VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA PODNIKATELSKÁ
ÚSTAV EKONOMIKY

FACULTY OF BUSINESS AND MANAGEMENT
INSTITUTE OF ECONOMICS

PROPOSAL OF AUTOMATIC RISK EVALUATION FOR BANKING CLIENT LOANS

NÁVRH AUTOMATICKÉHO HODNOCENÍ RIZIKA ÚVĚRU BANKOVNÍCH KLIENTŮ

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — Bc. JIŘÍ KOBELKA
AUTHOR

VEDOUCÍ PRÁCE — prof. Ing. PETR DOSTÁL, CSc.
SUPERVISOR

BRNO 2011

Brno University of Technology
Faculty of Business and Management

Academic year: 2010/2011
Institute of Economics

MASTER'S THESIS ASSIGNMENT

Kobelka Jiří, Bc.

European Business and Finance (6208T150)

Pursuant to Act. No. 111/1998 Coll., on Higher Education Institutions, and in accordance with the Rules for Studies and Examinations of the Brno University of Technology an Dean's Directive on Realization of Bachelor and Master Degree Programs, the director of the Institute of is submitting you a master's thesis of the following title:

Proposal of Automatic Risk Evaluation for Banking Client Loans

In the Czech language:

Návrh automatického hodnocení rizika úvěru bankovních klientů

Instruction:

Introduction
Executive summary
Theoretical basis of the work
Problem analysis and current situation
Proposals and contribution of suggested solutions
Conclusions
References
Apendices

Seznam odborné literatury:

DOSTÁL, P. Pokročilé metody analýz a modelování v podnikatelství a veřejné správě. 1. vyd. Brno: CERM, s.r.o., 2008. 340s. ISBN 978-80-7204-605-8.
DOSTÁL, P. Advanced Economic Analyses. 1. vyd. Brno: CERM, s.r.o., 2008, 80s. ISBN 978-80-214-3564-3.
ALIEV, A., ALIEV, R. Soft Computing and Its Applications. 1. vyd. World Scientific Pub. Ltd, UK 2002, 444s., ISBN 981-02-4700-1.
KLIR, G.J., YUAN, B. Fuzzy Sets and Fuzzy Logic, Theory and Applications. 1. vyd. Prentice Hall, New Jersy, USA, 1995, 279s., ISBN 0-13-101171-5.
THE MATHWORKS. MATLAB – Fuzzy Logic Toolbox – User's Guide. The MathWorks, Inc., 2008.

The supervisor of master's thesis: prof. Ing. Petr Dostál, CSc.

Termín odevzdání master's thesis is given by the Schedule of the Academic year 2010/2011.

L.S.

doc. Ing. Tomáš Meluzín, Ph.D. Director of the Institute	doc. RNDr. Anna Putnová, Ph.D., MBA Dean of the Faculty

Brno, 31.08.2011

Abstract

This Master’s thesis is focused on application of fuzzy logic on the process of automatic default client detection from the bank credit risk management point of view. Based on contemporary Credit Risk Management information system analysis author suggests changing approach in a loan client evaluation.

Abstract

Diplomová práce se zabývá aplikací fuzzy logiky na proces automatické detekce úpadkového klienta z pohledu řízení úvěrového rizika banky. Na základě analýzy stávajícího informačního systému Credit Risk Monitoring autor navrhuje změnu přístupu v hodnocení úvěrového klienta.

Keywords

Fuzzy logic, Credit Risk Management, Artificial Intelligence, Default client detection

Klíčová slova

Fuzzy logika, Řízení úvěrového rizika, Umělá inteligence, Detekce úpadkového klienta

Bibliographic Citation

KOBELKA, J. *Proposal of Automatic Risk Evaluation for Banking Client Loans*. Brno: Brno University of Technology, Faculty of Business and Management, 2011. 74 p. Supervisor Prof. Ing. Petr Dostál, CSc.

Declaration of originality

I hereby declare that this Master's Thesis is an original and has been written under the supervision of Professor Ing. Petr Dostál, CSc. All sources have been duly acknowledged in compliance with the relevant copyright legislation (Act. No. 121/2000 Coll., on copyright).

Brno, 31 August 2011

……..

Bc. Jiří Kobelka

Acknowledgements

The author would like to thank Professor Ing. Petr Dostál, CSc. for his valuable advice and support related to this Master's Thesis and the underlying research and to Ing. Pavel Kozák for his collaboration and for providing the data needed for the research.

Contents

1. Introduction

Volksbank CZ, a.s. (hereinafter Volksbank or the Bank) is a universal commercial bank. Bank activities include processing of loan products. For credit lines to be provided client risk needs to be assessed to determine whether the client in question will be able to meet his obligations to the bank and repay the loan plus the interest. The discipline which examines this issue is called credit risk management.

Every loan case is carefully examined by the bank to avoid exposure to the risk of loss, and Volksbank – like other banks – uses an automated system to carry out risk assessment of loan clients. The current loan clients are subject to monitoring with regard to their meeting the account turnovers required by their contracts or with regard to their inclusion in any of the black lists which are part of various bank registers.

The Bank currently operates a monitoring information system whose underlying principle is static. Fixed assessment ranges are set for individual information sources in line with the methodology of the Bank's Credit Risk Management Department. The resulting values are added together, and the sum represents a rough overview of the risk posed by a client in absolute value. In principle, the higher the number, the higher the potential risk.

However, this leads to a situation in which the degree to which risk assessment can be automated is low, as there is a disproportion between the client's absolute value and obligation. Larger clients will accumulate a higher number of risk points much more quickly although they need not be more risky.

The aim of this thesis is to analyze the current information system with respect to the automatic risk assessment of Volksbank´s loan clients and to put forward a new automatic solution to increase the relevance of risk assessment.

The thesis draws on actual data kindly provided by Volksbank. The data have been subjected to an irreversible modification consisting of the removal and change of client specific information and of text values. The numerical values have been preserved to avoid any

adverse effect on the conclusions of this thesis, which should, as a result, be able to provide a valuable solution to the Bank.

2. Executive summary

The aim of this thesis is to analyze Volksbank's capability for assessing the risk in existing loan clients by means of the Credit Risk Monitoring System and to put forward a proposal for improvement.

The thesis relies on qualitative research sources, especially on documents and books and non-standardized observations. Two interviews have also been conducted, one with Ing. Pavel Kozák from Volksbank's Development Department and the other with Mgr. Martin Vojtek, Ph.D. from the Czech National Bank's department of Banking Supervision.

The analysis of the methodology used by the Credit Risk Monitoring System has shown that the system has a low success rate in detecting high risk clients amounting from 17.7 to 19.12 percent. The low success rate is principally caused by the weak comparability of the individual data, which makes a relevant definition of the risk client threshold impossible. As a result, the system can only be automated to a limited extent and requires human intervention and correction.

The improvement on the existing risk assessment methodology put forward by the thesis draws on fuzzy logic. The variables for the transformation matrix have been selected based on an analysis of the discriminatory power of the individual variables established by means of information value calculation. The thesis divides clients in two groups: retail and corporate.

Calculation only concerns the retail group, as sufficient data for the corporate segment has not been available. The corporate segment has only been subjected to a theoretical analysis, which recommends the use of the neural network method for analyzing risk in existing corporate clients.

The fuzzy model drawn up for the retail group substantially improves on the accuracy of the existing Credit Risk Monitoring Methodology to the extent of up to 72.3%. This improvement in accuracy is, however, impaired by an increased error rate.

The increased error rate is probably caused by the lack of a highly discriminatory variable in the analyzed data, which would reduce the error rate while maintaining or improving the level of accuracy with respect to the identification of existing risk clients.

The author therefore recommends that the Bank should conduct a broader analysis of the information values of other indicators extractable from client data. A wider application of neural networks is also an option that should be considered.

3. Theoretical basis of the work

The thesis use qualitative analysis to examine the Credit Risk Monitoring System at Volksbank. Its aim is to conduct a research to analyze the existing situation and attempt to put forward a better solution. The goal of the thesis is thus to answer the following main research question.

The main research question is:

"Does the existing solution for assessing risk in loan clients of the Bank lend itself to automation and in what ways can the existing solution be improved?"

The thesis follows the qualitative research strategy. Hendl (1) notes that when employing this strategy, the researcher relies on a longer intensive contact with the situation on the ground or of the individual or group in question and attempts to obtain an integrated picture of the subject of research and the logic of its context. Qualitative research is characterized by the researcher gaining a picture of the situation during research and an inductive analysis of the data followed by their interpretation.

Qualitative research employs the following methods: document study and non-standardized observation. "Documents may be the only underlying data source for the study or they may provide support for data obtained through observation or interviews" (1 p. 204) Hendl differentiates between official documents, archive data, as well as mass media and virtual data.

With respect to the event examined, Hendl (1) distinguishes between contemporary documents (originating at the time of the examined event), retrospective documents (originating after the examined event), primary documents (drawn up by direct witnesses to the event) and secondary (drawn up based on primary documents). Contemporary and primary documents are the most suitable documents for the analysis attempted here; retrospective and secondary documents would be less reliable, as they themselves represent an interpretation of contemporary and primary documents.

Another important research method is the interview, which allows the researcher peep into the world of the interviewee. An interview is based on interaction between the interviewer and the interviewee. The direction of the interview is roughly sketched ahead of the actual interview by the interviewer, who may, however, also rely on spontaneous questions arising from the natural flow of conversation between him and the interviewee (2).

The Credit Risk Information System can be seen as consisting of two theoretical parts.

The first part encompasses the risk management methodology at the given bank, which is aimed at defining the individual risk, such as the Implicit Option Risk or Interest Rate Gaps. The theoretical framework used for risk management in this thesis is the 2010 book by Professor Joël Bessis entitled *Risk Management in Banking*. Professor Bessis gives a comprehensive survey of risk management across banks, and only selected parts of his framework will be relevant for the purposes of our analysis of Volksbank.

The second part consists in interpreting the individual risk areas and their contextualization. This can be done by using artificial intelligence, specifically by applying fuzzy logic and neural networks. In this respect, the thesis draws on a 2008 book by Petr Dostál entitled *Pokročilé metody analýz a modelování v podnikatelství a veřejné správě* ("Advanced analytical methods and modelling in business and public administration").

4. Problem analysis and current situation

This chapter provides basic information on Volksbank followed by an analysis of the current way the automated risk assessment of loan clients is operated. The chapter concludes with an analysis of the problematic parts of the information system model and makes recommendations for modifying the information system based on this analysis.

4.1. Basic information about Volksbank

Volksbank entered the Czech market in 1993 (3 p. 05). The owner of Volksbank is Volksbank International in Vienna (hereinafter VBI) (3 p. 05). In 2010, Volksbank reported an annual average of 622 employees including employees on maternity leave (3 p. 04). The overall volume of loans reached CZK 39.1 billion (3 p. 04). Credit risk management is the responsibility of the Bank's Credit Risk Management Department (3 p. 18).

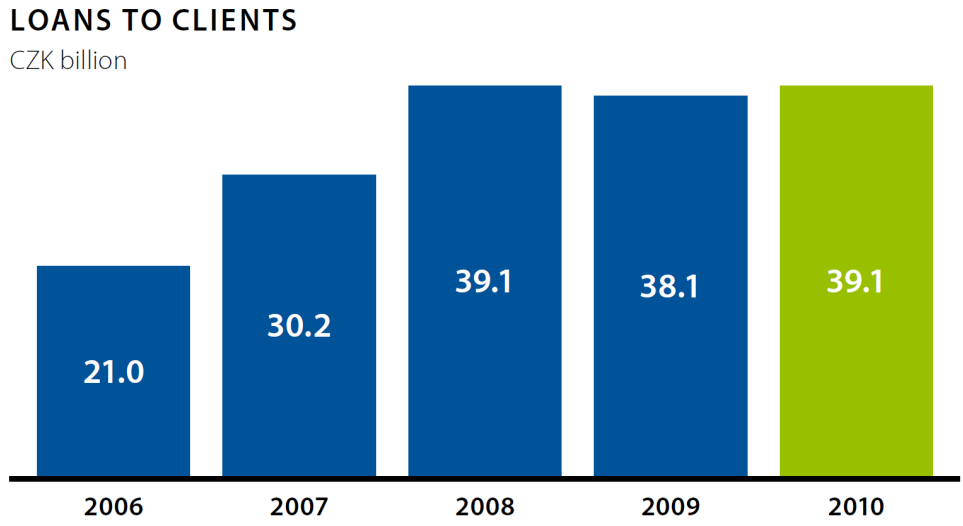


Figure 1 - Volumes of loans to clients of Volksbank (3 p. 4)

4.2. Analysis of the current situation

According to Dostál a Sojka (4 p. 6), "credit risk means the failure of debtor against the creditor; it means of not payment of debt, the creditor receives the loss". For the purposes of this thesis the Bank is deemed the creditor, while the loan client is deemed the debtor. It is the

interest of each creditor to assess his credit risk and to react to change in an adequate manner, for instance by writing down the loan or by increasing the interest rate to cover the risk (4 pp. 7-8).

Volksbank operates an in-house information system called Credit Risk Monitoring to monitor current loan clients. Individual client data from various sources form the basis for the assessment of client risk. Credit Risk Monitoring serves as an early warning system.

4.2.1. Definition of the loan client

The concept of a loan client is not as trivial as it may appear. Even within one bank one may encounter several interpretations of this term. Some may understand a loan client to denote a client with a mortgage or a consumer loan. Other users of the term may understand it to mean a client with an available credit line, such as a credit card. Yet other users may think of a client with an aggregate debit balance of below 0 at a given moment.

For the purposes of Credit Risk Monitoring, the Bank applies the last of the above methodologies. Thus 'loan client' denotes a client whose sum of debit balances on all both on- and off-balance accounts is below 0 at a given moment. Based on this criterion, the information system keeps track of all clients regardless of whether they are companies or individuals.

Although this definition may seem strange – it is not, after all, based on credit balance accounts –, there is a good reason for it. Normally, the bank may not perform netting against other client's accounts, although this may be changed contractually. The Credit Risk Monitoring information system, however, is based solely on real loan clients, namely those who have used at least a part of their credit line.

4.2.2. Default and non-default clients

Bessis describes the definition of default used by the Basel 2 banking rules as follows: "Basel 2 defines a default event as non-payment of debt obligations for 90 days"(5 p. 235). According to the Basel 2 rules, a default analysis should be carried out on an annual basis (5 p. 235). This thesis is based on the Basel 2 interpretation of default.

Accordingly, clients who have at least once been in default of the repayment of their loan for a period of more than 90 days will be considered default clients for the purposes of this thesis. Clients without such a recorded failure to repay their loan will, by contrast, be considered non-default clients.

4.2.3. Loan products

Let us now define the individual loan products, which will be referred to in the text below. Naturally, Volksbank's products needs to be taken into account. The loan products portfolio of this bank is pretty usual. It consists of mortgages, current account overdrafts, consumer loans, bank guarantees, tranche loans and investment loans. These are the products that this thesis focuses on, as these products are subject to processing by the Credit Risk Monitoring Information System.

Within the Czech context**, a mortgage loan** is a product available to individuals and suitable for financing the purchase of a real estate or – in the case of an 'American' mortgage – its loan without purpose (6 pp. 3-6). A mortgage loan will always be secured by a real estate; if the mortgage is taken out for a real estate that is about to be or being built, the real estate at its current (registered) construction stage will serve as a guarantee. Furthermore, the client and his joint debtors will have compulsory death insurance, and a bill of exchange of the corresponding value will be deposited at the bank (7). It might seem that this form of assurance is sufficient and that the bank cannot suffer any serious loss if the client breaches his contractual obligations to such an extent that the bank is forced to a write-down. However, as the global financial crisis has shown, the fall of the real estate market may cause the market value of the pledge to plunge, resulting in the bank's failure to satisfy its claims (5 pp. 3-18).

A **current account overdraft** is essentially a pre-approved credit line which the client may use as needed for a purpose of his choice and repay at any time in the future. However, interest accrues over the whole period and tends to be less favourable than that of a specific purpose loan such as a mortgage. A current account overdraft is available to the entire client portfolio, and may be acquired by private individuals, self-employed persons and companies (7; 8; 9).

A **bank guarantee** is a less commonly used bank product. Self-employed persons and companies might want to provide it to their business partners to give them the certainty that, even if they become insolvent, their business partner will still be able to satisfy its claims on them up to the amount of the bank guarantee. Thus, by issuing a bank guarantee the bank becomes a guarantor (8; 9).

A **tranche loan** is used by larger enterprises for inventory and production financing. By means of such a loan the bank enables its clients to repeatedly draw on, in individual tranches, the funds up to the amount of the approved credit line (8).

An **investment loan** is suitable for self-employed persons and companies in need of funds for the purchase of real estate, machinery and equipment or other fixed assets. The duration of its repayment should not exceed the depreciation period of the assets purchased (9; 8).

Credit cards operates on a principle which is similar to that of the current account overdraft, but differ in that the loan usually needs to be repaid within 30 to 60 days to avoid being charged a high interest.

Volksbank does provide other loan products as well, but the above mentioned products are the most important ones.

4.2.4. Functioning of the current Information System

The Credit Risk Monitoring Information System collects large amounts of data on loan clients from various sources on a daily basis. Each data source provides different information; generally, the data in question are either acquired from in-house bank systems or externally.

The following in-house data, among others, are further processed in the Credit Risk Monitoring Information System:

- Unapproved debits,
- Failure to meet the minimum obligatory credit turnovers,
- Failure to submit documents for the year-on-year status assessment of the company as of the contractual date.

External data includes, among others, data from the following sources:

- Interbank registers,
- Rating agencies,
- Government institutions (such as the Czech Ministry of Finance or the Czech Ministry of Justice).

Data from these sources are then processed and "events" are extracted from them. **Positive or neutral events** are kept track of but no risk points are awarded for them. Events considered improper by the bank are called **"negative events"** (see Appendix 1). Risk points are awarded for negative events in line with preset rules.

Several rule types exists. Generally, risk point award rules can be divided into fixed or scope-based on the one hand and into one-off or recurrent on the other. For instance, a negative entry in the Commercial Register, such as a distraint, would result in the client being awarded risk points based on a one-off fixed rule. By contrast, failure to fulfil annual, quarterly or monthly contractual turnovers will result in the award of risk points based on a recurrent rule adjusted for scope. This is because the Credit Risk Monitoring Information System uses what is called a rule based credit scoring methodology (10).

Despite the use of rule based credit scoring methodology, the aim of the Credit Risk Monitoring System is clearly not an automatic risk assessment. As Kozák notes: "Employees of the Credit Risk Management Department issue opinions on the individual events. The number of points and point ranges applicable to negative events are defined by the specialists of this department" (see Appendix 1).

The role of the system rather consists in enabling the staff of the Credit Risk Management Department to take all the relevant data into account and comment on the individual events. The information system should therefore not be perceived as a stand-alone tool for automated risk assessment, but rather as a utility which helps the Credit Risk Management Department carry out preventive detection of risk loans.

4.3. Test of the current Credit Risk Monitoring Information System

A test of a random sample of 1200 loan clients over a reference period of 12 months was carried out to verify the accuracy of the methodology currently used by the Credit Risk

Monitoring Information System in terms of default client detection. Out of these clients, 68 had been identified as actual default clients and 1132 as actual non-default clients based on the Basel 2 methodology.

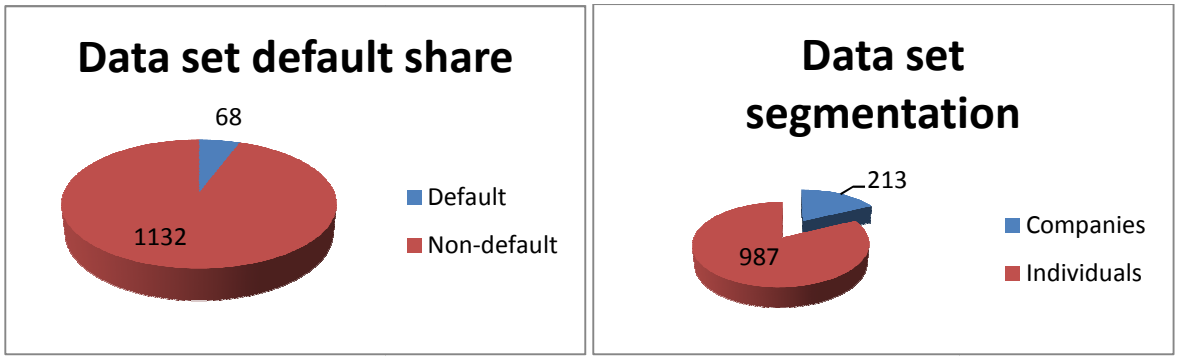


Figure 2 - Data set default share and segmentation

Since the Credit Risk Monitoring System uses an absolute point value (and not a percentage) to evaluate risk, a threshold for an automated assessment of default had to be set. The threshold was set at 50% of the maximum possible number of points over the reference period of 12 months. The description of the method for establishing the success-rate of the test is given in the following table.

Default client	Credit Risk Monitoring default	Result
TRUE	TRUE	TRUE
TRUE	FALSE	FALSE, TYPE I ERROR
FALSE	FALSE	TRUE
FALSE	TRUE	FALSE, TYPE II ERROR

Table 1 -Interpretation of the methodology for establishing the Credit Risk Monitoring Test success-rate

TYPE I ERROR indicates that an actual default client has gone unnoticed. TYPE II ERROR indicates that an actual non-default client has been incorrectly identified as a default client (5 p. 543).

The current information system did not perform well in the test. 58 actual default clients were identified as non-default clients (TYPE I ERROR), and 15 actual non-default clients were identified as default clients (TYPE II ERROR). Only in 10 cases did the Credit Risk

Monitoring Information System manage to correctly identify actual risk (default) clients. It also managed to correctly identify 1117 out of the actual 1132 well-performing clients.

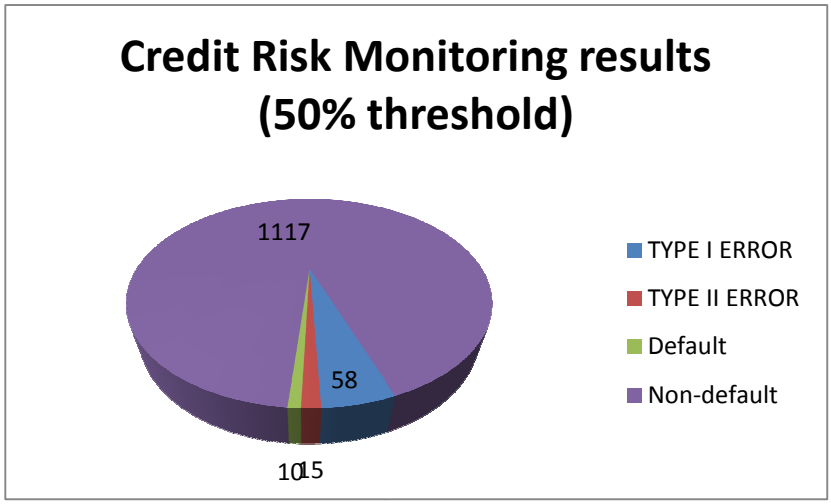


Figure 3 - Credit Risk Monitoring – test results (50% threshold)

The test result may be interpreted as showing the probability of correct automatic identification of a potential default client to be 14.7% and the probability of incorrect identification of an actual default client as non-default to be 1.33%.

To control for any errors resulting from the choice of the threshold based on which a client is assessed as default or non-default, the calculation for the same data sample over the same reference period of 12 months was performed again, this time with a threshold for the maximum number of risk points reduced from 50% to 25%.

However, there was no significant improvement. 13 clients were correctly identified as default, while 55 (actual default) clients were incorrectly identified as non-default (TYPE I ERROR). 1104 clients were correctly identified as non-default, while 28 clients were incorrectly identified as default (TYPE II ERROR).

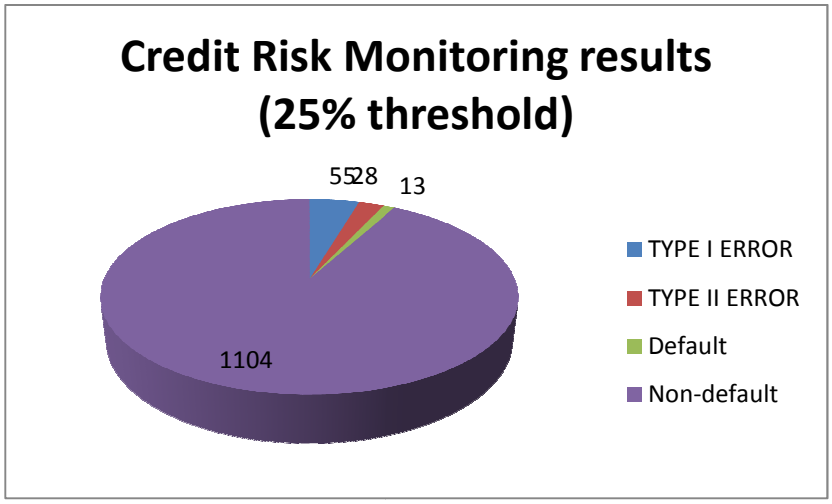


Figure 4 - Credit Risk Monitoring results (25% threshold)

The decrease in the sensitivity threshold obviously increased the assessment accuracy for a default client from 14.7% to 19.12%, which amounts to an improvement of 30.07% on the original result. At the same time, however, the error rate increased from 1.33 % to 2.47%, which means that the number of false alarms increased by as much as 85.71%.

4.4. Summary of the current situation

The Credit Risk Monitoring Information System clearly does not reflect the real status of default loans very accurately, and its very limited ability to predict default makes it practically useless for automated risk assessment. It does not differentiate between individual loan products or between corporate and retail clients. Its heavy dependence on the interpretation of information by humans dramatically impairs its automatic interpretation capability as such. Therefore, the architecture of the whole system should be changed.

5. Proposals and contribution of suggested solutions

As has been already pointed out in the chapter “Problem analysis and current situation”, the existing system for assessing risk in loan clients has a number of weaknesses:

1. The model it uses has a low accuracy of default identification.
2. The credit risk scoring model is not clearly defined.
3. The overall assessment is in absolute values: larger clients generally collect negative events “more easily” than smaller ones, yet the size of the client is not taken into account and the indicators are not weighted based on the relative size of the clients.
4. The system does lend itself to automation only to a limited extent and it still relies on human interpretation.

To eliminate these drawbacks, advanced statistical methods need to be employed to find a suitable model.

5.1. Data

The Bank provided a sample containing a higher-than-normal proportion of default clients for the purpose of examining the relationships between the individual variables and the extent to which these predict default. The sample includes 1178 clients, out of which 910 are retail and 268 corporate. Each client in the dataset has only one loan, which enables us to control for additional influences while testing the individual parameters. As a result, the outcomes of the calculations are not distorted by the impact of combinations of variables.

5.2. Retail clients

Retail clients include natural persons, chiefly non-entrepreneurs. For the purposes of automated risk assessment of loan clients, self-employed persons will also be considered retail clients.

5.2.1. Selection of an appropriate credit risk model

Although there are a variety of approaches to credit risk assessment of retail clients (11), this thesis adheres to the methodology put forward by Bessis in 2010. For the retail segment, Bessis differentiates between “behavioural scoring models” and “origination scoring models” (5 p. 546).

Bessis defines a behavioral scoring models as “an atempt to model the behavior of existing clients, when there is no new event that would change the debt level, given historical data of account and loan behavior. Behavioural models apply to existing clients for whom there is historical data, say, at least 6 months. It makes it easier to deal with existing clients than new clients for whom there is no credit history” (5 p. 546).

By contrast, the origination model is more suitable for assessing new clients: “There are two types of origination models. For new clients, there is less information, although all banks would collect a minimum set of data on the client, such as revenue, wealth and, eventually, historical behaviour of other existing accounts in other banks. Therefore, we cannot use the same attributes for modelling their risk as with the existing clients. Consider an existing client that requests a new loan. A second type of origination model is required, because we already have historical data on the client. In this case, we have a different origination model, which applies to a known client whose credit standing might be affected by a new loan. It is also an origination model because “originating” to this existing client is considered” (5 p. 546).

Since the role of the Credit Risk Monitoring Information System is to monitor current clients and their loan burdens, the behavioural scoring model is the more suitable one. However, this type of model is more demanding in terms of client data. Bessis lists the following data as suitable for analysis (5 p. 547):

- Time series of flows, measured by the absolute value of flows, both negative and positive, and averaged over a period of the past 6 months;
- Number of debit days, measured by the maximum of debit days over the past 6 months;
- Number of transactions suspended by a credit officer because they would have triggered an excess overdraft;

- Count of incidents over the past 6 months;
- Amount of liquid savings – measure of wealth often known by the bank, with some average calculated from the end of the month average of balances over the past 6 months;
- Leverage ratio – monthly payments of due/credit flows;
- origination of the account;
- Other personal wealth characteristics.

Different data sets may be used for actual modelling. Appendix 5 gives a summary of the various approaches to variables commonly used in retail credit scoring models.

5.2.2. Data attributes and their information value

The individual data attributes, i.e. variables, need to be subject to discrimination in order to determine their importance. Discrimination amounts to calculating the information value of the variable in question (5 p. 547; 12 p. 8). This chapter will provide a calculation of the information value of the individual attributes based on the data sample provided by the bank.

According to Kočenda and Vojtek (12) a variable´s information value can be expressed as follows:

$$IV_i = \ln\left(Odds_i\right)\left(\frac{Defaulted_i}{Defaulted} - \frac{Good_i}{Good}\right)$$

Where $Defaulted_i$ represents the clients identified as default based on the variable (attribute) in question, and *Defaulted* represents all default clients of the data set. Similarly, $Good_i$ represents non-default clients identified in the same way and *Good* the sum of non-default clients in the entire data set. The information value expresses the predictive power of the variable for the given group (12 p. 8).

$Odds_i$ is the value expressing the discrimination ability of the variable in question for the given group. The vaulue of $Odds_i$ is given by the following formula:

$$Odds_i = \left(\frac{Defaulted_i}{Defaulted}\right)\left(\frac{Good}{Good_i}\right)$$

The interpretation of the variables in this formula is identical with that of the values in the information value formula above.

Kočenka and Vojtek note that: ”In banking practice a value above 0.2 is taken as a sign of the strong predictability of a given variable.” (12 p. 8). This thesis will use this value as a benchmark.

Calculations for the individual available variables in the examined data sample can be easily performed in the above way. For the sake of clarity, the calculation of the information value of the attribute “female” with regard to client risk is shown below as an example:

$$Odds_F = \left(\frac{70}{261}\right)\left(\frac{649}{201}\right) = 0.268199 \times 3.2288557 = 0.8659758756$$

$$IV_F = \ln(0.8659758756)\left(\frac{70}{261} - \frac{201}{649}\right) = -0.1438982281 \times (0.268199 - 0.30970724)$$
$$= -0.1438982281 \times (-0.04150824) = \underline{\underline{0.00597296}}$$

The calculation clearly shows that, with regard to the assessment of risk in a loan client, the information value of the attribute “female“ is very low. Interestingly, the research conducted by Kočenda and Vojtek (12) at an undisclosed Czech bank found out the information value of gender to be approximately 0.0230161. Although this value does not exceed the threshold of 0.2 either, it is still about 4 times higher than the value arrived at based on the data sample for this thesis. This suggests that although gender may not be a decisive factor, it is obviously a highly volatile one and that, as a result, it might perhaps not be adequate to use one value for this indicator across all Czech banks.

For the sake of completeness, let us now calculate the information value of the attribute “male”:

$$Odds_M = \left(\frac{191}{261}\right)\left(\frac{649}{448}\right) = 0.731800766 \times 1.44866 = 1.06013102$$

$$IV_M = \ln(1.06013102)\left(\frac{191}{261} - \frac{448}{649}\right) = 0.0583925 \times (0.731800766 - 0.690292758) =$$
$$= 0.0583925 \times 0.041508008 = \underline{\underline{0.00242375635714}}$$

Based on the above calculations, we arrive at the following information value of the attribute "gender":

$$IV_{GENDER} = IV_F + IV_{M=} 0.00597296 + 0.00242375635714 = \underline{0.008396446}$$

In our case, the information value of the attribute "gender" has thus been shown to be below the considerable predictive power threshold, which is in line with Kočenda and Vojtek (12) who also consider gender a non-discriminatory value, at least in the Czech Republic. Dinh and Kleimeier, however, note that gender remains a discriminatory variable in developing countries (13 p. 483).

Information values for all other variables, for which data were made available for the purposes of this thesis, have been calculated and the non-discriminatory variables eliminated in a similar fashion.

Another data attribute is the **Length of Relationship** in years. It represents the duration of the relationship between the bank and the client at the time of the loan application to the date of calculation. The information value of this attribute over the data sample examined in this thesis looks as follows:

Years	Clients	Default	Non-default	Information value
0	124	52	72	0.051695
1	567	130	437	0.052838
1-3	68	27	41	0.019862
3-5	50	16	34	0.001401
5-10	74	25	49	0.004827
>10	27	11	16	0.00938
Total	**910**	**261**	**649**	0.140003

Figure 5 - Length of Relationship

As shown in the table above, the overall information value of the Length of Relationship between the bank and its client at the time of the loan application is significantly more important than gender.

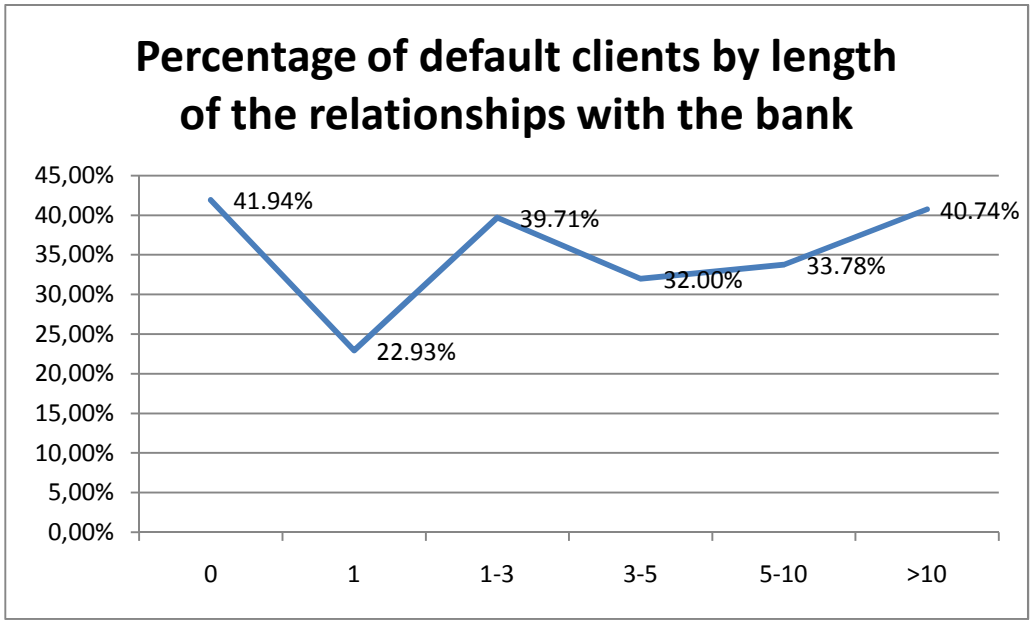


Figure 6 - Percentage of default clients by length of the relationships with the bank

The graph above clearly shows that the data sample does not allow for the conclusion to be made that a client who has been with the bank for a longer period of time at the time of his loan application is less likely to default on his loan than a client who has been with the bank for a shorter period.

Interestingly, the same calculation for a Czech bank carried out by Kočenda and Vojtek arrived at an information value of 0.601787, which implies very high predictive power. This suggests that the variable Length of Relationship is dependent on the specific client portfolio and may thus not be suitable for an indiscriminate application in risk assessment.

Kočenda and Vojtek (12) are strong supporters of making the "**Points**" variable part of the analysis. They define this variable as "the characteristics of a client's behavior in the current account" (12 p. 28). Unfortunately our data set does not comprise a sufficient amount of data, and there is no known way of constructing this variable retroactively and so, although Kočenda and Vojtek assert the information value of this variable to amount to 0.502122 (see Appendix 3), this thesis cannot take it into account.

Another attribute which could be used for risk analysis is the client´s **Date of Birth**. The following table gives a list of the calculated values.

Born up to year	Clients	Default	Non-default	Information value
1953	54	15	39	0.000117
1957	43	9	34	0.007489
1962	67	23	44	0.005330
1966	75	30	45	0.023051
1969	76	22	54	0.000014
1972	79	23	56	0.000039
1974	62	17	45	0.000263
1977	125	34	91	0.000732
1993	329	88	241	0.003300
Total	**910**	**261**	**649**	**0.040333**

Figure 7 - Date of Birth information value

Not surprisingly, in our case, the client´s age is not an absolutely decisive factor. However, the overall Date of Birth information value arrived at from our data is roughly identical with the information value reported by Kočenda and Vojtek, which is 0.047698 (see Appendix 3).

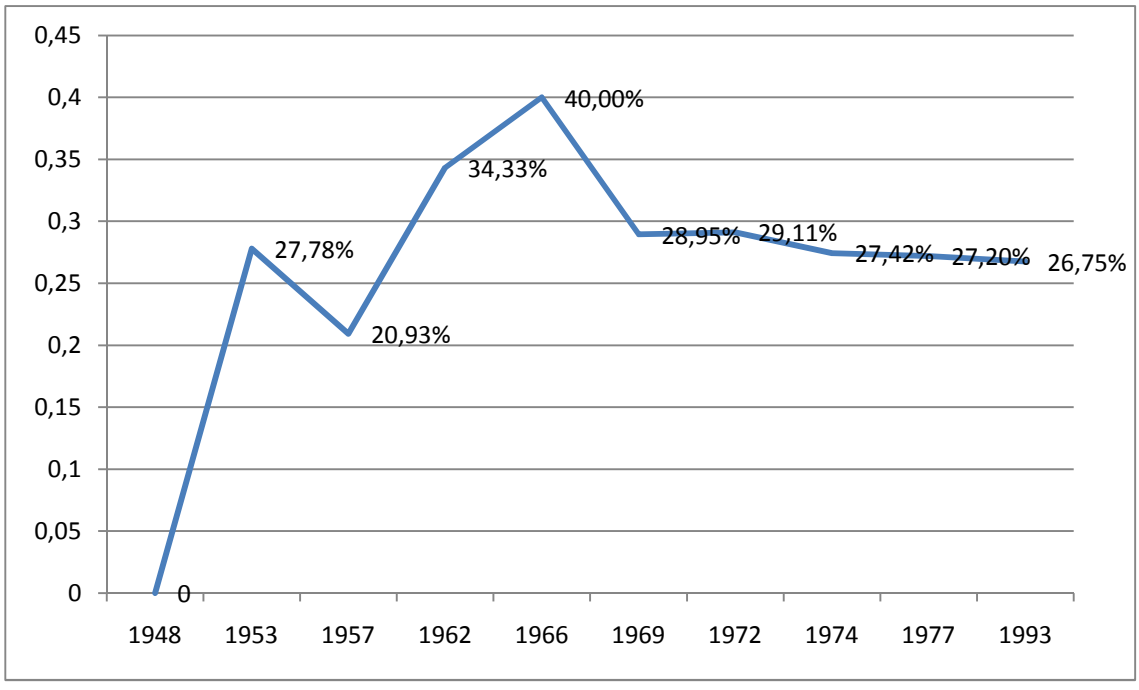


Figure 8 - Percentage of default clients by Date of Birth

The **Number of Years** from the client's opening of a current account at the bank proves to be an important attribute. The following table shows the calculations of the information values for the individual periods.

Years	Clients	Default	Non-default	Information value
1	124	4	120	0.422288
2	90	19	71	0.014909
3	240	62	178	0.005278
4	210	83	127	0.059395
5-6	121	44	77	0.017543
>6	125	49	76	0.03334
Total	**910**	**261**	**649**	**0.552754**

Figure 9 - Number of years from the opening of the current account as at 1 July 2011

Kočenda and Vojtek report an information value of 0.631346 for the attribute "number of years for which a person has been the bank´s client". The difference between the information value for this attribute arrived at from our data sample and the information value reported by Kočenda and Vojtek is not as substantial as the one observed for the Length of Relationship attribute. Consequently, the Number of Years attribute might be considered a generally strong variable, which could be used for defining scoring models at other Czech banks as well. The following graph illustrates an interesting growth trend: longer relationships between clients and the bank generally result in an increased client default risk for the bank.

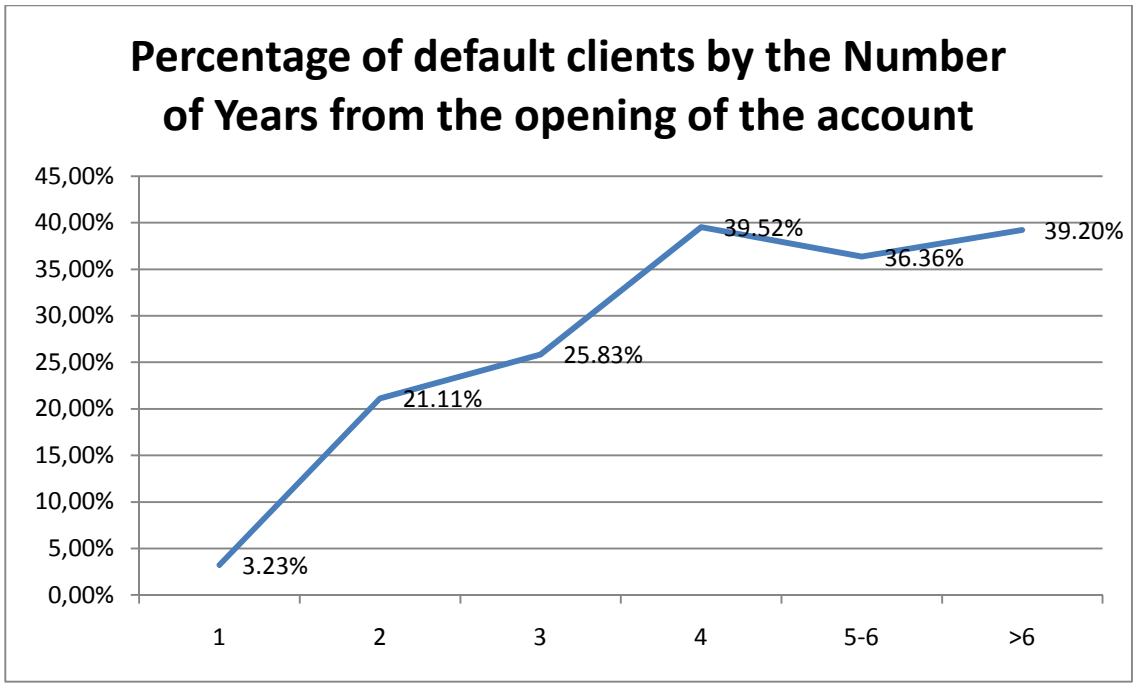


Figure 10 - Percentage of default clients by the Number of Years from the opening of the account

The information value for the **Amount of Loan** attribute was calculated by Kočenda and Vojtek at 0.123972 (see Appendix 3). In our case, however, the information value for this attribute is slightly higher, reaching the threshold of 0.2, at which the information value of a variable is considered high.

Amount of Loan in CZK	Clients	Default	Non-default	Information value
<100000	86	35	51	0.02967
>100000 and <300000	154	57	97	0.026142
>300000 and <800000	143	35	108	0.006975
>800000 and <1200000	133	26	107	0.032875
>1200000 and <1800000	142	38	104	0.001405
>1800000 and <3000000	147	28	119	0.040779
>3000000	105	42	63	0.032271
Total	**910**	**261**	**649**	**0.170118**

Figure 11 - Amount of Loan

The graph below shows default risk to be lowest between CZK 800,000 and 3,000,000.

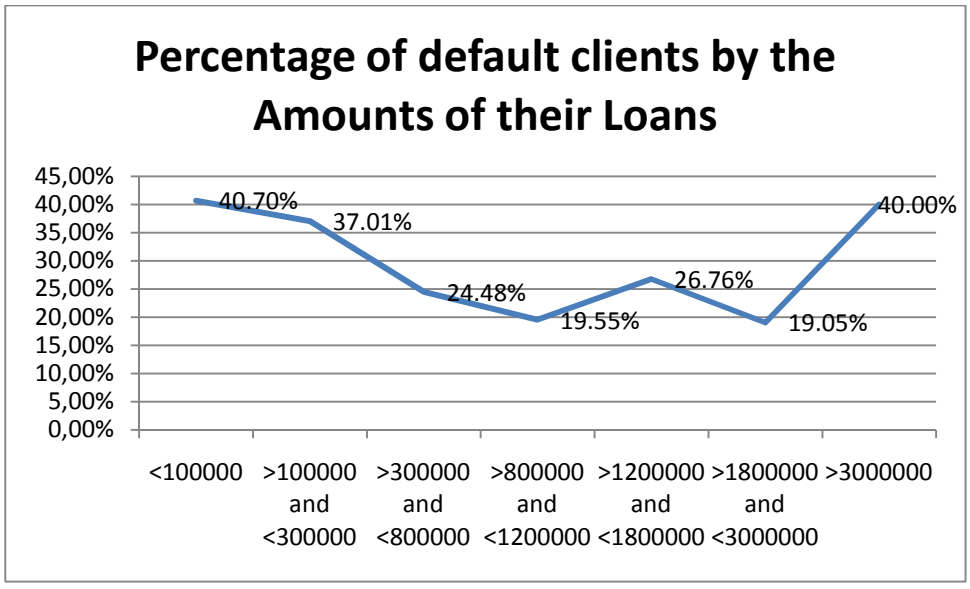


Figure 12 - Percentage of default clients by the Amounts of their Loans

For the **Type of Product** variable Kočenda and Vojtek report an information value of 0.022380 (see Appendix 3). Numbers from 1 to 4 are used to indicate different loan product types to ensure that data confidentiality is maintained and the data basis is clear.

Type of Product	Clients	Default	Non-default	Information value
1	23	10	13	0.011858
2	880	249	631	0.000346
3	3	1	2	0.000163
4	4	2	2	0.004173
Total	**910**	**262**	**648**	**0.016539**

Figure 13 - Type of Product

There is no major difference compared to the value arrived at by Kočenda and Vojtek (see Appendix 3), but the information value of the Product Type variable does not reach a value sufficient to consider it discriminatory.

Kočenda and Vojtek (12) also consider a **Region** attribute, which they ascertain from the postal code. However, the data available for this thesis do not include any information on regions; the region attribute will, therefore, not be included in our analysis. If Kočenda and Vojtek are right, then the omission of the Region attribute should not be a serious issue, as the information value they calculated based on their data only reaches 0.093896 (see Appendix 3), and is thus not discriminatory.

Marital Status is yet another variable used in our analysis. To protect client data, actual information on marital status has been transformed into a numerical index. This has, however, no bearing whatsoever on the process of variable discrimination.

Marital Status	Clients	Default	Non-default	Information value
1	177	66	111	0.032003
2	296	87	209	0.000390
3	320	75	245	0.024597
4	105	30	75	0.000003
5	12	3	9	0.000445
Total	**910**	**261**	**649**	**0.056993**

Figure 14 - Marital Status

The accuracy of risk calculation for a current loan client is obviously not too much affected by his or her Marital Status. Although Kočenda and Vojtek report a Marital Status information value of 0.112809 (see Appendix 3), the value in our case is about 50% lower.

The **Citizenship** attribute has proven irrelevant to risk assessment. The information values concerned are clearly presented in the following table.

Citizen of the CR	Clients	Default	Non-default	Information value
Yes	883	257	626	0.000415
No	27	4	23	0.016861
Total	**910**	**261**	**649**	**0.017276**

Figure 15 - Citizen of the Czech Republic

Likewise, the impact of the Residency attribute is minimal.

Residency	Clients	Default	Non-default	Information value
Czech Republic	893	258	635	0.00010326
Outside Czech Republic	17	3	14	0.00634405
Total	**910**	**261**	**649**	**0.00644731**

Figure 16 - Residency

By contrast, the information value of the **Number of Persons in Joint Household** attribute shows that this variable is highly discriminatory. Important note is that number of persons in joint household does not include applicant itself.

Persons in Joint Household	Clients	Default	Non-default	Information value
0	481	167	314	0.043611
1	200	52	148	0.003891
2	106	21	85	0.024610
3	63	9	54	0.042917
4	53	10	43	0.015304
5	7	2	5	0.000000
Total	**910**	**261**	**649**	**0.130333**

Figure 17 – Number of Persons in Joint Household

The following graph clearly shows that the repayment reliability increases, up to an extent, with the number of persons living in a joint household with the client. With 4 and more persons in a joint household, the chance of default starts rising again. This may be caused by the higher cost of living faced by more numerous households in real terms.

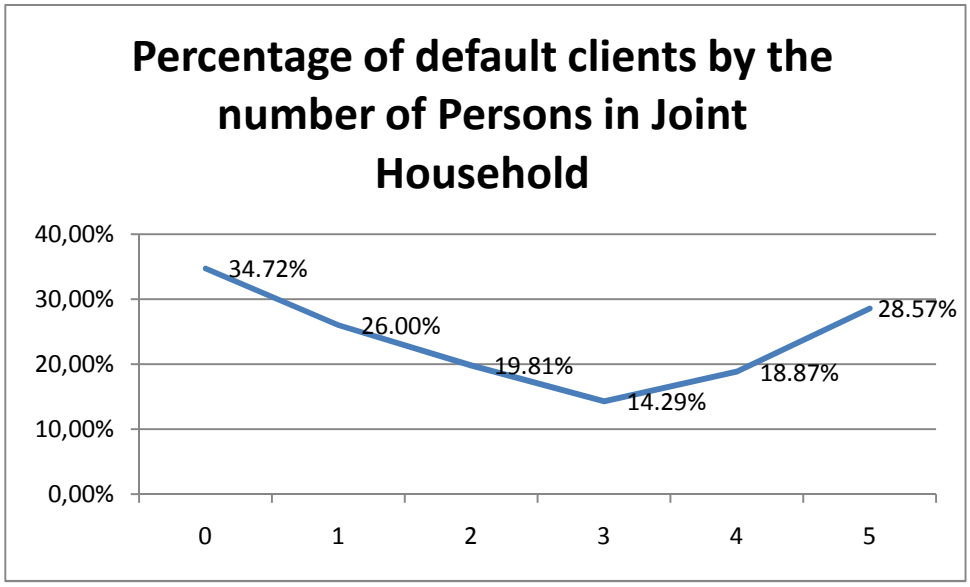


Figure 18 - Percentage of default clients by the number of Persons in Joint Household

The client's **Housing Type** proves to be a highly discriminatory variable. To protect client data, the values have been transformed into a numerical index, which does, however, not affect the values calculated. This variable clearly exceeds the recommended threshold. A comparison with another data set would be very interesting; unfortunately Kočenda and Vojtek (12) did not include this attribute in their analysis.

Housing Type	Clients	Default	Non-default	Information value
0	481	168	313	0.046591
1	205	28	177	0.154368
2	70	22	48	0.001351
3	93	27	66	0.000030
4	4	3	1	0.020002
5	15	1	14	0.030658
6	28	6	22	0.004237
7	14	6	8	0.009235
Total	**910**	**261**	**649**	**0.266471**

Figure 19 - Housing Type

In their paper Kočenda and Vojtek (12) emphasized Education as a major reliability indicator of a client. More educated clients tend to default on their loans less often.

Education	Clients	Default	Non-default	Information value
0	486	169	317	0.044842
1	10	6	4	0.022148
2	94	33	61	0.009622
3	176	34	142	0.045908
4	27	5	22	0.008413
5	115	13	102	0.123364
6	2	1	1	0.002087
Total	**910**	**261**	**649**	**0.256383**

Figure 20 - Education

In the case of Volksbank data, education clearly has a highly discriminatory value; although it does not reach the level of 0.359725 reported by Kočenda and Vojtek for their data set (see Appendix 3). The graph vividly demonstrates that the percentage of loan clients in default falls with higher education, notwithstanding the sudden rise of default at 6, which is to be attributed to the small size of the data sample (only two clients) and seems to be too insignificant to refute the whole trend. Again, the data have been transformed into numerical values to protect client privacy.

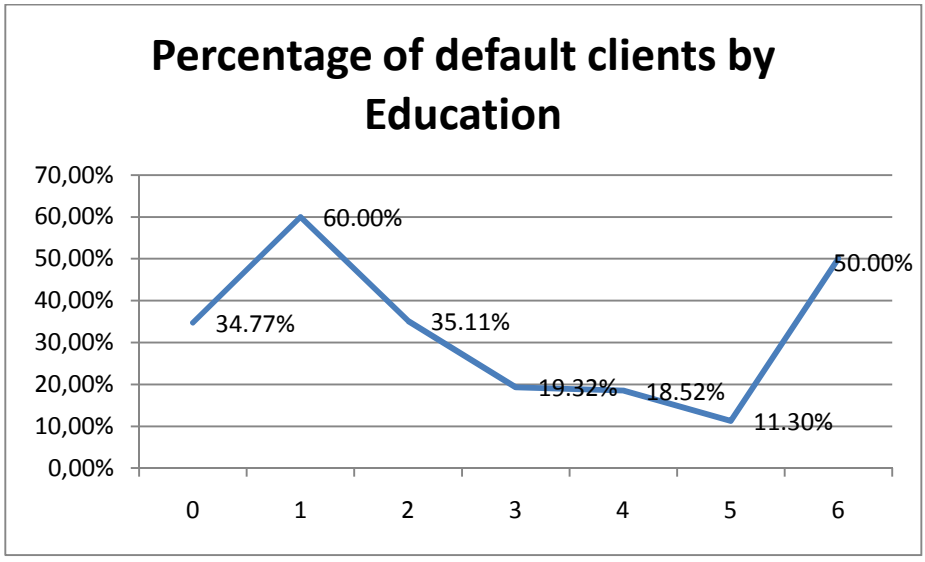


Figure 21 - Percentage of default clients by Education

The **Amount of Loan Instalment** is a very interesting variable. This variable approaches the high discrimination threshold. The following table shows the results of the calculation of the

variable´s information value. The individual instalment ranges are not included to protect client data confidentiality.

Amount of Instalment	Clients	Default	Non-default	Information value
Range 1	426	150	276	0.045004
Range 2	176	53	123	0.000935
Range 3	142	21	121	0.089062
Range 4	120	24	96	0.026605
Range 5	46	13	33	0.000021
Total	**910**	**261**	**649**	**0.161627**

Figure 22 - Amount of Loan Instalment

One would perhaps expect higher instalments to result in a higher chance of default. However, the graph clearly demonstrates a different trend. Paradoxically, the chance of default is highest for the lowest amounts of instalment. It should be noted that the data also includes consumer loans, which the bank may regard as more risky in terms of a possible default compared to, say, mortgages. As in the case of the number of Persons in a Joint household, however, the falling trend reverses, and in the medium range the number of defaults starts rising again. The Amount of Instalment should thus be taken into account when assessing the risk of a loan client defaulting.

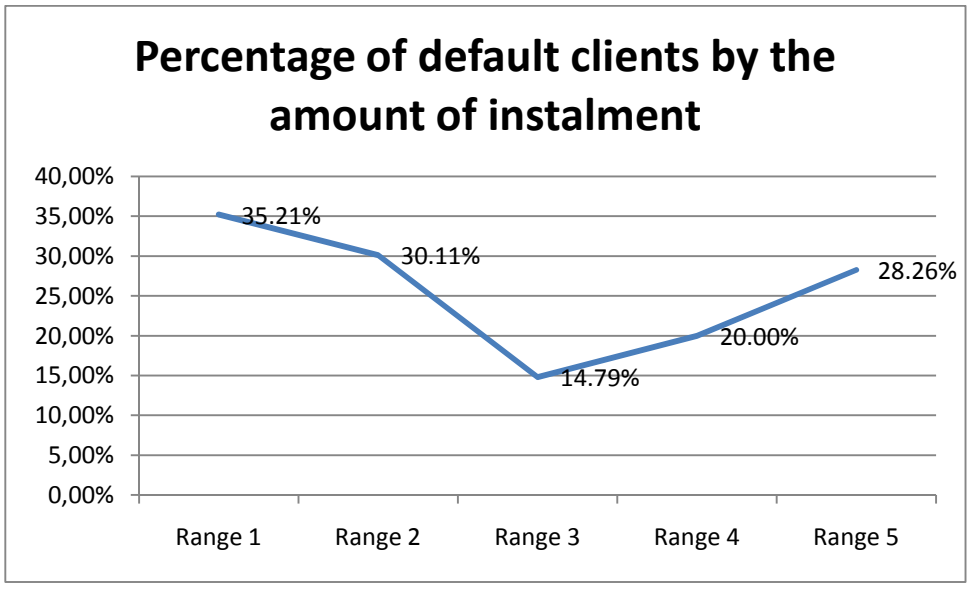


Figure 23 - Percentage of default clients by the Amount of Instalment

The following table shows that the **Employee** attribute, representing whether a loan client is an employee, is relatively discriminatory compared with other variables.

Employee	Clients	Default	Non-default	Information value
1	245	87	158	0.028243
2	469	100	369	0.073190
3	196	74	122	0.039264
Total	**910**	**261**	**649**	**0.140697**

Figure 24 – Employee

By contrast, the variables representing whether the client is an **Entrepreneur** (also in parallel with employment, if appropriate) or a member of the liberal profession both have almost identical information values and are discriminatory only to a minimum extent.

The following table shows the information value of the Entrepreneur attribute.

Entrepreneur	Clients	Default	Non-default	Information value
1	106	30	76	0.000040
2	608	157	451	0.013476
3	196	74	122	0.039264
Total	**910**	**261**	**649**	**0.052780**

Figure 25 – Entrepreneur

Similarly, the table below presents the calculation of for the **Liberal Profession** attribute. As can be seen, the overall information values are indeed almost the same.

Liberal Profession	Clients	Default	Non-default	Information value
1	709	185	524	0.012838
2	5	2	3	0.001537
3	196	74	122	0.039264
Total	**910**	**261**	**649**	**0.053639**

Figure 26 – Liberal Professions

Kočenda and Vojtek (12) consider the **Number of Employments** variable non-discriminatory with an information value of only 0.021004 (see Appendix 3). The result for Volksbank data is included in the following table. The actual number of employments is again in numerical

indexes, and the value 0 also includes the clients for which the number of employments is not known, as this may not have been a required piece of information for the provision of a loan.

Number of Employments	Clients	Default	Non-default	Information value
0	668	213	455	0.017471
1	4	2	2	0.004173
2	195	37	158	0.054990
3	41	8	33	0.010222
4	2	1	1	0.002087
Total	**910**	**261**	**649**	**0.088943**

Figure 27 - Number of Employments

Compared to the information value reported for this attribute by Kočenda a Vojtek (12) the information value in our case in about four times higher. That still does not make it highly discriminatory. This is, however, not the only difference between this thesis and Kočenda and Vojtek (12) with regard to this variable – they define it as: "The total number of employments in the last 3 years" (12 p. 26). The data Volksbank made available for the purposes of this includes the number of employments in the last 2 years. As a result a comparison of these values is not easily possible.

The graph below once again shows that the default rate is lower for medium values. Perhaps effective workers are able to change jobs reasonably often – not too often, and not too seldom. To have had two employers over the last 2 years is not terribly difficult. Consider a situation where the client asks for the loan, receives it and then, after 6 months, changes his job. This would bring the client to 2 employers in the last 2 years, while it is obviously not something negative.

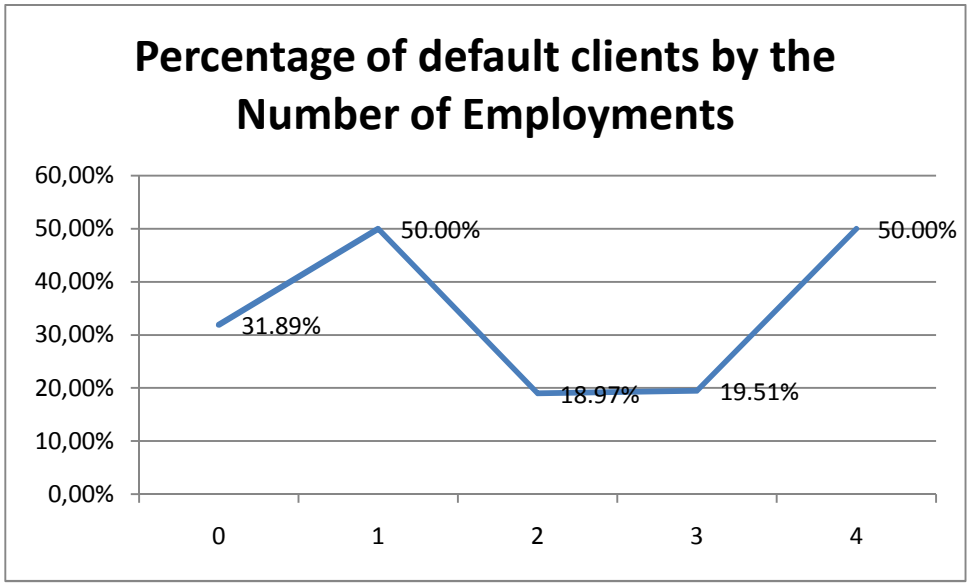


Figure 28 - Percentage of default clients by the Number of Employments

The **Employment Contract** attribute has a slightly higher information value; it indicates whether the client in question has an employment contract or not or whether the relevant information is at all available. Although banks tend to prefer their mortgage loan applicants to have employment contracts, the real information value of this attribute with respect to a client's default is not very discriminatory.

Employment Contract	Clients	Default	Non-default	Information value
1	550	187	363	0.038914
2	329	67	262	0.066550
3	31	7	24	0.003264
Total	**910**	**261**	**649**	**0.108727**

Figure 29 - Employment Contract

By contrast, **Salary** is a highly discriminatory variable. Kočenda and Vojtek (12) do not take this variable into account, but the research presented here suggests it is fairly discriminatory.

Salary	Clients	Default	Non-default	Information value
Range 1	596	206	390	0.051351
Range 2	148	41	107	0.000376
Range 3	166	14	152	0.266139
Total	**910**	**261**	**649**	**0.317867**

Figure 30 - Salary

To protect client data confidentiality, the salary bands are denoted by ”Range 1” to “Range 3” without actual figures being indicated. The following graph shows the direct link between salary and the risk of default.

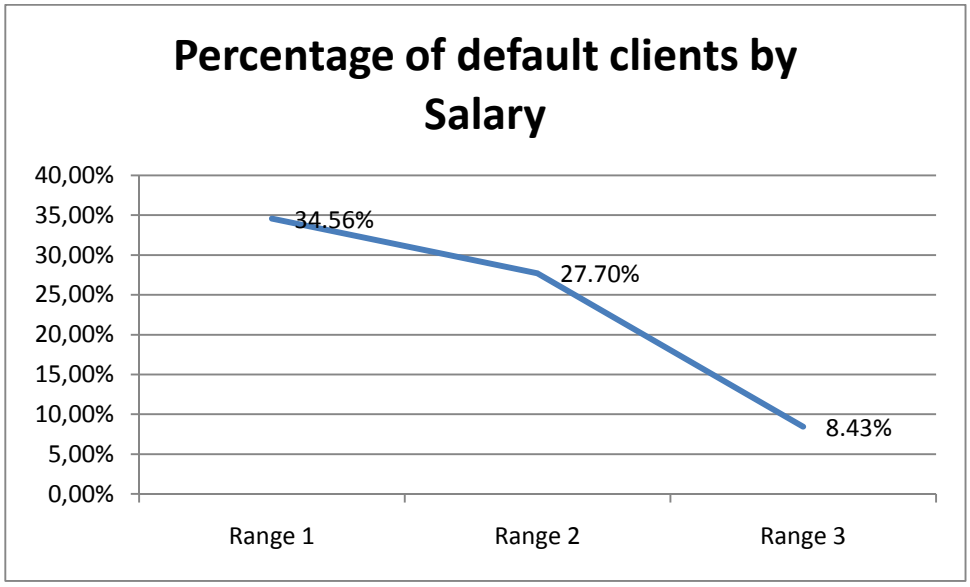


Figure 31 - Percentage of default clients by Salary

The information value of the **Current Account** variable amounts to a little more than half that of the Salary attribute. The Current Account variable indicates that a loan client also has a current account with Volksbank. The table below lists the information values calculated for the individual options.

Current Account	Clients	Default	Non-default	Information value
1	204	78	126	0.045164
2	399	78	321	0.098626
3	307	105	202	0.023364
Total	**910**	**261**	**649**	**0.167154**

Figure 32 - Current Account

Kočenda and Vojtek (12) consider the **Own Resources** attribute to be the most discriminatory variable with an information value of 1.462601. Nevertheless, Volksbank data, based on which the information value of **Own Resources** has been calculated at 0.095001 – well under the high discrimination threshold – show that a high information value of this attribute need not be a rule of thumb.

Own Resources	Clients	Default	Non-default	Information value
Range 1	786	245	541	0.012482
Range 2	79	10	69	0.069405
Range 3	45	6	39	0.035653
Total	**910**	**261**	**649**	**0.117539**

Figure 33 - Own Resources

Regular Income can also be considered an essential variable. For Volksbank data, the information value of this variable amounts to 0.310012, which means that this attribute is a highly discriminatory one. The values calculated for this attribute are shown in the table below.

Regular Income	Clients	Default	Non-default	Information value
Type 1	823	258	565	0.014983
Type 2	44	1	43	0.177927
Type 3	43	2	41	0.117102
Total	**910**	**261**	**649**	**0.310012**

Figure 34 - Regular Income

Unfortunately, it was not possible to retrieve other potentially relevant variables, such as Purpose of Loan or Sector of Employment, from the data available for the purposes of this

thesis. The overall table listing the information values of all attributes discussed in this thesis looks as follows:

Variable	Information value
Number of Current Account Years	0.552754
Salary	0.317867
Regular Income	0.310012
Housing Type	0.266471
Level of Education	0.256383
Amount of Loan	0.170118
Current Account	0.167154
Amount of Loan Instalment	0.161627
Employee	0.140697
Length of Relationship	0.140003
Number of Persons in Joint Household	0.130333
Own Resources	0.117539
Employment Contract	0.108727
Number of Employments	0.088943
Marital Status	0.056993
Liberal Profession	0.053639
Entrepreneur	0.052780
Date of Birth	0.040333
Citizenship	0.017276
Type of product	0.016539
Gender	0.008396
Residency	0.006447

Figure 35 - Information values arrived at

5.2.3. Fuzzy logic

Fuzzy logic, developed by L. Zadeh in 1965, works with what are called vague sets. These sets have a better correspondence to real world situations than the value sets used by classical

logic. As opposed to a logic with clear-cut criteria, where an element either is or is not part of a set, fuzzy logic differentiates between various degrees of set membership (14).

The difference between fuzzy logic and Boolean logic, where data needs to be categorized and weighting needs to be given to the individual categories, consist in the very approach to categories. In fuzzy logic, there are no categories: a fuzzy model does not weight values within the limits of distinct intervals but on a continuous basis (14).

Dostál notes that: “Fuzzy logic enables us to find a solution to a given case based on the rules defined for similar cases. The fuzzy method, which uses fuzzy sets, is a method that can be used in the area of corporate management“ (14 p. 8).

Dostál and Sojka describe fuzzy processing as an operation with three steps: “The fuzzy logic consists of three fundamental steps: fuzzyfication, fuzzy inference and defuzzification.” (4 p. 62).

Fuzzyfication transforms real variables into language ones. Language variables are based on linguistic variables: “The definition of language variables draws on linguistic variables, for instance the variable “Risk” can have the following attributes: zero, very low, low, medium, high, very high. Usually three to seven attributes are used for a variable” (14 p. 11).

Dostál and Sojka define fuzzy inference as: “System behaviour by means of the rules of the type IF THEN. The conditional clauses create these algorithms, which evaluates the input variables” (4 p. 63).

Defuzzification is understood by Dostál and Sojka as the verbal interpretation of the values arrived at: “The third step (defuzzification) means the transformation of numerical values to linguistic ones. The linguistic values can be, e.g. for variable risk very low, low, medium, high, very high risk” (4 p. 63).

5.2.4. Fuzzy model

To begin with, real variables to be used for fuzzyfication need to be defined. Variables with an information value higher than 0.2. can be selected based on Figure 35 – “Information values arrived at”. As there are only five such variables, the information value threshold needs

to be reduced, as in the case of Kočenda and Vojtek (12), to 0.1. This will give us a total of 13 variables with an overall information value of 2.839685.

Variable	Information value
Number of Current Account Years	0.552754
Salary	0.317867
Regular Income	0.310012
Housing Type	0.266471
Level of Education	0.256383
Amount of Loan	0.170118
Current Account	0.167154
Amount of Instalment	0.161627
Employee	0.140697
Length of the Relationship	0.140003
Number of Persons in Joint Household	0.130333
Own Resources	0.117539
Employment Contract	0.108727
Total	**2.839685**

Figure 36 - Variables selected for fuzzyfication

As a first step, a transformation matrix needs to be created. The transformation matrix needs to include the individual variables and numerically defined degrees of risk. As has been noted above, a variable's information value represents its predictive power. We will use that information value to define the degrees of risk.

If we know the total information value to be 2.839685, we can easily compute the percentage weighting of the individual variables in the transformation matrix as the quotient of the information value (IV) of a variable by the total information value. For instance, the weighting for the Number of Current Account Years looks as follows:

$IV_{Years} = 0.552754 / 2.839685 = 0.19465328$

Thus we arrive at a figure of about 19.47 %. The values for the individual variables listed in the table below have been calculated in a similar way.

Variable	IV	Result	Percentage
Number of Current Account Years	0.552754	0.194653	19.47%
Salary	0.317867	0.111937	11.19%
Regular Income	0.310012	0.109171	10.92%
Housing Type	0.266471	0.093838	9.38%
Level of Education	0.256383	0.090286	9.03%
Amount of Loan	0.170118	0.059907	5.99%
Current Account	0.167154	0.058864	5.89%
Amount of Instalment	0.161627	0.056917	5.69%
Employee	0.140697	0.049547	4.95%
Length of the Relationship	0.140003	0.049302	4.93%
Number of Persons in Joint Household	0.130333	0.045897	4.59%
Own Resources	0.117539	0.041392	4.14%
Employment Contract	0.108727	0.038288	3.83%
Total	**2.839685**	**1**	**100.00%**

Figure 37 - Percentage weighting of the variables in the transformation matrix

The graph below clearly shows the discriminatory power of the selected variables.

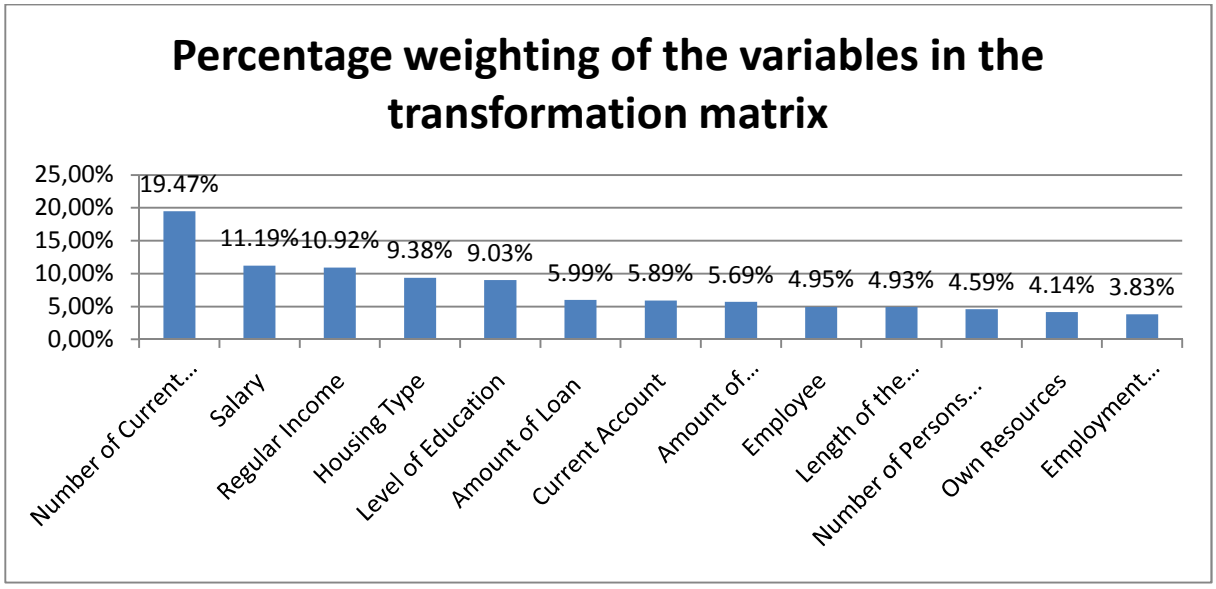


Figure 38 - Percentage weighting of the variables in the transformation matrix

Now that the weighting distribution has been set, the degree of risk needs to be expressed numerically. The calculation consists of performing a dot product operation based on a state

matrix where the percentage degree of total risk arrived at is given by the sum of the maximum numerical risk values associated with the variables in the transformation matrix.

This is why it is, to begin with, useful to select the sum of the maximum numerical risk values associated with the variables in the transformation matrix. The number 1000 has been selected for this purpose in order to keep matters simple while maintaining a sufficient level of detail. A simple calculation reveals the maximum value for each variable. For instance, for the Salary variable the result is: 0.1119 x 1000 = 111.9 and 112, if rounded up to the next integer. The results listed in the table below have been arrived at in a similar way.

Value	Max Value
Number of Current Account Years	195
Salary	112
Regular Income	109
Housing Type	94
Level of Education	90
Amount of Loan	60
Current Account	59
Amount of Instalment	57
Employee	50
Length of the Relationship	49
Number of Persons in Joint Household	46
Own Resources	41
Employment Contract	38
Total	**1000**

Figure 39 - Maximum values of the variables in the transformation matrix

These values are equal to the numerical degree of risk associated with the maximum variable value. Given that we know the individual numbers of well-performing and defaulting clients for the Salary variable, we arrive at the following table:

Salary	Clients	Default	Non-default
Range 1	596	206	390
Range 2	148	41	107
Range 3	166	14	152
Total	**910**	**261**	**649**

Figure 40 - Number default/non-default clients for the Salary variable

The percentage of the individual Salary ranges in the transformation matrix can be easily established: If we know the maximum value to be 112, we can find the highest value for the ratio of default clients in a specific range to all default clients.

Thus we get:

$\frac{206}{261} = 0.789272$

That means that the maximum value of the Salary variable represents 78.93 % of the client default risk. We can then easily compute the 100 % basis of this variable:

$$\frac{112}{0.789272031} = 141.9029$$

When this figure is rounded up, the sum of the numerical risk values reaches 142. Let us now check the calculation for correctness:

$142 \times 0.789272 = 112.076628402$

When the result is rounded up, we arrive at the original 112. The values for Range 2 can be calculated in the same way:

$$\frac{41}{261} = 0.157088123$$

$$142 \times 0.157088123 = 22.30651341 \approx 22$$

And the same goes for Range 3:

$$\frac{14}{261} = 0.053639847$$

$$142 \times 0.053639847 = 7.616858238 \approx 8$$

The correctness of the calculation can be checked by adding together Range1 + Range2 + Range3:

$$112 + 22 + 8 = \underline{\underline{142}}$$

In conclusion, the numerical risk values associated with the Salary variable are as follows:

Salary	Numerical risk value
Range 1	112
Range 2	22
Range 3	8
Total	142
Max	112

Figure 41 - Numerical risk values for the Salary variable

The numerical risk values for the other variables are arrived at in a similar way.

Variable	1	2	3	4	5	6	7	8	Sum	Max
Number of Current Account Years	9	45	146	195	103	115			613	195
Salary	112	22	8						142	112
Regular Income	109	1	1						111	109
Housing Type	94	16	12	15	2	1	3	3	146	94
Level of Education	90	3	17	18	3	7	1		139	90
Amount of Loan	37	60	37	27	40	30	4 4		275	60
Current Account	44	44	59						147	59
Amount of Instalment	57	20	8	9	5				99	57
Employee	44	50	37						131	50

Variable	1	2	3	4	5	6	7	8	Sum	Max
Length of the Relationship	20	49	10	6	9	4			98	49
Number of Persons in Joint Household	46	14	6	2	3	1			72	46
Own Resources	41	2	1						44	41
Employment Contract	38	13	2						53	38
Total									2070	**1000**

Figure 42 - Numerical risk values for selected variables

Given our knowledge of the transformation matrix, we can easily fill in the values for the individual possibilities. Let us now check the success rate of this transformation matrix.

The test of the information matrix has been carried out in Microsoft Excel. The values have been placed in the appropriate groups and the individual columns have been given a numerical risk value. For instance, the following formula has been used to calculate the Number of Current Account Years variable:

=IF(E270<2;9;IF(E270=2;45;IF(E270=3;146;IF(E270=4;195;IF(E270=5;103;IF(E270=6;103; IF(E270>6;115;9)))))))

Let us now demonstrate the entire calculation on an example. Note that the client used in the example is not an actual client of the Bank.

Variable	1	2	3	4	5	6	7	8	Points	Max
Number of Current Account Years	9	**45**	146	195	103	115			45	195
Salary	**112**	22	8						112	112
Regular Income	**109**	1	1						109	109
Housing Type	94	16	12	**15**	2	1	3	3	15	94
Level of Education	90	3	17	**18**	3	7	1		18	90

Amount of Loan	37	60	37	27	**40**	30	44		40	60
Current Account	44	**44**	59						44	59
Amount of Instalment	**57**	20	8	9	5				57	57
Employee	44	**50**	37						50	50
Length of the Relationship	20	49	**10**	6	9	4			10	49
Number of Persons in Joint Household	46	**14**	6	2	3	1			14	46
Own Resources	**41**	2	1						41	41
Employment Contract	**38**	13	2						38	38
Total									593	**1000**

Figure 43 - State matrix for an example client

By applying the scalar operation we arrive at the value of 593.

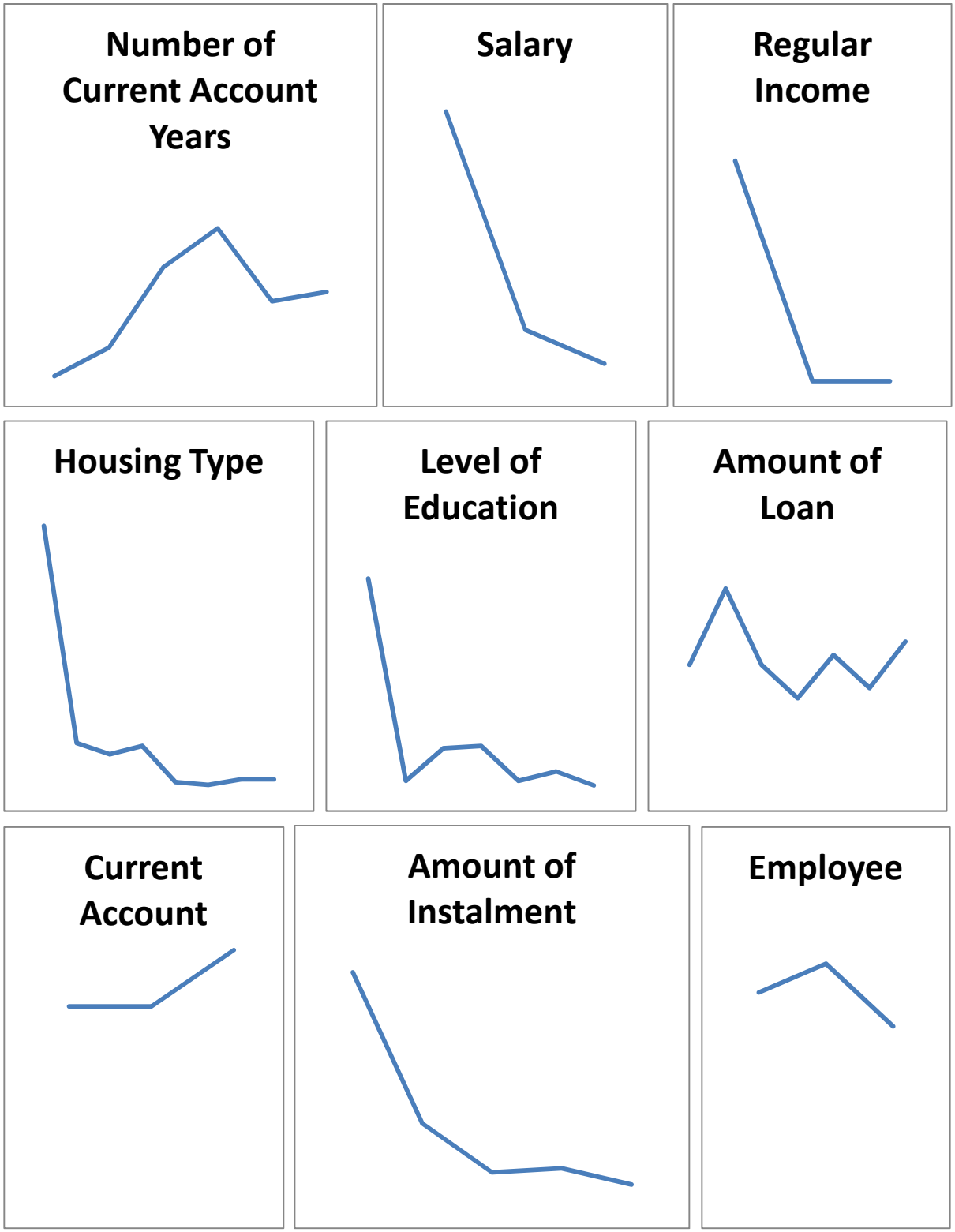
$$R = 1 \times 45 + 1 \times 112 + 1 \times 109 + 1 \times 15 + 1 \times 18 + 1 \times 40 + 1 \times 44 + 1 \times 57 + 1 \times 50 + 1 \times 10 + 1 \times 14 + 1 \times 41 + 1 \times 38 = \underline{\underline{593}}$$

By adding together all the values calculated for the variables in the transformation matrix together and dividing the sum of the maximum numerical risk values for each variable, we arrive at a percentage, which we will further interpret in a retransformation matrix.

For our example client the percentage arrived at 59.3%:

$$100 \times 593 \div 1000 = 59.3\ \%$$

Following figures represents membership functions for each variable (14 p. 14):


Number of Current Account Years
Salary
Regular Income
Housing Type
Level of Education
Amount of Loan
Current Account
Amount of Instalment
Employee

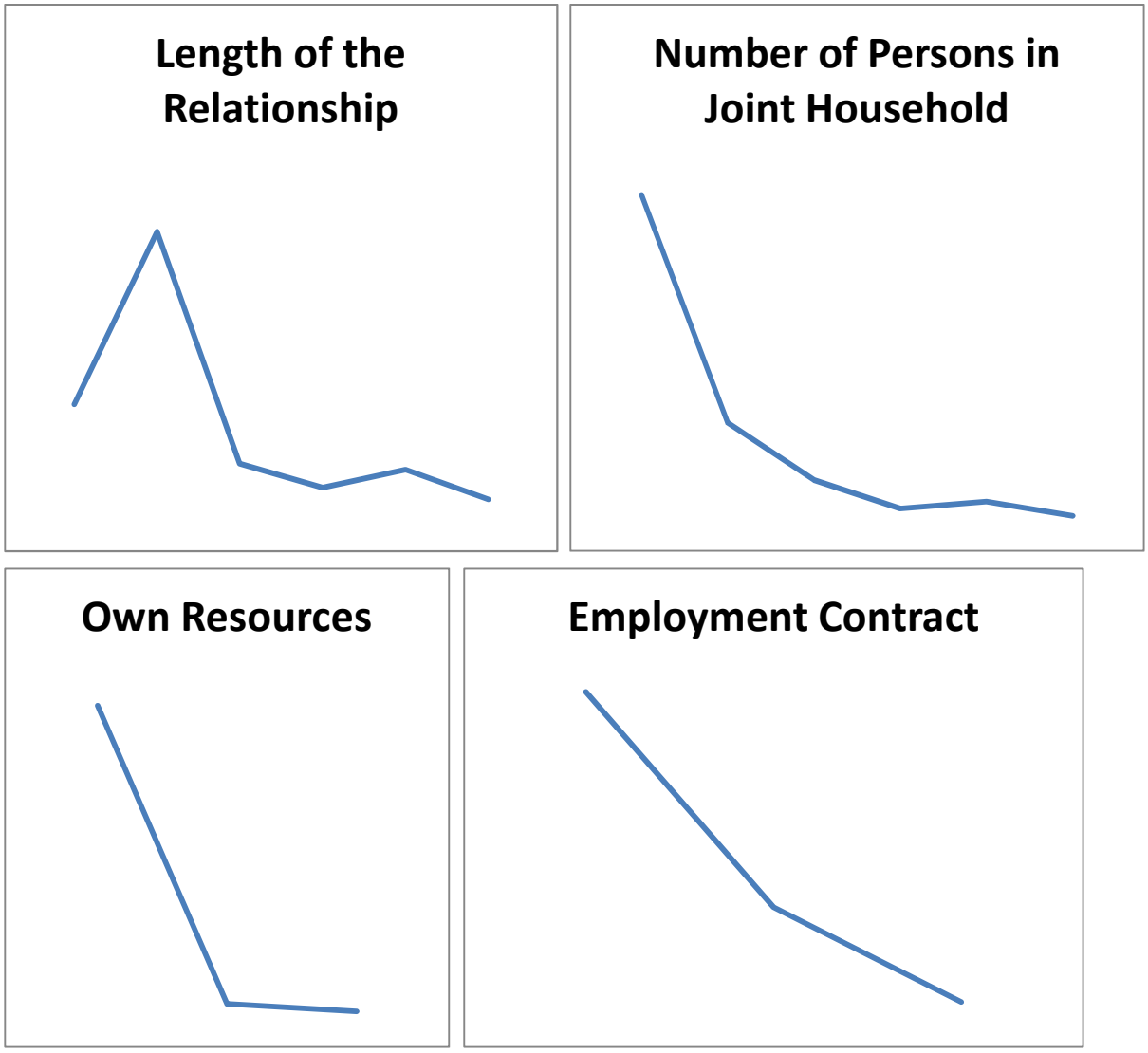


Figure 44 - Variables membership functions

5.2.5. Result of the fuzzy model

To be able to interpret the percentages arrived at in the way described above a retransformation matrix needs to be defined. In order to be able to compare the assessment produced by the original Credit Risk Monitoring Information System with the model proposed by this thesis, we will stick to the linguistic variables of Default and Non-default Clients.

Let there be the following retransformation matrix:

Percentage %	Linguistic variable
0-80	Non-default client
80-100	Default client

Figure 45 - Retransformation matrix

Having calculated the percentage for all selected clients and inserted the percentages in the retransformation matrix, we arrive at the following results:

Default	80 % Threshold	Result	Number of clients
TRUE	TRUE	TRUE	166
TRUE	FALSE	FALSE, TYPE I ERROR	95
FALSE	FALSE	TRUE	365
FALSE	TRUE	FALSE, TYPE II ERROR	284

Figure 46 - Results for 80% threshold

The test of the existing Credit Risk Monitoring Information System in Section 4.3 showed a success rate of default client detection of 14.7 %. The fuzzy model has been based on a data sample of 261 default and 649 non-default clients and its success rate of default client detection is about 63.6 %.

This rate is of course higher than that of the existing Credit Risk Monitoring Information System; however, the fuzzy model has also a considerably higher error-rate: the percentage of actual non-default clients identified as default (TYPE II ERROR) is 43.76 %.

The percentage of actual default clients not identified as such by the fuzzy model (TYPE I ERROR) is about 36.4 %. The percentage of successfully identified non-default clients is 56.24 %. To obtain a general picture of the success rate, we can take the sum of the successfully identified non-default and default clients and divide it by the total number of clients:

$$\frac{365 + 166}{910} \times 100 = 58.35\ \%$$

This way, we obtain an overall success rate of the fuzzy model of 58.35 %. However, we also need to consider the results generated by retransformation matrices with different thresholds. The table below indicates the values for a default threshold of 50 % and more.

Default threshold %	Non-default	Default	TYPE I ERROR	TYPE II ERROR
50	179	240	21	470
60	300	200	61	349
70	340	177	84	309
75	352	169	92	297
80	365	166	95	284
85	460	106	155	189
90	532	65	196	117
95	608	27	234	41

Figure 47 - Fuzzy model values for various retransformation matrices

The graph clearly shows that the accuracy of non-default client detection increases with the default threshold:

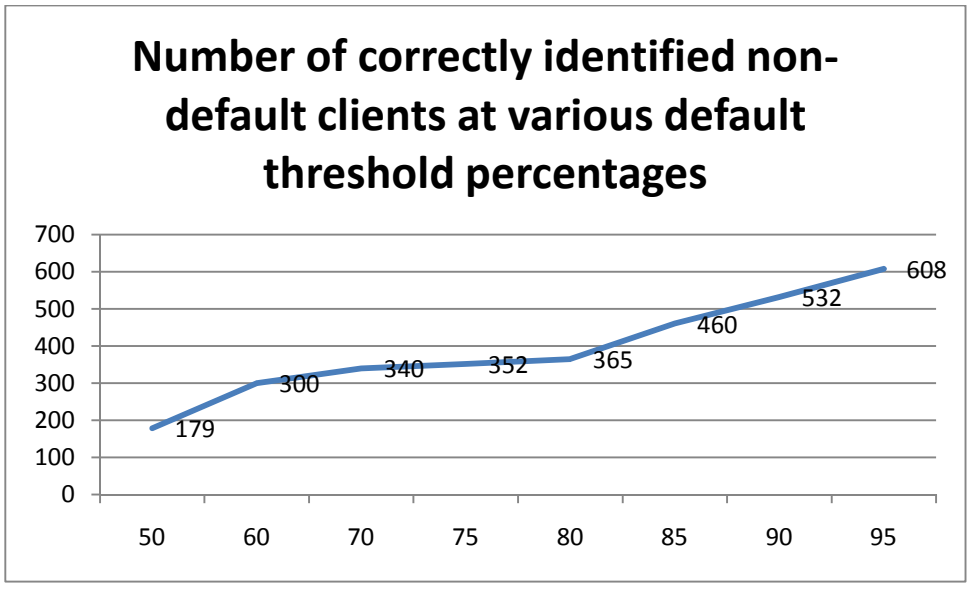


Figure 48 - Number of correctly identified non-default clients at various default threshold percentages

On the other hand, the number of successfully identified default clients decreases substantially as the percentage threshold for default client identification increases.

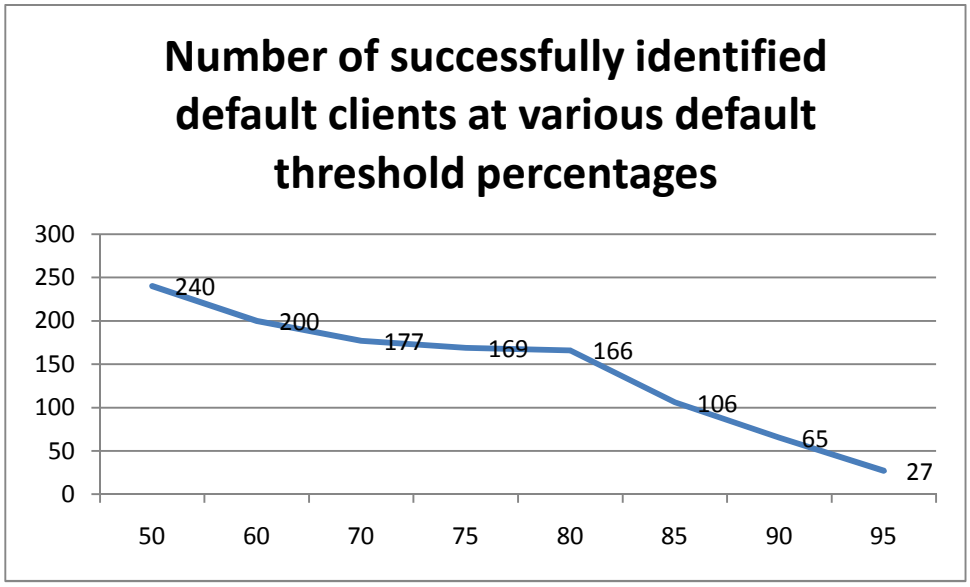


Figure 49 - Number of successfully identified default clients at various default threshold percentages

The curve for default clients not successfully identified (TYPE I ERROR) is very similar to the curve for successfully identified non-default clients. This is logical: given the fact that the curve for successfully identified default clients slopes downwards with an increasing default threshold, the trend of the TYPE I ERROR curve needs to be exactly opposite.

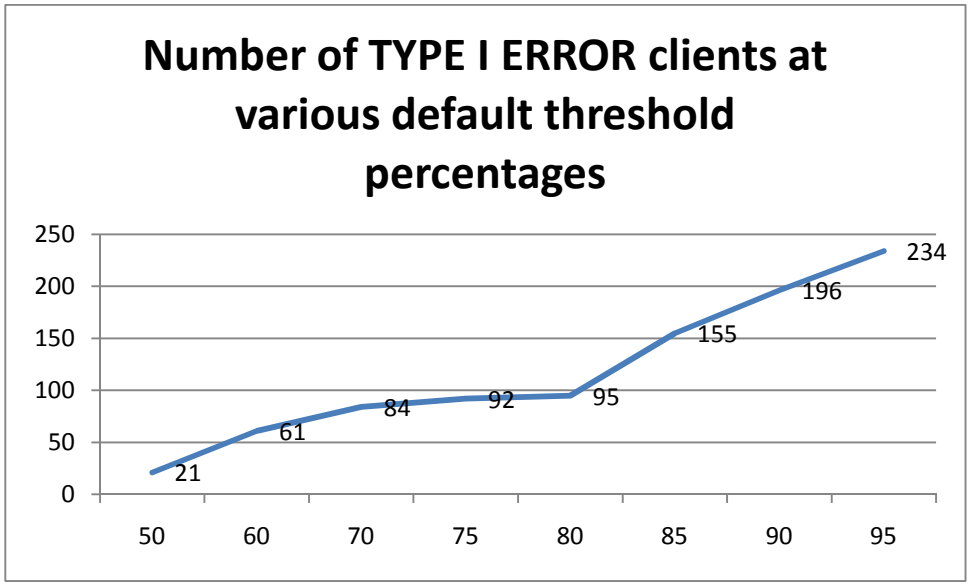


Figure 50 - Number of TYPE I ERROR clients at various default threshold percentages

The trend of the curve for TYPE II ERROR clients, that is to say actual non-default clients identified as default, is the same as the trend of the curve for successfully identified default clients.

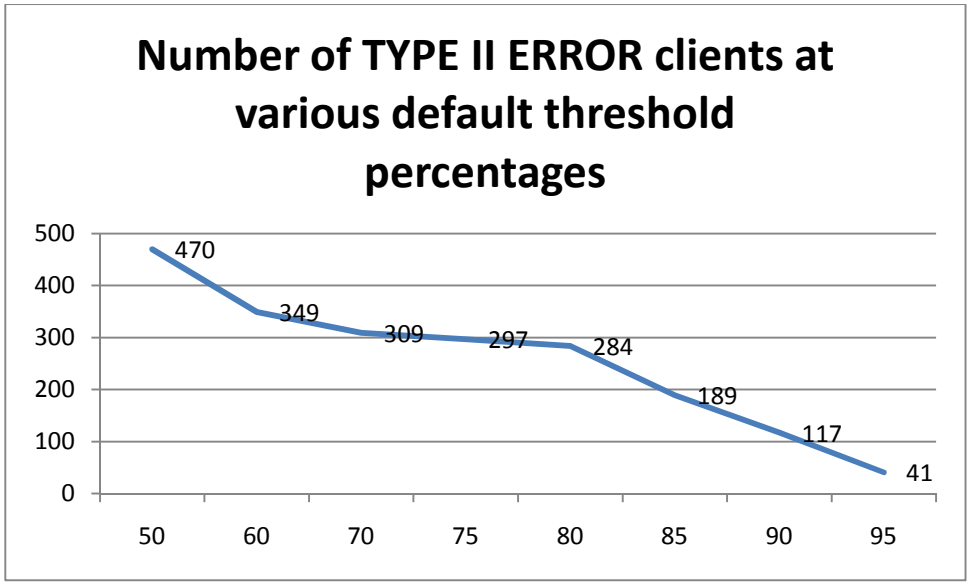


Figure 51 - Number of TYPE II ERROR clients at various default threshold percentages

Plotting all the above functions on a single graph helps to gain a better picture of the curves for the individual functions. This way, the relationship between the individual curves is easily visible.

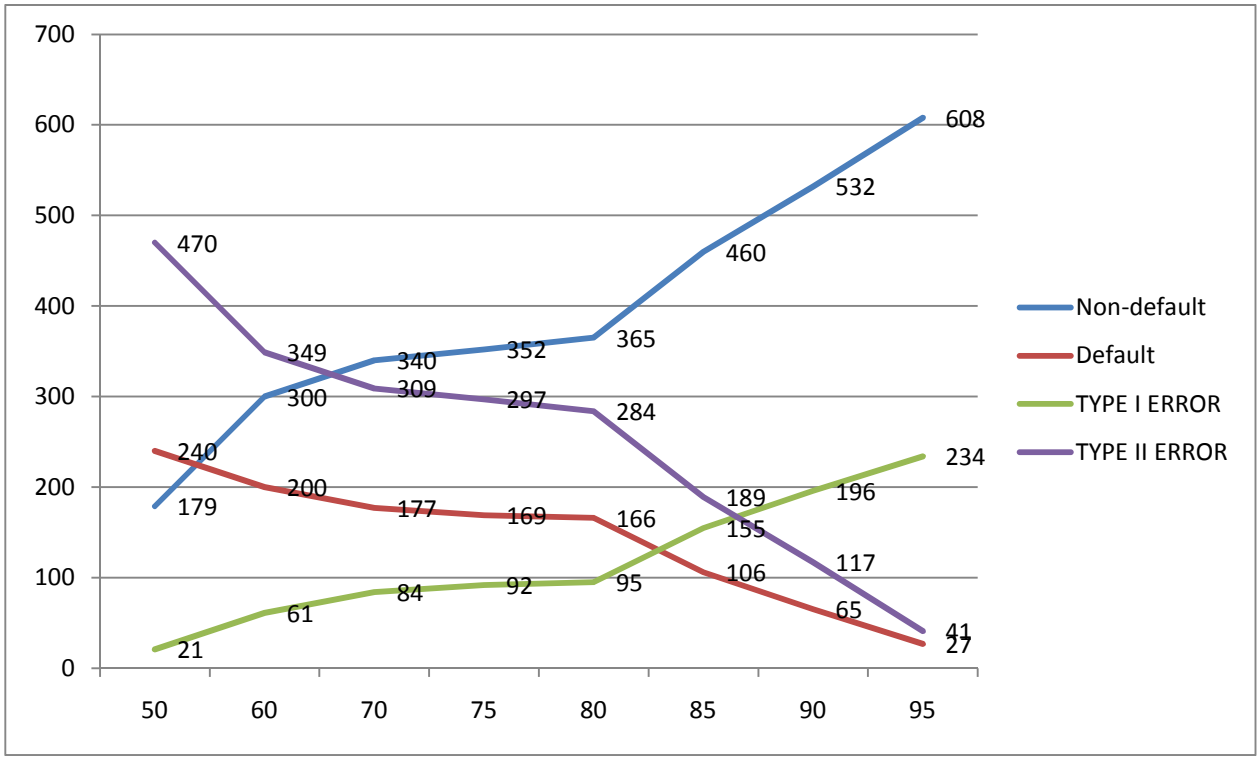


Figure 52 - Curves for the individual functions by default threshold percentage

The graph above shows the accuracy of non-default client identification increasing with the rising default threshold percentage. The TYPE II ERROR curve is almost inverse to the non-default client curve. Similarly, the default client curve is essentially inverse to the TYPE I ERROR curve. It is advisable to normalise the curve trends to be able to compare the direct relationship between them.

This is done by transforming the values into percentages (of the basis). The success rate of non-default client identification is then expressed as the number of identified non-default clients for the given default threshold divided by the total number of non-default clients. For a threshold of 50%, the success rate of non-default client detection is:

$$\frac{470}{649} = \underline{\underline{72.419\ \%}}$$

Similarly, the number of identified default clients for a given default threshold percentage needs to be divided by the total number of default clients, that is 261, in order to determine the success rate of default client identification. Division by the total number of default clients

is also used to compute the TYPE I ERROR rate, while TYPE II ERROR rate requires a division by the total number of non-default clients. The results are listed in the table below:

Default threshold	Non-default	Default	TYPE I ERROR	TYPE II ERROR	Total success rate
50 %	27.58%	91.95%	8.05%	72.42%	46.04%
60 %	46.22%	76.63%	23.37%	53.78%	54.95%
70 %	52.39%	67.82%	32.18%	47.61%	56.81%
75 %	54.24%	64.75%	35.25%	45.76%	57.25%
80 %	56.24%	63.60%	36.40%	43.76%	58.35%
85 %	70.88%	40.61%	59.39%	29.12%	62.20%
90 %	81.97%	24.90%	75.10%	18.03%	65.60%
95 %	93.68%	10.34%	89.66%	6.32%	69.78%

Figure 53 – Indicator-to-basis ratios

The total success rate means the indicator (which has been mentioned before) of the sum of correct default and non-default identifications divided by the total number of clients in the examined sample. As is apparent from the following graph, the Total Success Rate indicator has a tendency to grow as the threshold percentage increases.

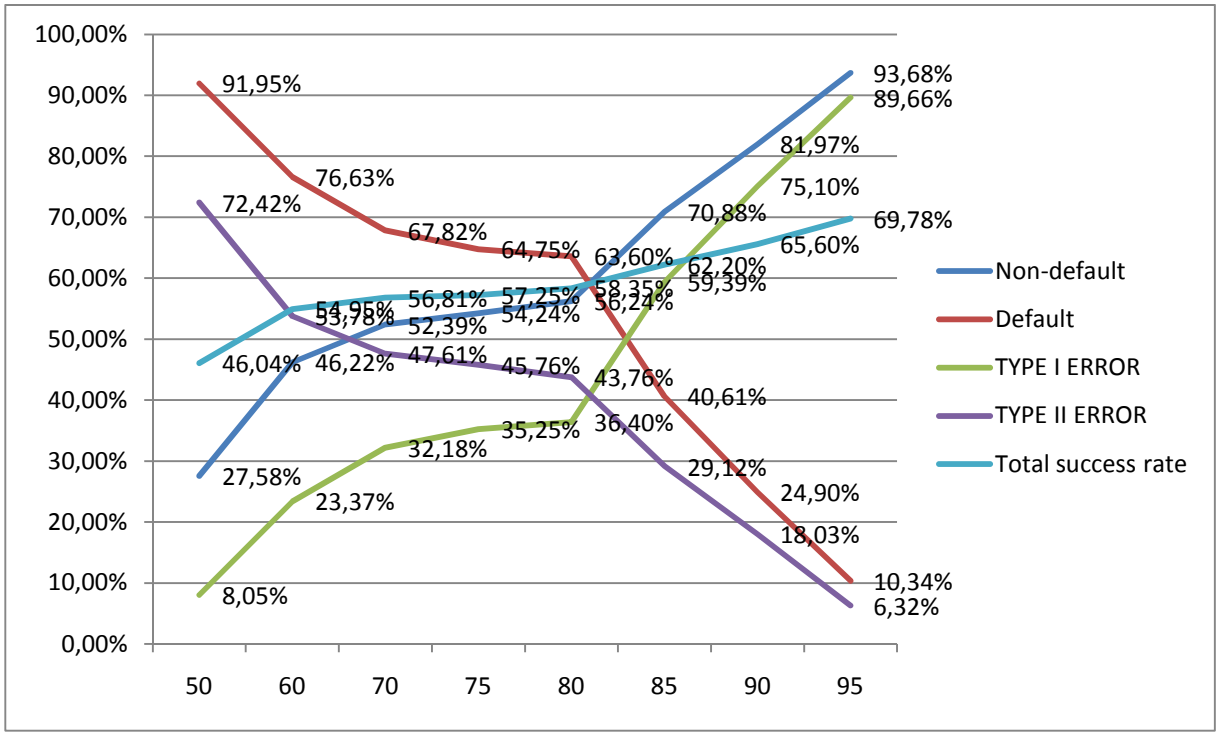


Figure 54 - Curves showing the identification success rate (in %) for the individual functions by the default threshold percentages

Although the total success-rate curve is sloping upwards, it should be noted that while the non-default identification increases with the increasing threshold, the success rate of default identification decreases considerably at the same time. Therefore, the preferences of the Bank staff need to be considered when defining the fuzzy model retransformation matrix.

If a maximum capacity to identify default clients is required – even at the cost of a higher error rate resulting in an increased number of non-default clients being identified as default –, then a lower default threshold is appropriate – in our case a threshold of 50%. A balanced evaluation could use the intersection of the non-default and default curves that is the 80 % default threshold.

The testing conducted as a part of this thesis has shown that the success rate of default detection based on the fuzzy model is higher than that of the existing Credit Risk Monitoring System but also that this higher efficiency is impaired by the higher error rate of the fuzzy model. Of course, identifying an actual non-default client as a potential default client seems better than not registering a potential threat of default at all. However, the fuzzy model is clearly lacking key variables, which would increase the accuracy of client identification and,

at the same time, eliminate errors. Another ways to improve the accuracy of the results include the use of other statistical methods, such as a neural network.

Our analysis, like the one carried out by Kočenda a Vojtek (12) has considered clients with one single loan only. There are a variety of ways of addressing the issue of clients with more than one loan. Some variables, such as Salary, Number of Persons in Joint Household, would remain the same. There may, however, be some differences with respect to other variables, such as the Amount of Instalment or the total Amount of Loan. These variables can be combined or, as the case may be, added together. However, the Length of Relationship before applying for the loan variable needs to be conceptualized in a different way and the assessment needs to be more comprehensive.

Such an assessment should take into account the fact that combinations of certain variables for different loans can:

- have a fixed character,
- have combinatorial character, or
- be void of any relationship between each other.

An analysis of different approaches to assessment has not been possible due to the lack of data for testing. The simplest approach to multiple-loan-client assessment consists in always considering the highest percentage and its linguistic variable interpretation. This means that, where a client has 3 loans, and the risk associated with the first one is 45%, the second one 56% and the third one 80%, the maximum risk value, i.e. 80%, should be selected for further analysis.

5.3. Corporate clients

Corporate clients include all companies satisfying the criterion of being a legal entity (“legal person”). The risk in companies can take the form of problematic (financial) discipline, as other factors, such as off-balance-sheet commitments need to be taken into account. A company in good financial health can easily lose a lawsuit and become insolvent.

Unfortunately, sufficient data is not available for analyzing and testing the corporate client model, this chapter will therefore only discuss an appropriate credit risk model theoretically.

5.3.1. Commercial credit scoring products

The easiest option, of course, is to determine the credit score of companies based on the services provided by rating agencies, such as Standard and Poor's, Moody's or Fitch. Unfortunately, these renowned rating agencies scarcely cover Czech and small enterprises and the room for their use is thus very limited.

Other specialized products can be used for assessing Czech companies. These include "*Firemní Lustrátor"* ("Corporate Screening") by Creditinfo Czech Republic (15). This product is a commercially available and can be used either via a web browser or as a web service enabling its integration in in-house information systems. The use of the system is subject to a fee: client use credits to pay for the display of specific data or pays a flat rate.

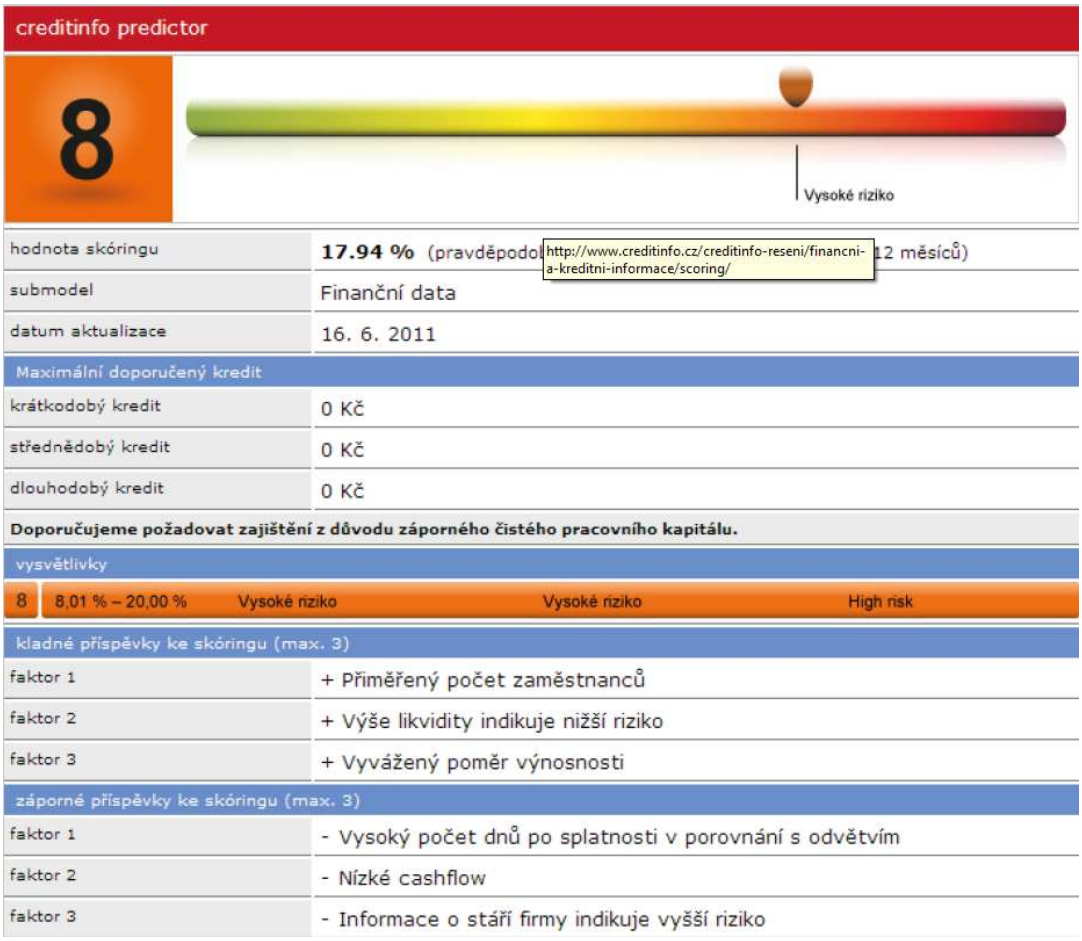


Figure 55 - Screen capture of the "Firemní Lustrátor" application, adopted from (16)

Firemní Lustrátor, however, has a couple of drawback. It covers only a limited range of companies, especially companies from the Czech and Slovak republics to begin with. Banks' clients may include foreign companies, but banks relying solely on this system essentially cannot assess them. Another disadvantage is that the principles based on which the system operates are hidden. In practice, *Firemní Lustrátor* is a total black box, with Creditinfo Czech Republic, s.r.o., responsible for its operation and development but not publishing any models or other details related to how scores are actually calculated.

Also, any bank using the system makes itself dependent on its provider, which carries risk. If the *Firemní Lustrátor* is unavailable or should Creditinfo Czech Republic, s.r.o. close down, the Bank's Credit Risk Monitoring Information System would be left without data on corporate clients. However, if the Bank wishes to outsource its credit scoring operations, *Firemní Lustrátor* is a suitable product.

5.3.2. Risk assessment models

The Bank may deem it more appropriate to have its own credit scoring model for existing loan clients – this is the typical scenario today. Vojtek notes that “all banks design their own rating models precisely because of the fact that the weighting and variables need to correspond to the time of development of the model and to the country where the data comes from“(please see Appendix 4).

Several models can be used to assess the risk in existing corporate loan clients (11). Aziz and Dar (11 pp. 6-22) give the following classification of these models:

- Statistical models
- Artificially Intelligent Expert System (AIES) models
- Theoretic models

According to Aziz and Dar “statistical models include univariate and multivariate analyses of which the latter dominates and uses multiple discriminant, linear probability, logit, and probit models.“ (11 p. 5)

Statistical models include Univariate analysis, Multiple Discriminant Analysis, Linear Probability Model, Logit model, Probit model, Cumulative Sums procedure and Partial adjustment process.

The **AIES** model chiefly draws on the principles of artificial intelligence. “Humans use their intelligence to solve problems by applying reasoning based on the knowledge possessed in their brains. Hence, knowledge plays the pivotal role in human intelligence. AI, in order to be as competitive as human intelligence or at least comparable, should benefit from similar knowledge in application of its reasoning to the problem posed. Expert systems (ES) were developed to serve this purpose for AI“ (11 p. 12).

Aziz and Dar (11) further divide AIES into:

- Recursively partitioned decision trees (Inductive learning model)
- Case-Based Reasoning (CBR) model
- Neural Networks (NN)

- Genetic Algorithms (GA)
- Rough sets models

Aziz and Dar describe **theoretical models** as “able to predict bankruptcy by looking at distress conditions present in the firms. However, another way of approaching this problem is to look at the factors that force corporations to go bankrupt“ (11 p. 18).

Aziz a Dar distinguish the following theoretical models:

- Balance Sheet Decomposition Measure (BSDM) / Entropy theory
- Gambler’s Ruin theory
- Cash management theory
- Credit risk theories
- Balance Sheet Decomposition Measure (BSDM) / Entropy theory
- Gambler’s Ruin theory
- Cash management theory
- Credit risk theories

A much discussed method is the Altman Z-score bankruptcy model. The model was developed by Edward I. Altman in 1968 and involved 66 companies divided in 2 groups. Each group included 33 companies. The bankruptcy group – Group 1 – chiefly consisted of companies included in the bankruptcy petition in the National Bankruptcy Act between 1946 and 1965. The non-bankruptcy group – Group 2 – was made up of companies with assets between USD 1 and 25 million. The average value of corporate assets in Group 2 amounting to USD 9.6 million was just a little bit higher than the average value for Group 1. Group 2 companies were still in business at the time of the analysis.

However, Altman’s bankruptcy model is obsolete. This is also noted by Vojtek: “Altman’s model was developed based on a sample of companies at a specific time and in a specific country. No bank is using it“ (please see Appendix 4).

As the retail client analysis has shown, the fuzzy model is not sufficiently accurate and has a higher error rate. Although this has in all probability been caused by the absence of a highly discriminatory variable, neural networks could still prove to be a much better option (14 pp. 235-242).

All models have their drawbacks and advantages. Aziz and Dar give the following summary of disadvantages of neural networks (NN) listed in the paper by Altman and Varetto: "…long processing time to complete the NN training stage, requirement of having a large number of tests to identify appropriate NN structure, and the problem of over fitting can considerably limit the use of NN" (11).

5.4. Summary of proposals and suggested solutions

This chapter examines the possibility of an automated assessment of risk in existing retail and corporate loan clients. Based on the data sample made available, a calculation of the information value was conducted in order to identify highly discriminatory variables.

These variables were then inserted into a transformation matrix in keeping with the principles of fuzzy logic. Next, all variables were entered in and percentages calculated for all state matrices and then interpreted, based on a retransformation matrix, by linguistic variables. The values for retransformation matrices were calculated for default thresholds between 50% and 95%.

Although the capability of the fuzzy model for detecting default clients is much higher than that of the existing Credit Risk Monitoring System, which has been the subject of analysis in the preceding chapter, the higher error rate of the proposed fuzzy model reduces the value of the solution arrived at. The higher error rate may be caused by an unidentified variable with a considerable discriminatory power which has not been included in the data provided by the Bank.

The proposal for an automated solution for assessing the risk in corporate clients has been discussed on a theoretical level only, as relevant data has not been available to duly test the model. As in the case of retail clients, it is advisable to consider the use of neural networks in future.

6. Conclusion

The main question put forward by this thesis was:

“Does the existing solution for assessing the risk in loan clients of the Bank lend itself to automation and in what ways can the existing solution be improved?“

In conclusion, the inappropriateness of the existing solution for automation can be confirmed. The existing solution requires a personal approach. The scoring method used in the solution results in a considerable discrepancy between individual clients and makes it very difficult to clearly establish the boundary between default and non-default clients. The success rate of the existing Credit Risk Monitoring Information System is only 14.7% and 19.2% at default thresholds of 50 % and 25 %, respectively.

This thesis has developed a fuzzy model based on fuzzy set theory in order to optimize the accuracy of default client identification. To achieve this, data must be analysed as a first step and then transformed into a format suitable for further use. An analysis of the information value of variables was conducted with the aim of finding the highly discriminatory variables and then a transformation matrix was created where the individual ranges were given weightings.

As only retail client data was available in sufficient quantity, the main part of the research focused on retail, while the use of neural networks was recommended for corporate clients. The final success rate of default client identification arrived at under the fuzzy model for the retail sector was considerably higher than that of the existing Credit Risk Monitoring System, achieving a success rate of up to 91.95%. Unfortunately, this increased success rate came at the expense of much higher error rate, especially for TYPE II ERROR, that is to say for the type of error which involves non-default clients being identified as default.

This does not need to be a major problem, as the bank might be better off with a false alarm than with a risk client which has not been identified. The thesis included a presentation of the results for retransformation matrix default thresholds of 50% to 95% in several steps.

The higher error rate of the fuzzy model is caused by the absence of a highly discriminatory variable in the analyzed data. Such a variable would ensure an increased accuracy of the outcomes, especially in terms of lower error rates. It should also be noted that automated processing requires quality data and is not able to take exceptions into account. It is therefore necessary to have access to appropriate and clean data, otherwise there may be unnecessary errors.

In conclusion the accuracy of the fuzzy model with respect to the detection of default clients has been confirmed to be better than results of the existing model; however further data analysis is needed to identify other highly discriminatory variables. Besides the fuzzy logic, neural network is deemed to constitute a suitable risk analysis method.

7. References

1. Hendl, Jan. *Kvalitativní výzkum : Základní metody a aplikace.* 1st edition. Praha : Portál, 2005. 408 p. ISBN 80-7367-040-2.

2. Disman, Miroslav**.** *Jak se vyrábí sociologická znalost.* Praha : Karolinum, 2008. p. 374.

3. Volksbank CZ**.** *Volksbank Czech Republic* [online]. 2011-05-03 [cit. 2011-06-12]. ANNUAL REPORT 2010. Available from WWW: <http://www.volksbank.cz/vb/public/cd/24/4e/61/20300_76192_Annual_Report_Volksbank_CZ_2010_WEB.pdf>.

4. Dostál, Petr, Sojka, Zdeněk**.** *Financial Risk Management.* Zlín : Univerzita Tomáše Bati ve Zlíně, 2008. ISBN 978-80-7318-772-9.

5. Bessis, Joël**.** *Risk Management in Banking.* 3st edition. Wiltshare : John Willey & Sons, 2010. 822 p. ISBN 978-0-470-01912-2.

6. Fabozzi, Frank J**.** *The handbook of mortgage-backed securities.* 5th edition. New York : McGraw-Hill, 2001. 950 p. ISBN 0-07-135946-X.

7. Volksbank CZ**.** *Volksbank Czech Republic* [online]. 2010 [cit. 2011-06-30]. Citizens. Available from WWW: <http://www.volksbank.cz/vb/jnp/en/obcane/index.html>.

8. Volksbank CZ**.** *Volksbank Czech Republic* [online]. 2010 [cit. 2011-06-30]. Companies. Available from WWW: <http://www.volksbank.cz/vb/jnp/en/firmy/index.html>.

9. Volksbank CZ**.** *Volksbank Czech Republic* [online]. 2010 [cit. 2011-06-30]. Entrepreneurs. Available from WWW: <http://www.volksbank.cz/vb/jnp/en/podnikatele/index.html>.

10. Credit & Management Systems**.** *Credit Research Foundation* [online]. 1999 [cit. 2011-06-18]. Rules Based Credit Scoring Methodology. Available from WWW: <http://www.crfonline.org/orc/cro/cro-15-1.html>.

11. Aziz, Adnan M. and Dar, Humayon A. *Predicting Corporate Bankruptcy: Whither do We Stand?* [online]. *Loughborough : Loughborough University,* 2004 [cit. 01 07 2011]. Available

from WWW: https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/325/3/DepartmentalPaper_AzizandDar_.pdf.

12. Kočenda, Evžen, Vojtek, Martin. *CESifo Working Paper No. 2862* [online]. 2009 [cit. 2011-06-12]. Default Predictors and Credit Scoring Models for Retail Banking. Available from WWW: <http://www.cesifo-group.de/DocDL/cesifo1_wp2862.pdf>.

13. Dinh, Thi Huyen Thanh and Kleimeier, Stefanie**.** *Credit Scoring for Vietnam's Retail Banking Market: Implementation and Implications for Transactional versus Relationship Lending*. International Review of Financial Analysis. 2007, Vol. 16, 5. pp. 471-495.

14. Dostál, Petr. *Pokročilé metody analýz a modelování v podnikatelství a veřejné správě.* Brno : AKADEMICKÉ NAKLADATELSTVÍ CERM, 2008. 344 p. ISBN 978-80-7204-605-8.

15. Creditinfo Czech Republic**.** Creditinfo Czech Republic [online]. 2011 [cit. 2011-07-16]. Firemní Lustrátor. Available from WWW: <http://www.creditinfo.cz/creditinfo-reseni/financni-a-kreditni-informace/creditinfo-firemni-lustrator/>.

16. Creditinfo Czech Republic**.** Creditinfo Czech Republic [online]. 2011 [cit. 2011-07-16]. Firemní Lustrátor. Available from WWW: < http://www.creditinfo.cz/library/Files/Product-Sheets/cz_firemni-lustrator_produktovy-list.pdf>.

17. Diaz-Serrano, Luis**.** *Income volatility and residential mortgage delinquency across the EU*. Journal of Housing Economics. September 2005, Vol. 14, 3. pp. 153-177.

18. Goodman, Allen C. and Smith, Brent C**.** *Residential mortgage default: Theory works and so does policy*. Journal of Housing Economics. December 2010, Vol. 19, 4. pp. 280-294.

19. Karakoulas, Grigoris**.** *Empirical Validation of Retail Credit-Scoring Models*. The RMA Journal. September 2004. pp. 56-60.

8. List of abbreviations and symbols

Abbreviation	Explanation
IV	Information Value
s.r.o.	A Czech private company form roughly corresponding to the British private limited company (Ltd.).
NN	Neural network
AIES	Artificially Intelligent Expert System

9. List of figures

10. List of appendices

11. Appendices

11.1. Appendix 1 – Interview with Ing. Pavel Kozák

Interview with Ing. Pavel Kozák, Project Manager of Volksbank in charge of the development of the Credit Risk Monitoring Application. The interview was held on 16 May 2011.

Jiří Kobelka: Hello, may I ask you a few questions regarding the Credit Risk Monitoring Application?

Pavel Kozák: Hello, sure.

Jiří Kobelka: Could you please briefly describe the business role of this application?

Pavel Kozák: I work as an IT Project Manager, and my perspective is, of course, limited to what I pick up from the communication with the "client", the end user, which is the Credit Risk Management Department at our bank, but I do know that the main impetus for the creation of this application was the fact that the bank had a very elaborate system of credit risk assessment for before accepting that risk – that is to say for the evaluation of a client's loan application – while after the loan was given out, that is to say during its repayment the very same information sources were either no longer taken into account or considered only to a limited extent or with a periodicity that was too long. Depending on what the individual information sources allow for, this application enables us to update data, ideally on a daily basis, and to immediately respond to the situation. Of course the application monitors only current loan clients from all client segments. Neither potential nor former loan clients are kept track of by the application.

Jiří Kobelka: And in case a client falls within the category of those with a loan default risk? What steps can be taken by the bank then?

Pavel Kozák: It depends on the product, the type of client, specific contractual arrangements as well as on the amount of the client's total liabilities and the quality of the security. There is, of course, a difference between a consumer loan and large investment loans or development

projects. The rationale of the Credit Risk Monitoring application is more that of providing early warning and support for the decision making process. The final decision will not be based only on this. We try to respond to our clients´ needs as much as we can, but the bank must have the capability to monitor risk and respond to it adequately.

Jiří Kobelka: On what data are the application's assessments based?

Pavel Kozák: Generally speaking, on data either from the in-house systems of the Bank, such as the main banking system, or from external sources, both sources available to the public or non-public interbank registers. The individual sources are then divided into positive or neutral ones, which are shown in the application but are not further assessed, and into negative ones which are subject to assessment. There are events of a constant nature for which there is no numerical range; these include changes in the company's record in the Commercial Register. Such changes result in a fixed number of "penalty" points. Then there are negative events that are numerical for which individual point ranges are defined. These include the amount of liabilities overdue, how long these have been overdue, or the extent of non-compliance with the agreed contractual conditions and the like.

Jiří Kobelka: How exactly does scoring work and how are points defined?

Pavel Kozák: In principle, scoring is very simple. As I have said, scoring is either fixed, which means a constant number of points is awarded for a negative event, or based on the scope of the event. Scoring usually takes place only once as of the date when the negative event in question is fed in the application. The application enables older events to be filtered by the application. Employees of the Credit Risk Management Department issue opinions on the individual events. The number of points and point ranges applicable to negative events are defined by the specialists of this department.

Jiří Kobelka: Does that mean that the Credit Risk Monitoring application uses no specific scoring model?

Pavel Kozák: It does not use a specific model just yet. We have used the application for our in-house purposes for a relatively short period of about two years, so we are trying to analyze

its benefits and considering potential future improvements. But implementing a scoring model would be perfectly adequate in this case.

Jiří Kobelka: Does the Credit Risk Assessment methodology used by the application differentiate between individual client segments, such as retail, SMEs, municipalities?

Pavel Kozák: Not directly, scoring and coefficients are always the same, except that certain data is not available for some types of clients and is therefore not used. To give you an example: there are differences between companies and individuals in terms of what data they disclose. When working with the application the specialists of course take the client's segment into account.

Jiří Kobelka: When interpreting the results of the preliminary analysis provided by the application, do you divide them into sample groups, for instance between ten templates to which every client can be assigned?

Pavel Kozák: We have not looked into such an interpretation yet but we have considered a similar solution which would help assign clients automatically according to pre-defined templates. Currently, we have only drawn up a couple of standard scenarios, which look at the correlation of negative events in more sources. Based on that correlation the real situation of the client is easier to estimate. This allows us to detect cases which may not necessarily stand out from the rest in terms of the total number of "penalty points" collected, but whose nature makes it obvious that the firm's health has deteriorated.

Jiří Kobelka: What loan products are monitored by the application? Mortgages, current account overdrafts, revolving credits? And what is the definition of a loan client?

Pavel Kozák: We monitor all current loan clients regardless of the product type. A loan client is any individual or entity for which there are active provisions recorded in the balance sheet or off-balance sheet of the bank, both before or after maturity.

Jiří Kobelka: When a Bank monitors the risk associated with its current loan clients, what can it, as a matter of course, actually do under the existing agreements if it finds out that the risk associated with the client is significantly higher than it originally was at the time when the loan was granted?

Pavel Kozák: The goal is not to punish the client in any way, the goal is simply to protect the Bank’s claims effectively and early. The application does not change the procedures with respect to the client, it only makes them faster. More precisely, it improves the accuracy of their targeting. All possible options are naturally included in the contractual conditions.

11.2. Appendix 2 – Variable definitions by Kočenda and Vojtek

The following list of variable definitions is adopted from Kočenda and Vojtek papers (12).

Socio-demographic variables

Sex	Sex of the client, categorized variable
Marital status	Status of the client, single/married, categorized variable
Date of Birth	Date of birth of client
Sector of employment	The sector in which the client is employed, categorized variable
Type of employment	Type of client's employment, categorized variable
Education	The highest attained education of client, categorized variable
Number of employments	The total number of employments in the last 3 years
Employment position	The position of client in employment, categorized variable
Years of employment	The number of years in the current employment
Credit ratio 1	Ratio of Expenditures/Income of client
Credit ratio 2	Ratio of (Income-Expenditure)/Living Wage of client
Region	Post Code of region of client's address

Bank-client relationship variables

Type of product	Type of product/loan
Number of co-signers	The Number of co-signers for the current loan
Purpose of loan	The declared purpose of loan, categorized variable
Loan Assurance	The type of credit risk mitigation, categorized variable

Points	The characteristics of client's behaviour at the current account
Own resources	Declared own resources, in percentage of total amount needed
Amount of loan	The total amount of loan granted
Date of account opening	The year when client opened an account in the bank
Date of loan	The year in which the loan was granted
Length of the Relationship	The length of client/bank relationship at the time of loan application

11.3. Appendix 3 – Information values of variables by Kočenda and Vojtek

The following list of information values of variables is adopted from Kočenda and Vojtek papers (12).

Own resources	1.462601
Date of account opening	0.631346
Length of the relationship	0.601787
Points	0.502122
Education	0.359725
Purpose of loan	0.279959
Years of employment	0.136041
Sector of employment	0.188681
Credit ratio 1	0.175810
Number of co-signers	0.131135
Amount of loan	0.123972
Marital status	0.112809
Region	0.093896
Employment position	0.063872
Type of employment	0.055486
Credit ratio 2	0.052161
Date of Birth	0.047698
Sex	0.039528
Loan assurance	0.036422
Type of product	0.022380
Number of employments	0.021004

11.4. Appendix 4 – Interview with Mgr. Martin Vojtek, PhD.

Interview with Mgr. Martin Vojtek, PhD., co-author of Default Predictors and Credit Scoring Models for Retail Banking (12)**. Mgr. Martin Vojtek, PhD., works as the Head of Quantitative Validation Team at the Financial Market Supervision Department of the Czech National Bank.**

Jiří Kobelka: The paper which you co-authored describes the data sample you used as retail clients of an undisclosed bank with a single loan. Have these clients been checked in terms of whether they have another loan at another bank institution in parallel, for instance by means of the BRCI (Bank Register of Client Information, "*Bankovní registr klientských informací*")?

Mgr. Martin Vojtek, PhD.: The paper is unfortunately of a slightly older date, and the sample is therefore older as well (the sample is from 2000–2006, if I remember that correctly). At least at the beginning of that period the BRCI was not fully operational. Also, I am not absolutely sure what the setup of the bank procedures looked like, and so I cannot tell you whether such verification was carried out at the bank. When we received the data, it had already been rendered anonymous, and so we could not verify that ourselves (via BRCI or another institution).

Jiří Kobelka: Based on what did you select the socio-demographic and the bank-client relationship variables? If it had been possible, would you have used other variables as well?

Mgr. Martin Vojtek, PhD.: The long list of variables we used was essentially everything the bank had at its disposal at that time: we of course used the socio-demographic variables from the applications (the bank probably did not collect more data than that); moreover the bank was not really prepared for a reasonable collection of behavioural characteristics (account turnovers, etc.), the main reason being that it was not entirely a standard retail bank. I would have certainly used more behavioural variables if it had been possible.

Jiří Kobelka: How can the results of your work be applied in a situation where the client has multiple loans? If the socio-demographic variables remain the same, then by, say, means of a simple selection of the lowest value of each single variable with respect to the resulting values of every loan?

Mgr. Martin Vojtek, PhD.: We probably need to differentiate between multiple loans at one bank and at several banks. In the first case (if the bank has complete information) more loans can be taken into account by means of assessing the client's creditworthiness (if all of them are repaid in instalments). It also makes sense to develop a specific model for each loan type (mortgages, credit cards, consumer loans, etc.), and evaluate the client based on these partial models, where appropriate. In the other case the only option is probably to rely on the information from the BRCI and similar sources.

Jiří Kobelka: Do you think that Altman's bankruptcy model for corporate clients is still up to date and suitable for assessing risk in existing loan clients?

Mgr. Martin Vojtek, PhD.: Altman's model was developed based on a sample of companies at a specific time and in a specific country. No bank is using it: all banks design their own rating models precisely because of the fact that the weighting and variables need to correspond to the time of development of the model and to the country where the data comes from. In reality, however, there are regular structural changes and there is therefore no reason why a model calibrated 40 years ago should work today.

Jiří Kobelka: Do you think that it makes sense to look for patterns in the products of existing clients?

Mgr. Martin Vojtek, PhD.: It certainly does make sense, the behaviour of clients, for instance as far as mortgages are concerned, differs radically from their behaviour with respect to credit cards (to make a long story short: clients tend to give out their last penny for their home, because they may just as well lose it, while for credit cards they have found out that they can get away with relatively little damage when they do not pay). My experience is that it is very common to develop retail models based on specific products precisely because of these differences in client behaviour. A more refined distinction does, however, seem useless, I do not think that the behavioural patterns are that much different for, say, consumer loans for cars and consumer loans for furniture.

11.5. Appendix 5 - Variables commonly used in retail credit scoring models

Table 1
Variables commonly used in retail credit scoring models
When available, a rank indicating the importance of the variable in the credit scoring model is given in brackets. (NI) indicates that a variable was considered but finally not included. The first colum is based on Crook et al. (1992).

commonly used variables in industrialized countries	variables included in the credit scoring model		
	Crook et al. (1992) - UK	Schreiner (1999) - Bolivia	Vigano (1993) - Burkina Faso
Postcode	Postcode (1)	Date of disbursement	Customer's personal characteristics
Age	Employment status (2)	Amount disbursed	(age, sex, religion, marital status, education,
Number of children	Years at bank (3)	Type of guarantee	employment sector and place, etc)
Number of other dependants	Current account (4)	Branch	Data on the enterprise
Whether an applicants has a home phone	Spouse's income (5)	Loan officer	(type, professional skills, number of employees,
Spouse's income	Residential status (6)	Gender of the borrower	productivity, profitability, etc)
Employment status	Phone (7)	Sector of the firm	Profitability
Employment category	Years at present employment (8)	Number of spells or arrears	(main and secondary revenue, revenue stability)
Years at present employment	Deposit account (9)	Length of the longest spell of arrears	Amount and composition of assets
Income	Value of home (10)		(total assets including money and deposits)
Residential status	Outgoings (11)		Financial situation
Years at present address	Number of children (12)		(initial and current amount of loans received,
Estimated value of home	Applicant's income (NI)		defaults, loans granted)
Mortgage balance outstanding	Mortgage balance outstanding (NI)		Investment plans
Years at bank	Charge card (NI)		(presence of investment plan, other sources of
Whether a current account is held			finance)
Whether a deposit account is held			Customer's relationship with bank
Whether a loan account is held			(past loans with bank, savings account with
Whether a check guarantee card is held			bank, etc)
Whether a major credit card is held			Bank's control of credit risk
Whether a charge card is held			(loan distination, disbursement form, method
Whether a store card is held			of repayment, loan amount and maturity, collateral,
Whether a building society card is held			contractual conditions on interest rate, etc)
Value of outgoings			

Figure 56 - Variables commonly used in retail credit scoring models, source: (13)

11.6. Appendix 6 – Formulas for the calculation of the individual variables in the transformation matrix

Number of Current Account Years

=IF(E270<2;9;IF(E270=2;45;IF(E270=3;146;IF(E270=4;195;IF(E270=5;103;IF(E270=6;103;IF(E270>6;115;9)))))))

Salary

=IF(AZ270="NULL";112;IF(AZ270=0;112;IF(AZ270="NULL";112;IF(AND(AZ270>0;AZ270<20001);22;IF(AZ270>20000;8;112)))))

Regular Income

=IF(CI270="NULL";109;IF(CI270=" NULL";109;1))

Housing Type

=IF(AN270="NULL";94;IF(AN270="NULL";94;IF(AN270=1;16;IF(AN270=3;15;IF(AN270=2;12;IF(AN270=4;2;IF(AN270=5;1;IF(NEBO(AN270=6;AN270=7);3;94)))))))

Education

=IF(AO270="NULL";90;IF(AO270="NULL";90;IF(AO270="NULL";90;IF(NEBO(AO270=1;AO270=4);3;IF(AO270=2;17;IF(AO270=3;18;IF(AO270=5;7;IF(AO270=6;1;90)))))))

Amount of Loan

=IF(O270<100000;37;IF(AND(O270>99999;O270<300000);60;IF(AND(O270>299999;O270<800000);37;IF(AND(O270>799999;O270<1200000);27;IF(AND(O270>1799999;O270<3000000);30;IF(AND(O270>1199999;O270<1800000);40;IF(O270>2999999;44;

Current Account

=IF(BZ270="NULL";44;IF(BZ270=" NULL";44;IF(BZ270=0;44;IF(BZ270=1;59;44))))

Amount of Loan Instalment

=IF(AP270="NULL";57;IF(AP270=0;57;IF(AP270="NULL";57;IF(AND(AP270>0;AP270<5001);20;IF(AND(AP270>5000;AP270<9001);8;IF(AND(AP270>9000;AP270<20001);9;IF(AP270>20000;5;57)))))))

Employee

=IF(AR270="NULL";37;IF(AR270=" NULL";37;IF(AR270=0;44;IF(AR270=1;50;37))))

Length of the Relationship

=IF(Z270<=0;20;IF(AND(Z270>0;Z270<=1);49;IF(AND(Z270>1;Z270<=3);10;IF(AND(Z270>3;Z270<=5);6;IF(AND(Z270>5;Z270<=10);9;IF(Z270>10;4;20))))))

Number of Persons in Joint Household

=IF(AM270="NULL";46;IF(AM270="NULL";46;IF(NEBO(AM270=3;AM270=5);2;IF(AM270=1;14;IF(AM270=2;6;IF(AM270=4;3;46))))))

Own Resources

=IF(CH270="NULL";41;IF(CH270=0;41;IF(CH270="NULL";41;IF(AND(CH270>0;CH270<50000);2;IF(CH270>499999;1;41)))))

Employment Contract

=IF(AX270="NULL";38;IF(AX270=" NULL";38;IF(AX270=1;0;IF(AX270=2;2;38))))