COLLECTION OF THESES

Beatrix Oravecz

SELECTIONAL BIAS AND ITS REDUCTION BY CREDIT SCORING MODELS

PhD dissertation

Consultant:
László Hunyadi, Ph.D.
Professor

© Beatrix Oravecz
Content

I. Research history and justification of the topic .................................................. 4
II. Methods applied ................................................................................................ 7
III. The results of the dissertation ........................................................................ 9
IV. Main references ................................................................................................ 11
V. List of own publication, in connection with the topic ...................................... 18
I. Research history and justification of the topic

In the last 15-20 years there have been revolutionary changes on the market of financial services. Banks started to use automatic decision-making methods and decision support models to be able to speed up credit approval decisions.

An information asymmetry exists between creditor and applicant. One of the greatest risks for banks is the credit risk, which expresses the risk that the credit applicant does not pay, or pays only partly the borrowed capital and/or its interests back, and thus the bank suffers loss. The basic interest of the banks is to gain more information of better quality possible about the customers, and to get more information from them about the liquidity and willingness to pay with the help of different data mining systems. Credit scoring, used for rating serves this aim.

Credit scoring played a very important role in the explosive growth of the stock of consumer credits. Without an accurate and automatic risk analyzing system banks could not have increased their retail placing in such a big measure.

Despite the wide application of credit scoring methods, the methodology still has aspects, which are not paid enough attention to neither in the special literature, nor in practice. The question of representativity of the model building sample is a field like this. Scoring models are usually built on a non-representative sample, as in this case we have a total data base typically only at those customers, who have come through a credit-review process and have been accepted. The credit scoring model, used to accept/ reject the applications by and by looses its actuality, accuracy so needs to be re-built. If the model is not refreshed, it does not follow the forthcoming changes in the population and the effect of explanatory variables, and the original model loses its predictive power. On the other hand, however, if only the data of customers accepted is used to refresh the model, the validity of the new model will be questionable, as the distribution of accepted and rejected will probably differ as a result of the systematic judgment process, so the accepted do not represent all applicants, indicating the whole population.

This phenomenon is called reject bias, or more generally selectional bias.
Model building, using the features of rejected (reject inference) may serve an answer for the dilemma. This is actually the estimation of how the rejected applicant would have behaved, if he had been granted the credit.

An often quoted example is the old offender. Old offender applicants are almost always rejected. If all were rejected, then without reject inference this criterion would not appear in the final model. The fact that the majority is rejected often means, that the minority, who is accepted, disposes really special features, and usually does not represent old offenders at all. So, if a model is constructed only on the performance of the accepted, the final model will be too optimistic.

In the dissertation we deal with methods, suitable for the reduction of selectional bias in case of credit scoring models. The analysis of the phenomenon is almost completely missing from the Hungarian special literature, and could be met only on the level of mentioning.

Beyond its curiosity, the choice of topic is reasonable because of its practical importance. As if the performance of the model can be improved a little bit, it may result in a huge increase in profit or a decrease of risk for the banks, as it is about placing of great volumes. The more accurate appraisal of risk is at the same time advantageous for the customers too, as it makes the reduction of additional price of risk for good debtors possible, or those, who have been rejected so far, can get a credit of an adequate additional price of risk.

Summarizing former researches in connection to the topic, we can state, that the adoption of rejected during the model building can be a sensible and useful solution only, if certain conditions come true for the accepted and rejected population. These solutions may work in practice as assumptions are usually reasonable, or at least show to a good direction. For instance it is a rational assumption that the ratio of bad is higher within the rejected, than within the accepted (even with the same score), even if it cannot be correctly defined numerically how much it is greater.
The benefit of the application of the real and imputed data of the rejected depends on the rejection rate, the distribution within the population and sample and the fulfillment of applied statistical conditions. There are some portfolios, where the ratio of rejected is really low (for example the market of mortgages). In these cases dealing with the rejected may be unnecessary, as their ratio within the population is negligible, so bias caused by them does not need any correction. On the other hand in case of portfolios of greater risk, for example in case of crediting small- and beginner companies, the rejection rate may be really high, so selectional bias can not be neglected.

The best solution, applicable may be occasionally different (by customer groups, products). There is no finished theoretical background in reference, whether the dropout of which conditions cause significant bias in parameter estimations. Such a general principle would be difficult to lay down, as bias is greatly dependent on the database.

According to some statisticians, the problem of conclusion from a non-random sample can be solved with the right imputation of the missing collapse data of the rejected. (Joanes 1993/4, Donald 1995, Copas and Li 1997, Greene 1998). Generatives of scorecard already apply reject inference techniques abroad, in which they are supported by statistical software packages (e.g. SAS). However, they usually work as black boxes, because the underlying principles and assumptions are not clear for the users.

If certain assumptions are acceptable, and the rejected are applied with some kind of imputation, we face a question: how can our model be validated and how can the improvement be measured? Only a few relevant studies were made in this topic, as the majority of data bases, used for testing is not complete, or was simulated (Donald 1995, Feelders 1999, Manning et al. 1987). Hand and Henley (1993/4) revealed that solutions, used in business life are problematic, as they are usually based on really doubtable assumptions.
II. Methods applied

I examined the applicability of some methods in my dissertation, and then chose the one, which seemed to be the best applicable in case of the concrete research data base.

Selectional bias, appearing in case of credit qualifying models is a problem, deriving from missing data, as in case of customers, previously rejected, the value of the variable, describing credit risk (re-payment) is missing (is not observable), so in chapter I. the types and possible methods of handling of missing data is taken one after another.

In the next chapter (II.), the tasks of credit scoring, the most often used methods and indices, suitable to value them are shortly looked over. In practice the application of logistic regression is most widespread in case of credit scoring models; therefore I also used logit models during the empirical research to estimate the non-payment possibility of the customers.

In chapter III. methods, presented in the special literature and serving the reduction of selectional bias appearing in case of scoring model are reviewed. All methods use somehow the information available about the rejected.

The effective re-payment information of the rejected is unknown, that’s why – as information cannot arise from nothing – if we want to use them for the model building, we need to use assumptions, or additional information needs to be gathered about their re-payment behavior.

In this chapter the theoretical background of such techniques (reject inference) will be introduced, highlighting the assumptions applied or the method of gathering and using additional information and sum up the practical experiences so far.

The fulfillment of the conditions applied cannot be generally tested, so -after studying the special literature- I came to the conclusion that the only robust and
effective way to eliminate bias, is to credit a part of rejected and their behavior and possible collapse is observed in this way.

It is undoubted, that the model could be corrected with the use of *additional information*, as this time we lean on more information during model building. This way, however, can not always be realized, because of the money and time need of the solution. The application of *gate opened for a mouth* with a kind of cost optimal sample distribution is a possible way to decrease the costs of the procedure. This means, that all customers, who are otherwise to be rejected have chance to get into the sample, but not with the same possibility. Those, whose expected loss is higher, can get a loan with a lower possibility and with a higher possibility those, where this expected loss is smaller. So we get a stratified sample with a certain sort of cost-optimal sample distribution. Finally a sample, representing the whole population is achieved by *reweighting* without taking charge of huge costs by allowing everybody in.

In terms of *the empirical research* I examined on a real bank data base (retail credit card data) the improvement, costs and benefits attainable with the method of gate opened for a mouth, on the scoring model, built with the help of logistic regression.
III. The results of the dissertation

- Making the special literature known I have introduced methods categorized adequately for missing data mechanisms, suitable for the reduction of selectional bias appearing by credit scoring models.

In terms of the **empirical research** I examined the improvement, costs and expected benefit of the method of gate opened for a mouth on a real bank database (on retail credit card data). As a result of the model calculations utilizable recommendations are formed for practical experts.

- During the empirical research we experienced, that *models of lower performance can be built in case of higher rejection rate (strong and not completely random selection)*, than in case of a lower-ratio rejection. One of the reasons for this is that in this case only a few bad customers are included in the portfolio, making the recognition of the characteristics of bad for the models more difficult. The other reason is that certain values of otherwise significant variables do not get into the sample as a result of selection; therefore the explanatory variable will not be significant.

- In such cases one method of collecting additional information may help, if new observations are gained from internal source with the application of gate opened for a mouth. We have seen that *the performance of the model was improved by the method of gate opened for a mouth, and as a result the profit, attainable on the product increased.*

- We found that if the aim is to maximize profit, it is *better to use a cutoff value determined theoretically*, opposite to the method of empirical definition, which is in practice widespread.
According to our results the degree of model improvement and increase of profit was the highest in the first step. So other customers, close to those customers, who are otherwise to be accepted and only a little bit worse than them worth being allowed in with the help of gate opened for a mouth. This first step, extensive improvement of the model and increase in the profit is probably only the characteristic of the data base, but other general considerations suggest this strategy too. Our estimations are much better close to the acceptance range. The ratio of bad can probably be well estimated here, therefore the costs of the extra sample can be more easily estimated, and are lower, than taking the sample from a further range of the sample.

Finally we can say, that techniques, theoretical- and practical considerations reviewed in the dissertation can be applied not only in the fields of credit scoring, but in case of many data mining problems, including similar sample selectional mechanism.
IV. Main references


T. ASTEBRO ÉS I. BERNHARDT [2001], “Bank Loans as Predictors of Small Start-up Business Survival”, manuscript


BÓDY SÁNDOR, SULYOK PAP MÁRTA [1997], “Cégminősítés”, *Nemzetközi Bankárképző Központ*, Budapest


11


D.J. FORGARTY [2005], “Multiple Imputation as a Missing Data Approach to Reject Inference on Consumer Credit Scoring”, *Manuscript*, DavidF1967@email.uopx.edu


HAJDU OTTÓ [2003]: „Többváltozós statisztikai számítások”, KSH, Budapest


D.J. Hand [1997], „Construction and Assessment of Classification Rules,” *Chichster: Wiley*.


Hunyadi László [2001], “A mintavétel alapjai”, *Egyetemi Jegyzet SZÁMALK*, Budapest

Hunyadi László és Vita László [2002], “Statisztika közgazdászoknak”, *Közponi Statisztikai Hivatal*, Budapest


N.M. KIEFER and C.E. LARSON [2003], “Specification and Informational Issues in Credit Scoring”, manuscript


E. MAYS [2004], “Credit Scoring for Risk Managers” South Western Thomson Learning


RUDAS T. [1998],”Hogyan olvassunk közvélemény-kutatásokat?”, Új Mandátum Könyvkiadó, Budapest


V. List of own publication, in connection with the topic

Articles:


Article in a book:


Others:
