SELECTION BIAS REDUCTION IN CREDIT SCORING MODELS

Josef Ditrich

Abstract

Credit risk refers to the potential of the borrower to not be able to pay back to investors the amount of money that was loaned. For loans to individuals or small businesses, credit risk is typically assessed through a process of credit scoring. For these purposes, credit scoring models are built. It involves using different statistical techniques and historical data from the accepted applicants. However, the scorecard is designed to be used on all applicants and therefore parameter estimates of credit risk models may be biased due to the selection bias. Reject inference is a technique that tries to mitigate the consequences of this phenomenon. One of the possibilities how selection bias can be reduced is to grant loans to a part of rejected applicants and analyse their behaviour (*enlargement* method). This approach is time-consuming and costly especially. We introduced a modification of the method with the costs optimization. Our results show that involving rejected cases positively affects forecast accuracy of credit score as well as the discriminative power of models. Finally, we discuss the expected costs and benefits of the modified approach.

Key words: credit scoring models, reject inference, selection bias, enlargement method, additional information

JEL Code: C13, C24, C51

Introduction

The use of credit risk models that serve the banking and non-banking institutions to measure the riskiness of loan applicants has already become a common practice in the financial sector. Credit scoring is an important component for maintaining profitability and transparency of the entire lending process. Given the volumes with which lenders normally operate, even a slight improvement of the discriminatory and predictive abilities of these models may generate significant additional gains.

While credit scoring models are applied to the entire population of credit applicants, for their creation or for modification of the existing decision rules are usually used only the information of those applicants who have been granted a loan in the past and whose payment

discipline could be actually analysed. This discrepancy leads to *reject bias*, or more generally to *selection bias*. The consequence of the presence of this bias may be an erroneously selected acquisition strategy to expand the banking portfolio and the resulting lower-than-expected profits or even significant losses (Verstraeten and Van den Poel, 2005).

The methods aiming to eliminate or at least reduce this phenomenon are collectively referred to as *reject inference*. Most of these methods are based on the principle that they attempt to predict how the rejected credit applicants would have behaved had they been granted their loan. Based on this estimate, rejected applicants are then taken into account in creating new models.

Reject inference methods can be divided into multiple "logical" groups. The first group includes methods such as *parcelling* or *re-classification*. These are extrapolation methods which are very easy to implement into the development of the model, however, the improvements they bring are at least questionable (Kiefer and Larson, 2006). A more sophisticated method is *augmentation*. An extensive discussion on the success rate of the method can be found in Banasik and Crook (2007).

The second group of methods include techniques based on *Heckman's two-stage bias correction* (Heckmann, 1979). Here, the basic prerequisite for their applicability is full specification of two mechanisms - classification and selection. Nevertheless, although many empirical studies from the past have shown that it is a theoretically sound method which may bring certain improvement, it is unreliable and very data-sensitive. Moreover, it is based on the assumption of normality, which is usually invalid in credit scoring (Banasik et al., 2003).

Recently, there have been mentions of some new methods in the literature. One of them is the method *bound and collapse*, which stems from the bayesian theory (Chen and Astebro, 2012). This is a robust method that computes extreme probability distributions based on probability intervals. Another way is the use of approaches based on the *support vector machine* (e.g. Maldonado and Paredes, 2010).

The last group are the methods which use additional information to predict the behaviour of rejected applicants. Collectively, they are often referred to as *supplemental data methods*. Such information may be obtained from internal or external sources (Siddiqi, 2006). External sources can be, for example, data from credit bureaus, insolvency or distraint registers, or "bartering" with other financial or non-financial institutions. Analysing large credit bureau databases is the subject of the study by Barakova et al. (2013). Information from internal sources can be obtained by granting loans to a part or to all rejected credit applicants and subsequent analysing their behaviour. This approach is known as the

enlargement method (Hand, 1998). Using additional information is considered to be the most effective method. Yet, it should be noted that this method may be very expensive, which should be taken into account when implementing it.

This paper aims to present a method which could reduce the financial demands of the enlargement method while maintaining its contribution to the quality of models created. The efficiency of method modification is illustrated by comparison of the discriminative power and forecast accuracy of the models created based on a real banking database. The related financial impacts are also evaluated.

The rest of this paper is organized as follows: The next section presents the methodology and data used for impact calculation. Section 2 discusses analysis results, and the final section presents conclusions and recommendations for further development.

1 Methodology

This chapter introduces the method of selection of rejected credit applicants to the portfolio, which aims to mitigate the effects of selection bias while achieving lower financial cost than simple random sampling as used by the enlargement method. The efficiency will be tested by selected quality indicators. Also the cost of this approach and expected benefits will be discussed.

The proposed modification of the enlargement method is based on stratified random sampling. A set of rejected credit applicants is sorted in ascending order according to their probability of default (*PD*). Subsequently, the set of rejected applicants is divided into several (approximately) equally large groups (PD groups). For each group, there is set a proportion of applicants from the group to be randomly selected (who will be accepted). The proportions are selected so as to be decreasing towards the most risky applicants. The reason for the use of decreasing proportions is that the lower the credit score the higher the number of risky applicants. It can therefore be assumed that this measure will result in lower cost for obtaining new information. The worse the applicant, the more likely they get in default, which directly contributes to the loss arising from the client (see (1)).

The second possible method of sorting rejected credit applicants is directly according to their expected loss (*EL*) and their subsequent division into EL groups. This method, in particular, can be expected to result in significant reduction of additional costs for obtaining new information. Expected loss (*EL*) can be expressed as follows:

$$EL = PD \cdot LGD \cdot EAD, \qquad (1)$$

where *LGD* is loss given default, and *EAD* is exposure at default.

To calculate the expected revenue (*ER*), the following equation was created:

$$ER = k \cdot PD \cdot i_r \cdot Avg\_ON\_Balance, \qquad (2)$$

where k is the proportion of clients who pay interest on loan funds granted, i_r is annual interest rate, and *Avg_ON_Balance* is average drawing of a credit card.

Expected profit (*EP*) is obtained by subtracting expected loss from expected revenue:

$$EP = ER - EL. \qquad (3)$$

All models were created using binary logistic regression with forward likelihood ratio algorithm. The probability of entry was set to 5%.

To evaluate the discriminative power of models, i.e. the ability to distinguish between good and bad clients, were selected the following indicators: Kolmogorov-Smirnov statistics, divergency, AUROC, Gini coefficient. The higher values of each indicator, the better separation of both groups of clients model provides.

To determine the accuracy of forecast models, i.e. the accuracy of estimation of probability of default, were used the following indicators: Brier score, logarithmic score. The lower values of both indicators are calculated, the more accurate in predictions of probability of default model is.

2 Case Study

2.1 Data

Data for this research was provided by one of the largest banks in the Czech market. It contains information about clients who were randomly approached with an offer for a credit product (a credit card) as part of the bank's campaign in 2012. The only group of clients rejected were the riskiest ones with significantly negative records in bank registers, such as distraints, personal bankruptcy, etc. No other selection rule was applied. As a result, the rate of rejected credit applicants was around 5%.

As BAD were identified such clients which in the first 12 months of existence of the loan reached at least 90 days past due with a total amount of at least CZK 500. All other clients were marked GOOD. An overview of other information available about clients and loans is given in Table 1. The database contains information on 3,858 applicants, of which 316 (8.19%) the bank identified as BAD and the remaining 3,542 as GOOD. For the purpose of modelling and testing the models developed, the entire database was divided into the development and validation samples in the ratio of 2:1.

The resulting database also contains nine socio-demographic variables "normally[1]" gathered for standard loan applications, six variables calculated from the information from the credit bureau, and three variables calculated from bank records. All the explanatory variables were categorical.

Tab. 1: Overview of Information Available about the Client and Loan

Variable	Description
ID	artificial identifier
Request_Date	date of a contract
Target	explanatory variable (0=good, 1=bad)
Avg_ON_Balance	average drawing of a credit card
AR_Score	credit score calculated by the original AR model (approved-rejected model)
PD	probability of default
LGD	loss given default
EAD	exposure at default
Interest_Rate	annual interest rate

Source: own

2.2 Data Preparation

First, a model was created on the development sample, for which the above indicators of quality were calculated on the validation sample. Theoretically, this model could be considered the best possible because it uses all available observations. Later in this paper, the model is referred to as E-model (etalon model) and is used as a benchmark.

To test the efficiency of the solution proposed, a simulated situation was created where the rate of rejection of incoming population stands at 50%. To this end, the development sample was divided into two halves based on the credit score. Those with higher scores were marked as "accepted" and served for creating model M(50). The model illustrates a situation where – for the purpose of modelling – there would be available only information about accepted clients. This situation is common for creating scoring models in practice. In the next step was applied the enlargement method (models marked "rnd") and subsequently also its proposed modification (models marked "pd" and "el").

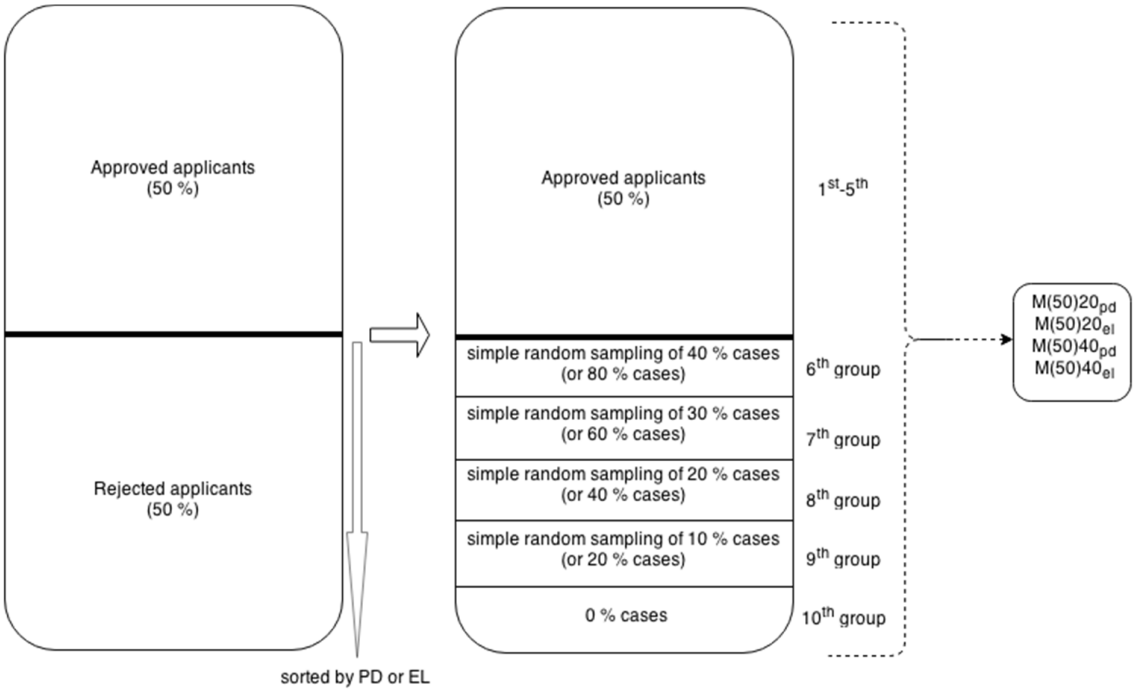
Improvements to the default model M(50) were carried out in two steps. In the first step was selected and added 20% of the rejected applicants, in the second step another 20%. The method of selecting applicants from the set of the rejected applicants depended on the specific approach tested. To make the obtained results more relevant (all tested approaches

[1] Information available about the clients was such usually appearing in questionnaires of other banks, i.e. job, education, family status, income, age, etc.

are based on simple random sampling), the selection and model designs for each method tested were replicated 100 times. The obtained results were then averaged.

Figure 1 shows both settings of the proportions of selected rejected applicants for proposed modification of enlargement method. If the bank has set aside fewer funds for improving their model, it should at the same time set a lower proportion of selected rejected applicants. Conversely, if the bank invests more funds for this purpose, it may increase the proportion of rejected applicants to be included in the final selection.

Fig. 1: Setting of Random Selections in Groups of Rejected Applicants



Source: own

The proportions were designed to diminish linearly towards the worst applicants according to the given indicator (*PD* or *EL*). The assumption is that the probability of default increases roughly linearly towards the worst applicants, along with the expected loss, and therefore with regard to economic optimization it is necessary to accept proportionately less of these applicants.

Note to (2): The value of k indicates the percentage of clients who fail to repay borrowed funds during the grace period, which is usually between 30 and 60 days for credit card products on the Czech market. This information was not available in the database provided. Therefore, the value was set at $k = 50\%$.

2.3 Results

Quality indicators of all developed models calculated on the test sample are listed in Table 2. The results clearly show that the performance of model M(50) is very weak and significantly different from the theoretically best possible model (E-model). Selecting only 20% of rejected applicants and adding them to the development sample for M(50) caused significant improvement in all three methods in both of the monitored areas of quality. Selecting and adding further 20% of rejected applicants brought slight improvement in terms of forecast accuracy, but only minimal in terms of discriminative power indicators.

Tab. 2: Indicators of Model Quality

Indicator	E-model	M(50)	M(50) 20rnd	M(50) 20pd	M(50) 20el	M(50) 40rnd	M(50) 40pd	M(50) 40el
K-S statistics	0.4227	0.2451	0.4037	0.3758	0.3927	0.3998	0.3804	0.3947
Divergency	0.976	0.209	0.568	0.479	0.483	0.668	0.579	0.618
AUROC	0.761	0.695	0.753	0.739	0.747	0.756	0.743	0.749
Gini coefficient	0.522	0.390	0.506	0.478	0.493	0.513	0.486	0.499
Brier score	0.062	0.085	0.071	0.072	0.071	0.070	0.070	0.070
Logarithmic score	0.156	0.211	0.183	0.189	0.201	0.176	0.181	0.194

Source: own

In terms of discriminative power and forecast accuracy, the best results are achieved by models which are based on samples "enriched" with the original enlargement method – M(50)20rnd, M(50)40rnd. Here all chosen indicators

This is thanks to the very accurate estimate of the proportion of bad applicants in each PD groups (see Table 3).

Tab. 3: Average Proportion of Bad Clients in PD Groups

PD group	Real bad rate	Dev. sample M(50)20rnd	Dev. sample M(50)20pd	Dev. sample M(50)20el	Dev. sample M(50)40rnd	Dev. sample M(50)40pd	Dev. sample M(50)40el
6	9.34%	9.80%	9.62%	5.71%	9.71%	9.27%	6.25%
7	12.02%	11.76%	11.39%	6.78%	11.54%	11.61%	5.88%
8	8.95%	7.84%	9.43%	6.45%	8.82%	8.82%	6.50%
9	11.28%	9.80%	11.54%	7.69%	11.54%	12.00%	7.69%
10	20.16%	21.15%		20.00%	20.39%		18.87%

Source: own

With regard to economic optimization, the proportions were designed in such a way so that none of the rejected applicants was selected from the riskiest group (10th PD group) or most loss-making group (10th EL group). Given that Spearman's rank correlation coefficient

between *PD* and *EL* is less than 1 (ρ_S=0.769), the groups of rejected applicants formed by the two different methods contain different observations. This also causes that the both development samples for "el" models include clients from the 10th PD group. The consequence is that "el" models have higher discriminative power than "pd" models. In order to optimize costs, the numbers of bad applicants in each PD group are significantly underestimated, causing weaker forecast accuracy of M(50)20el and M(50)40el.

The cost of approaches can be expressed in the form of losses arising from the acquisition of additional information. The amount of cost incurred to acquire new information depends on the method of selection of additional applicants. Table 4 clearly shows that the method of random selection of applicants from the set of rejected applicants (i.e. "rnd" models) is the costliest. However, it provides the highest accuracy in both classification of applicants and accuracy of estimates of probability of default.

Tab. 4: Average Expected Losses and Revenues

Model	EL	ER	EP	Index (EL)
M(50)20rnd	260,269	1,192,254	931,985	100.00
M(50)20pd	201,361	1,205,134	1,003,773	77.37
M(50)20el	146,623	861,578	714,955	56.33
M(50)40rnd	518,329	2,387,243	1,868,914	100.00
M(50)40pd	394,457	2,370,168	1,975,711	76.10
M(50)40el	293,376	1,725,170	1,431,794	56.60

Source: own

If the bank had a very limited budget or aimed to reduce cost to a minimum, it would do best if it selected rejected applicants based on their expected loss ("el" models). Despite the relatively strong dependence between *EAD* and *EL* (ρ = 0.55), this method allows to minimize the expected cost (loss) compared to simple random sampling (enlargement method) on the database analysed by about 44%. On the other hand, it is to be expected that the estimated proportion of bad applicants in the set of rejected applicants will be significantly underestimated, causing the model to estimate the probability of default of applicants very optimistic.

A compromise option is to select applicants based on their probability of default. "Pd" models are in both directions qualitatively slightly worse than "rnd" models. As regards the database tested, the expected decrease of cost would be approximately 23%.

Conclusion

The proposed approach is based on the principle that each rejected applicant still has a chance to get into the bank portfolio (and also into the development database for new models), but not with equal probability. More likely will be accepted those with lower probability of default or expected loss and less likely those with higher probability of default or expected loss. The measure aims to enable the bank to better optimize its cost of obtaining additional information.

The results of the empirical study show that the benefits of the modified enlargement method for the quality of models are considerable. If the aim is only to improve the discriminative power of models, it is effective to utilize a selection of rejected applicants based on their expected loss. This presents a quality model for maintaining very low additional cost. If the bank also needs to improve forecast accuracy, which is in practice a more common requirement, it is preferable to use a selection of rejected applicants based on their probability of default. Financial savings will not be high in this case, however, the model will be qualitatively almost comparable with a model built on data of rejected applicants selected randomly (i.e. using the enlargement method).

We are aware of the fact that the method of selection of rejected applicants based on expected loss is not always applicable. While the probability of default of an applicant is always known at the time the loan application, loss given default and exposure at default is often estimated only for clients who are already in the institution's portfolio. Therefore, we see room for further research in the use of other variables affecting the financial demands of the method, such as taking into account the amount of the loan requested instead of *EAD*. Also, it would be useful to determine *LGD* values in another way. For credit cards, it should be sufficient to use, for example, the average rate of drawing credit on the existing portfolio; for other types of products, to set a constant value.

Another direction for the development of both approaches could be analysing the number of formed groups which include rejected credit applicants. Alternatively, development of models could incorporate the augmentation method. Parnitzke (2005) succeeded in combining both methods (enlargement and augmentation) on simulated data with very good results.

Finally, it would be useful to focus on setting percentages for selection from groups of rejected applicants. In this paper, the proportions were set linearly. It would be favourable to try also different layouts.

References

Banasik, J., Crook, J., Thomas, L. (2003.). Sample Selection Bias in Credit Scoring Models. *Journal of the Operational Research Society*, 822-832.

Banasik, J., Crook, J. (2007). Reject Inference, Augmentation, and Sample Selection. *European Journal of Operational Research*, 1582-1594.

Barakova, I., Glennon, D., Palvia, A. (2013). Sample Selection Bias in Acquisition Credit Scoring Models: an Evaluation of the Supplemental-Data Approach. *Journal of Credit Risk*, 9, 1–43.

Chen, G. G., Astebro, T. (2012). Bound and Collapse Bayesian Reject Inference for Credit Scoring. *Journal of the Operational Research Society*, 63, 1374-1387.

Hand, D. J. (1998). Reject inference in credit operations. In Mays, E. F.: *Credit Risk Modelling: Design and Application*. Global Professional Publishing, 181-190.

Heckmann, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, 47, 153-161.

Kiefer, N. M., Larson, C. E. (2006). Specification and Informational Issues in Credit Scoring. *International Journal of Statistics and Management Systems*, 1, 152-178.

Maldonado, S., Paredes, G. (2010). A Semi-supervised Approach for Reject Inference in Credit Scoring Using SVMs. *Advances in Data Mining: Applications and Theoretical Aspects*, 6171, 558-571.

Parnitzke, T. (2005). Credit Scoring and the Sample Selection Bias. Institute of Insurance Economics, working paper.

Siddiqi, N. (2006). *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. Hoboken, N.J.: Wiley.

Verstraeten, G., Van den Poel, D. (2005). The Impact of Sample Bias on Consumer Credit Scoring Performance and Profitability. *Journal of the Operational Research Society*, 56, 981-992.

Contact

Josef Ditrich

University of Economics, Prague

W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic

xditj04@vse.cz