
<td>0.119093</td>
<td>0.00087972</td>
<td>0.008479</td>
</tr>
<tr>
<td>Max</td>
<td>26.36882</td>
<td>16.51885</td>
<td>0.48659377</td>
<td>0.935895</td>
</tr>
</tbody>
</table>

Intervals between minimum and maximum values were divided into 8 ranks (Table 3).
This question. Posterior probabilities to default in each event are the intersection of 2 events so multiplication rule can be used to compute the joint probabilities. In tree diagram company's rating (AAA – D) and company is reliable (R) or not reliable (N). Each of outcomes is the product of probabilities on the branches were calculated joint probabilities. The probabilities for each branch at step 1 are prior probabilities and at step 2 – conditional probabilities.

According to sum of ranks one of possible credit ratings were attributed for every client (Table 4).

The tree diagram (Figure 3) illustrates the classification process with 16 outcomes possible: company's rating (AAA – D) and company is reliable (R) or not reliable (N). Each of outcomes is the intersection of 2 events so multiplication rule can be used to compute the joint probabilities. In tree diagram the probabilities for each branch at step 1 are prior probabilities and at step 2 – conditional probabilities. Multiplying probabilities on the branches were calculated joint probabilities.

Figure 3. Probability tree

According to this available information bank can calculate the posterior probabilities: if company defaults what is the probability that company has particular rating. Bayesian theorem can be used to answer this question. Posterior probabilities to default in each rating were calculated according Bayesian theorem:
\[ P(C_i | X) = \frac{P(X | C_i)P(C_i)}{P(X)} \]  

(12)

where P(C_i | X) is the posterior probability of default of each rating (C_i);
P(C_i) is the prior probability of each rating (C_i);
P(X | C_i) is the probability of default (X) conditioned on rating (C_i);
P(X) is the prior probability of default (X).

\[ P(X) = \sum P(C_i \cap N) = 0.234373 \]  

(13)

Posterior probabilities of default in group of reliable clients are calculated in Table 5.

### Table 5. Posterior probabilities of default in group of reliable clients

<table>
<thead>
<tr>
<th>Rating</th>
<th>AAA</th>
<th>AA</th>
<th>A</th>
<th>BBB</th>
<th>BB</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Posterior probability of default (%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6.67</td>
<td>0</td>
<td>26.67</td>
<td>33.33</td>
<td>33.33</td>
</tr>
</tbody>
</table>

Clients classified by discriminant functions as reliable have 33.33% probability of default with ratings D and C, 26.67% with rating B. Low probability of default have clients with ratings from AAA to BB.

### Table 6. Classification accuracy rates of the developed model

<table>
<thead>
<tr>
<th>Rate</th>
<th>Discriminant analysis</th>
<th>Discriminant analysis and ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCR</td>
<td>84.0</td>
<td>98.0</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>30.0</td>
<td>2.0</td>
</tr>
<tr>
<td>( \beta )</td>
<td>2.0</td>
<td>2.0</td>
</tr>
</tbody>
</table>

If bank clients with ratings B, C and D numbers as not reliable, the correct classification rate of model increases to 98.0% (Table 6). Also the type II error rate decreases to 2.0%. It is evident that credit risk estimation model was improved. So banks using discriminant analysis and simple Bayesian classifier can measure default probability of their clients and evaluate their credit risk.

### Conclusions

1. The general objective of discriminant analysis within a credit assessment procedure is to distinguish solvent and insolvent borrowers as accurately as possible using functions which contain several independent creditworthiness criteria.
2. This research has shown that banks using discriminant analysis and simple Bayesian classifier can measure default probability of their clients.
3. The developed model analyzes 2 year financial data of companies. Overall 84% of companies were classified correctly by discriminant functions. Created rating scale increased this rate to 98%.
4. 97.22% of companies, classified by discriminant functions as not reliable, actually defaulted. Clients classified by these functions as reliable must be rejected as not reliable credit applicants if they have ratings B, C and D. That is conditioned by high posterior probability of default.

### References


