VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA PODNIKATELSKÁ
ÚSTAV EKONOMIKY

FACULTY OF BUSINESS AND MANAGEMENT
INSTITUTE OF ECONOMICS

THE APPLICATION OF FUZZY LOGIC FOR RISK EVALUATION OF CORPORATE CLIENTS

APLIKACE FUZZY LOGIKY PRO HODNOCENÍ RIZIKOVOSTI FIREMNÍCH KILENTŮ

DIPLOMOVÁ PRÁCE
MASTER‘S THESIS

AUTOR PRÁCE Bc. LUBOŠ MLČOCH
AUTHOR

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

BRNO 2012

Brno University of Technology
Faculty of Business and Management

Academic year: 2011/2012
Institute of Economics

MASTER'S THESIS ASSIGNMENT

Mlčoch Luboš, 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:

The Application of Fuzzy Logic for Risk Evaluation of Corporate Clients

In the Czech language:

Aplikace fuzzy logiky pro hodnocení rizikovosti firemních klientů

Instruction:

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

List of literature:

DOSTÁL, P. Advanced Decision Making in Business and Public Services. Brno : CERM, 2011. 168 s., ISBN 978-80-7204-747-5.
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.
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.

Deadline for submission master's thesis is given by the Schedule of the Academic year 2011/2012.

L.S.

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

Brno, 30.06.2012

Abstract

This thesis focuses on application of fundamental principles of fuzzy logic in the process of risk evaluation of corporate clients. Based on data from a real bank, the author designs two distinct models that serve as default detection tools for corporate clients. Both models and their performance are thoroughly evaluated.

Abstrakt

Diplomová práce se soustředí na aplikaci principů fuzzy logiky v procesu hodnocení rizikovosti firemních klientů. Na základě reálných dat poskytnutých bankou autor navrhnul dva různé modely, které slouží jako nástroje pro detekci úpadkových firemních klientů. Oba modely a jejich výkonnost jsou řádně otestovány.

Keywords

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

Klíčová slova

Fuzzy logika, Řízení rizik, Umělá inteligence, Detekce úpadkového klienta

Bibliographic citation

MLČOCH, L. Application of Fuzzy Logic for Risk Evaluation of Corporate Clients. Brno: Brno University of Technology, Faculty of Business and Management, 2012. 84p.

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, 27 August 2012

…………………………

Bc. Luboš Mlčoch

Acknowledgements

I’d like to thank my supervisor – Professor Ing. Petr Dostál, CSc. for his invaluable advice and support throughout the whole process of creating this Master’s Thesis.

I also want to thank Mgr. Tomáš Magyar for his collaboration and willingness to help.

Last but not least, I would like to thank my parents for everything they have done for me over the years.

Contents

1 Introduction

Banks play a very important role in the society and economics of today. There are a number of noteworthy banks in the Czech Republic, and one of these banks has kindly provided me with data sample that contains data from real clients. This bank (it will be referred to as the bank hereinafter) wishes to stay anonymous and I will honour that request.

The data sample is to be analysed, and eventually used as a basis for creation of tools that could numerically express the risk factor of a SME (small and medium sized enterprises) client. SMEs are considered the backbone of the economy of many countries all over the world (8).

One of the primary functions of a bank is to offer various loan products. However, before a bank commits to a loan, it needs to evaluate the (potential) client and decide whether it is viable (i.e. the client will be able to repay both the loan and the interest) to provide the loan or not. A bank needs some kind of risk evaluation system that helps with the decision. The discipline that is concerned with this issue is called credit risk management. A bank is generally monitoring a number of variables, both financial and non-financial, of every already active client as well. These clients may wish to increase the loan, renew, or take advantage of a different product.

The main aim of this thesis is to apply the fundamental principles of fuzzy logic and design a system that would be able to predict whether a particular client possess a risk (of defaulting against the creditor) or not.

However, this thesis focuses solely on the analysis and risk evaluation of the non-financial variables. According to Lehmann (7), very little research is available on the role of non-financial (soft facts) data in internal credit rating systems.

As mentioned already, this thesis takes advantage of real-world data sample that has been kindly provided by the bank. However, the sample was modified in a way that makes any client identification impossible (i.e. all personal information has been removed), but the actual numerical ratings of each client have been preserved so that there is no negative impact on the solution proposed in this thesis.

In return for the bank’s goodness of providing a real data sample, this thesis should come up with tools or findings that may be of actual use for the bank.

2 Executive summary

The main aim of this thesis is to apply fundamental principles of fuzzy logic in the process of risk evaluation of SME (corporate) clients by means of creating a fuzzy model, and at the same time confirm the appropriateness of application of fuzzy logic in this process.

The results of this thesis are two distinct fuzzy models that are capable of default client detection, i.e. assessing the risk level. Each model is created using a different approach. Both approaches stem from the analysis of the data sample provided by the bank.

The first model, called the Excel model, draws on the fundamental principles of fuzzy logic. The variables suitable for use in the transformation matrix of this model have been chosen based on analysis of the predictive power of the non-financial categories, which was determined using the information value calculation.

It is concluded, that the Excel model is suitable for daily use due to several reasons. The Excel model is easy to modify and maintain, and it can be used on any computer with standard office pack installed. The Excel model also handles large data input easily.

The default detection success rate of the model differs based on the selected default threshold, i.e. retransformation matrix, and peaks at 88%, however this figure is impaired by a very high error rate. For that reason, the thesis offers a collection of retransformation matrices and their results.

Overall, the performance of the Excel model is deemed satisfactory considering the fact that the data sample provided by the bank is relatively small, and no financial data of any corporate clients has been made available to the author.

The second model, called the MATLAB model, offers one big advantage over the Excel model – it is capable of processing float numbers, which naturally leads to more accurate and reliable results. Even when using integer variables on input, the MATLAB model is more accurate than the Excel model, albeit based on a much smaller sample size. The default detection success rate of the model peaks at 90%, with relatively low error rate of respectively 10% and 25.5%. Disadvantages of the MATLAB model include the fact that it requires the MATLAB suite to run and can't be modified as easily.

As such, the appropriateness of application of fuzzy logic in the banking sector, SME sector to be exact, has been confirmed in the thesis.

3 Theoretical basis of the work

This Master's thesis follows the qualitative research strategy as outlined by Hendl (9).

The main aim of this thesis is to design a tool capable of assessing the risk level of a corporate client. The tools (models) will be based on principles of fuzzy logic.

However, it is imperative to build the theoretical base first, as the models will stand on this base.

This chapter sums up the necessary theory base needed for creating a fuzzy logic-based model, which is the main goal of this thesis.

3.1 Fuzzy logic

Fuzzy logic was first introduced by Professor Lotfi Zadeh of California University in 1965. As opposed to classical logic, which interpolates the input into a crisp set, fuzzy logic has an ability to classify elements into a continuous set using the concept of degree of membership (1).

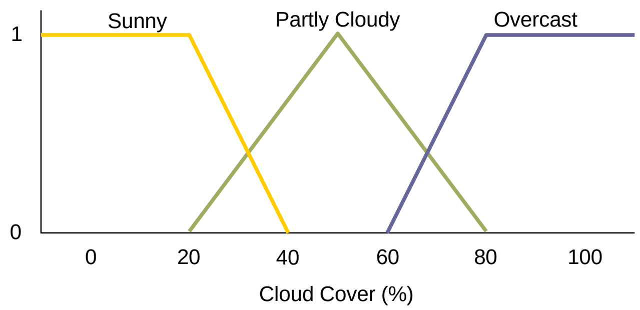

Figure 1 – Demonstrating membership levels

As shown in the Figure 1 above, (e.g. 25% Cloud Cover belongs to the sunny group by 0.8, and to the partly cloudy group by 0.2), unlike classical logic, membership function of fuzzy logic not only gives two states – true or false (0 or 1), but it can also give values between 0 and 1. Therefore, it is obvious that fuzzy logic greatly differs from classical logic in basic principles and potential use.

Fuzzy logic finds its use in a wide array of fields ranging from the car industry and electrical household appliances to corporate management and decision making enhancement. Additionally, Dostál (2) states that the fuzzy method (using the fuzzy sets) can be used in the area of risk management.

Fuzzy processing

According to Dostál and Sojka (3), fuzzy processing consists of three fundamental steps, namely fuzzification, fuzzy inference and defuzzification.

Figure 2 – The three fundamental steps of fuzzy processing

Fuzzification is the conversion of real variables into language variables (2).

Dostál (2 p.11) defines language variables as follows: “The definition of language variables draws on linguistic variables, for instance the variable “Risk” can have the following variables: zero, very low, low, medium, high, very high. Usually three to seven attributes are used for a variable”.

Dostál and Sojka (3 p.63) describe fuzzy inference in the following way: “System behaviour by means of the rules of the type IF THEN. The conditional clauses create these algorithms, which evaluates the input variables”.

Defuzzification is defined as the conversion of numerical values to linguistic ones (3). Dostál and Sojka also state that “The linguistic values can be, e.g. for variable risk very low, low, medium, high, very high risk” (3 p.63).

3.2 Selecting the appropriate credit risk model

Even though there are a number of approaches for assessing credit risk corporate clients (6), the methodology published by Bessis (5) is used for the purposes of this thesis.

Bessis makes a distinction between behavioural scoring models and origination scoring models (5).

According to Bessis (5 p.546), behavioural scoring model is “an attempt to model the behaviour of existing clients, when there is no new event that would change the debt level, given historical data of account and loan behaviour. 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”.

Bessis (5) also claims that the origination model is better suited for assessment of new clients. This thesis will use the behavioural scoring model as it is more suitable than the origination model for the purposes of this thesis.

3.3 Default clients

Bessis states that Basel 2 banking rules define default event as: “non-payment of debt obligations for 90 days“. (5 p.235) Bessis recommends that a default analysis is carried out on an annual basis (5). By contrast, clients who do not fail to meet payment of debt obligations in 90 day long window can be considered non-default clients.

4 Problem analysis and current situation

This chapter focuses on analysing the current situation in the bank and establishes the main goal of this thesis. The relevant problems of Risk Management are discussed. The chapter concludes with a summary of the current situation and the problem.

4.1 Current situation in the bank

As mentioned already, the bank requested that its name remain classified, and this thesis respects that request. However, it is still possible to discuss the current situation in general terms so that the bank's name is not disclosed, but the problems are clearly outlined.

Every bank must deal with the credit risk. Dostál and Sojka (3) state that credit risk refers to risk that a debtor will default on any type of debt by failing to make payments to a creditor.

For the purposes of this thesis, the bank is the creditor, while the SME client is the debtor.

Every creditor needs to assess the level of credit risk in case of every individual SME client, and react to it in an appropriate way (e.g. increasing the interest in order to cover the credit risk). Therefore, every bank also operates some kind of system that assesses the risk level of every individual client (7).

Lehmann (7 p.3) states that "Bank internal credit risk evaluation systems go by a number of names, such as expert systems, credit scoring or credit risk rating. These methods differ in the degree of subjectivity contained in the decision making process, i.e. the degree to which the system can be adapted to the individual case and, thus, the degree to which the decision is influenced by the credit analyst's or relationship manager's personal opinion.".

4.2 Problem analysis

Based on the provided data sample, it is obvious that the bank evaluates its SME clients based on sixteen non-financial categories (variables). The number of evaluated financial variables is unknown, but that is not relevant to this thesis anyway.

This shows a need for a tool capable of assigning a numerical value to the risk level of a client for the non-financial part of the overall risk evaluation. At the same time, the appropriateness of application of fuzzy logic in corporate banking will be tested.

4.3 Summary

Tools capable of assessing the risk level of a SME client need to be designed. These tools will be created on the fundamental principles of fuzzy logic. Based on the evaluation of performance of both tools, the appropriateness of application of fuzzy logic in corporate (SME) banking will be tested.

The desired features of the two tools include accuracy, easy maintenance, ease of use, and compatibility with various operating systems.

5 Proposals and contribution of suggested solutions

This chapter thoroughly analyses the data sample and determines the predictive power of each variable by calculating the information value of all sixteen non-financial categories (variables), as well as the share of default clients in each rating within each category.

Based on the analysis, two solutions for detecting default clients are designed. Both models are created using two different, yet similar, approaches and estimation techniques.

The first model for default client detection is built on the establishment of base of variables with high predictive power by the means of calculating the information value of all non-financial variables. The final version of the model is created in Microsoft Excel.

The second model is created using the Fuzzy Logic Toolbox of the MATLAB application suite. The backbone of this model is based on a set of fuzzy rules, with the parameters set according to analysis of the data sample.

Both models and their performance are thoroughly evaluated.

5.1 Analysis of the data sample

The first step in the process of creating a fuzzy-logic based risk model is the analysis of the data sample, which was kindly provided by the bank.

The data sample includes over 4200 instances (245 default clients and 3982 non-default clients) of various SME (small and medium sized enterprises) clients. Ideally, the share of default clients would be closer to that of non-default clients, but unfortunately, such a sample was not made available to the author. Even then, the sample does contain sufficiently high number of default instances.

SME clients from the data sample are evaluated based on 16 non-financial categories and a number of financial categories (variables).

However, as mentioned already, this work focuses solely on the non-financial categories. That does not make the models any less valuable though. Recent literature (7) concludes that financial variables are not sufficient to predict SME default and that including non-financial variables improves the prediction power of a solution.

The non-financial categories are thoroughly described and discussed later in this chapter.

In order to lessen the impact of the fact that financial categories are not available for analysis, the bank was asked to provide data sample containing entries from SME clients with similar financial figures.

In order to construct a fuzzy risk model capable of detecting default clients, it is necessary to find patterns that clearly differentiate default clients from non-default clients. The following sub-chapter is concerned with calculation of the information value of all the non-financial variables.

5.2 Information value of the non-financial variables

As mentioned already, the data sample consists of 16 non-financial variables.

However, it is important to note that not all variables have the same or even comparable impact on default client detection. The importance of each variable for the purposes of detecting default clients greatly differs.

The information value (*Ival*) is one of the estimation techniques that can help determine the importance of each variable. The information value expresses the predictive power of a variable (4).

In their research paper, Kočenda and Vojtek (4) put forward the following equation for calculation of the information value (*Ival*) of a variable:

$$Ival\ i = \ln(Odds\ i) \cdot \left(\frac{Defaluted\ i}{Defaluted} - \frac{Good\ i}{Good}\right)$$

Defaulted i represents clients identified as default based on the given variable (i), and *Defaulted* is the sum of all default clients in the entire data set.

Similarly, *Good i* represents non-default clients identified as non-default based on the given variable (i), and *Good* is the sum of all non-default clients in the entire data set.

Odds i is a value that expresses the discrimination ability of a variable for the whole group. *Odds i* is calculated in the following way:

$$Odds\ i = \frac{Defaluted\ i}{Defaluted} \cdot \frac{Good}{Good\ i}$$

The variables in this formula are defined in the same way as the variables in the previous equation.

The whole equation, then, looks like this:

$$Ival\ i = \ln\left(\frac{Defaluted\ i}{Defaluted} \cdot \frac{Good}{Good\ i}\right) \cdot (\frac{Defaluted\ i}{Defaluted} - \frac{Good\ i}{Good})$$

In their research paper, Kočenda and Vojtek (4) specify that in the banking sector, category (variable) needs to have the total information value equal or greater than 0.2 to be considered a variable with high predictive power. This thesis uses the same information value threshold of 0.2 as the basis for creating the final version of the risk model.

Thanks to the equation mentioned above, it is possible to calculate the information value for all sixteen non-financial variables.

5.2.1 Recent Development of the Financial Situation

The first non-financial variable is called **Recent Development of the Financial Situation**. The recent development of the financial situation can be either negative or positive, depending on various external factors (economic policies of the state, EU legislation, etc.), and also factors such as competitors, assets and liabilities etc.

This category is rated on a scale that ranges from *A* to *E* (best to worst). A Not Available (*N/A*) rating is used in case that the necessary data is not available. In the actual data sample, which was provided by the bank, the ratings are represented by numerical values.

To clarify the process of calculating the information value, this variable will be used to demonstrate how the information value is calculated.

Based on the data sample, the following table was created:

Rating	Total Clients	Default	Non-Default
N/A	23	2	21
A	278	7	271
B	2808	105	2703
C	666	55	611
D	372	46	326
E	80	30	50
Total:	**4227**	**245**	**3982**

Table 1 – Client spread across all ratings within the category

Using the aforementioned equation, it is possible to calculate the information value for each of the six available ratings in the variable.

$$Ival\ i = \ln\left(\frac{Defaluted\ i}{Defaluted} \cdot \frac{Good}{Good\ i}\right) \cdot \left(\frac{Defaluted\ i}{Defaluted} - \frac{Good\ i}{Good}\right)$$

Information value for rating *N/A*:

$$Ival\ N/A = \ln\left(\frac{2}{245} \cdot \frac{3982}{21}\right) \cdot \left(\frac{2}{245} - \frac{21}{3982}\right)$$

$$Ival\ N/A = 0.0013$$

Information value for rating *A*:

$$Ival\ A = \ln\left(\frac{7}{245} \cdot \frac{3982}{271}\right) \cdot \left(\frac{7}{245} - \frac{271}{3982}\right)$$

$$Ival\ A = 0.0343$$

The calculation is done in the same way for all the remaining ratings in this variable. The total information value for the whole variable is calculated by adding up numerical values of information value of each rating.

Rating	Total Clients	Default	Non-Default	Information value
N/A	23	2	21	0.0013
A	278	7	271	0.0343
B	2808	105	2703	0.1151
C	666	55	611	0.0270
D	372	46	326	0.0879
E	80	30	50	0.2503
Total:	**4227**	**245**	**3982**	**0.5158**

Table 2 – Overview of information value of the first non-financial variable

As we can see, the total Information value of this variable is very high. This is mainly thanks to the rating *E*, which boasts the highest predictive power out of all ratings within this variable. The *E* rating would pass the pre-set predictive power threshold of 0.2 on its own.

Overall, this variable's information value easily passes the pre-set predictive power threshold, which makes it suitable for use in the eventual risk models.

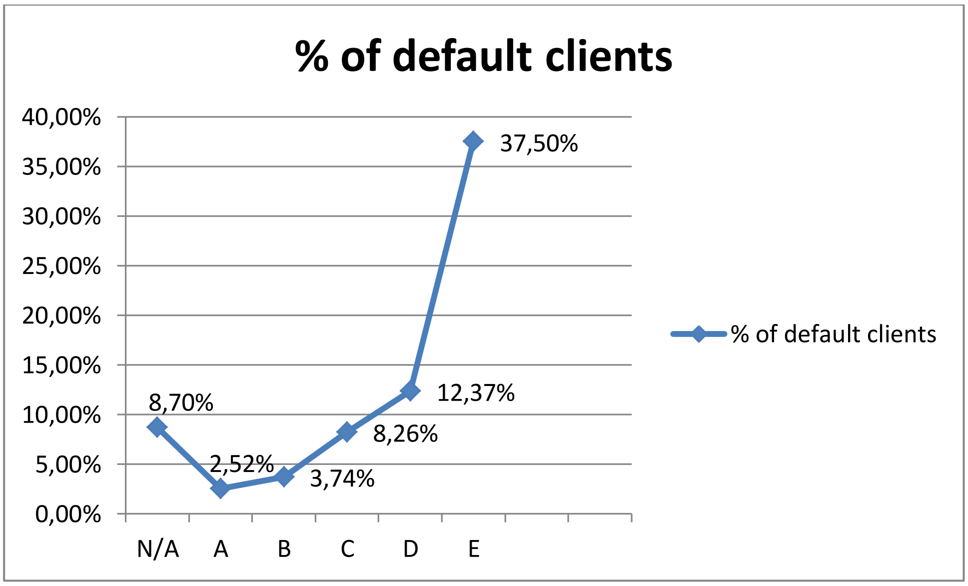

Figure 3 - Share of default clients in each rating within the first variable

Figure 3 depicts the share of default clients in each rating of this variable. As expected, the *E* rating boasts the greatest share of default clients at 37.5%.

The development outlined in Figure 3 is to be expected – the share of default clients increases as the ratings get worse.

The only exception is rating *N/A*, which has higher share of default clients than ratings *A*, *B*, and *C*. However, this is caused mainly by the fact that only very few clients actually received the *N/A* rating, and it does not refute the observed trend. However, for obvious reasons it is preferred that rating *N/A* is used as little as possible.

It is also interesting to note, that more than 10% of all the client instances in the data sample have received either *D* or *E* rating. This sub-set of clients contains more than 30% of all default clients of the entire data set.

5.2.2 Line of Business

Another variable is the **Line of Business**. Line of business significantly affects current and future economic situation of a client. The attractiveness of a particular line of business is enhanced by high profitability, potential growth rate of the sector etc.
Much like the previous variable, this one is also rated on a scale that ranges from *A* to *E* (best to worst) with an extra Not Available (*N/A*) rating.

Rating	Total Clients	Default	Non-Default	Information value
N/A	0	0	0	0.0000
A	0	0	0	0.0000
B	135	5	130	0.0057
C	3899	216	3683	0.0021
D	185	22	163	0.0384
E	8	2	6	0.0112
Total:	**4227**	**245**	**3982**	**0.0575**

Table 3 - Overview of information value of the second non-financial variable

As shown in Table 3 above, the information value of the variable called *Line of Business* is significantly lower than that of the previous variable.

Since the total information value is clearly under the pre-set predictive power threshold (0.2), this variable will be ignored for the purposes of creating the risk models.

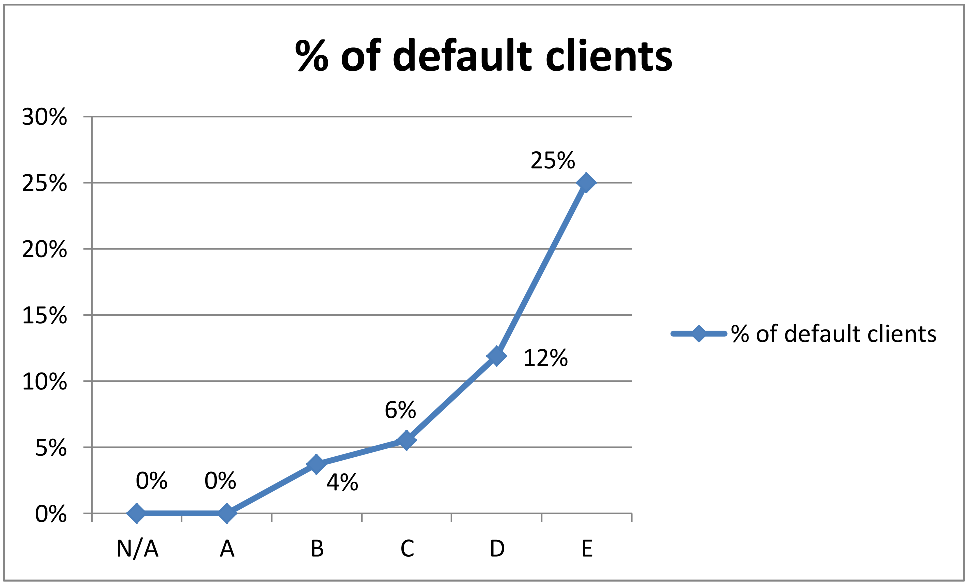

Figure 4 - Share of default clients in each rating within the second variable

The development represented in the curve above (Figure 4) is not surprising – the share of default clients increases as the ratings get worse – just like in the case of the previous variable.

However, it should be noted that only 8 out of 4227 clients have received the *E* rating, which greatly skews the results – just two default clients are enough to result in a 25% share of default clients within the rating.

Considering the low information value of the variable and the fact that the rating *C* is by far the most commonly received (or given) rating, it is not surprising that vast majority of default clients (216 out of 245) from the entire data sample have received this particular rating.

5.2.3 Market Position

The next variable is called **Market Position**. Market position must be evaluated in relation to the economic growth, i.e. the first step is to select the relevant market. Market ranges from international to regional. The selection of the market for evaluation is based on the company and its operations. Once the market is selected, the market position of a particular SME needs to be evaluated.

Rating scale of this variable does not differ from the previous variables in any way.

Rating	Total Clients	Default	Non-Default	Information value
N/A	0	0	0	0.0000
A	0	0	0	0.0000
B	169	7	162	0.0043
C	3232	185	3047	0.0001
D	817	52	765	0.0020
E	9	1	8	0.0015
Total:	**4227**	**245**	**3982**	**0.0079**

Table 4 - Overview of information value of the third non-financial variable

Table 4 above indicates that the information value of this variable is really low. It is even significantly lower than that of the previous variable - Line of Business, which is ignored for the purposes of creating risk models.

Therefore, this variable will be ignored as well. Once again, the rating *C* was given to the majority of the clients from the data sample, albeit not to the degree as in case of the previous variable.

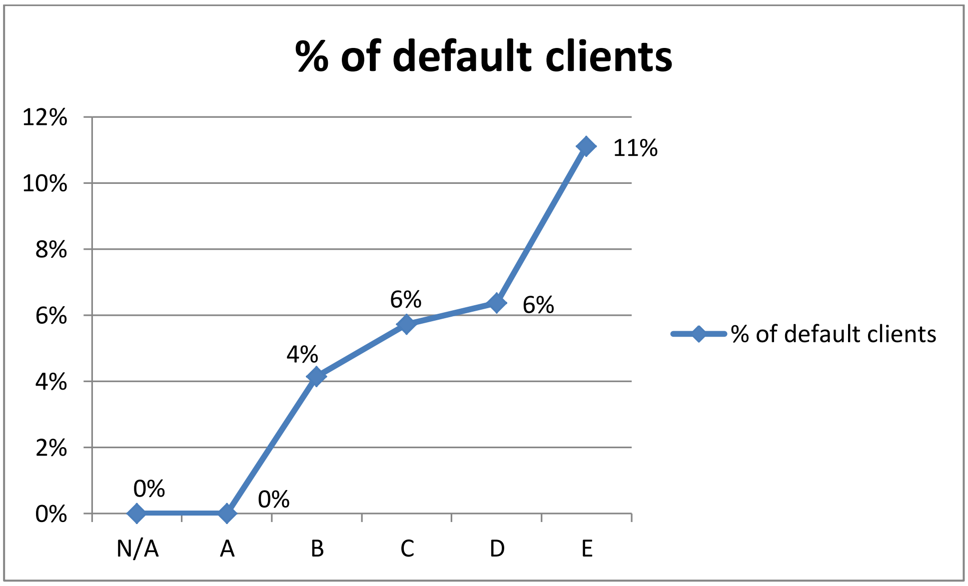

Figure 5 - Share of default clients in each rating within the third variable

Figure 5 above shows the expected trend – the share of default clients increases as the ratings get worse. However, as in the case of the previous variable, only very few clients actually received the *E* rating. On the other hand, almost 20% of all clients received the *D* rating, which still indicates weak market position.

5.2.4 Client's Perspectives

Another variable is called **Client's Perspectives**. This variable evaluates the future prospects of the client, but not in terms of their Line of Business, which is rated in the previous non-financial category, but in a sense of client's perspectives themselves. For example, a client modernized the manufacturing process in the company, but the positive impact of that investment has not yet fully manifested.

This category is rated on a scale from *A* to *E* (best to worst). A *Not Available* rating is available as well.

Rating	Total Clients	Default	Non-Default	Information value
N/A	5	1	4	0.0043
A	10	1	9	0.0011
B	296	27	269	0.0209
C	3395	132	3263	0.1177
D	499	75	424	0.2108
E	22	9	13	0.0810
Total:	**4227**	**245**	**3982**	**0.4358**

Table 5 - Overview of information value of the fourth non-financial variable

As we can see in Table 5 above, this variable is significantly above the set threshold, thus its predictive power and importance for the creation of risk models is very high. As in the case of the very first variable, one rating would pass the pre-set predictive power threshold on its own – this time, it is the *D* rating.

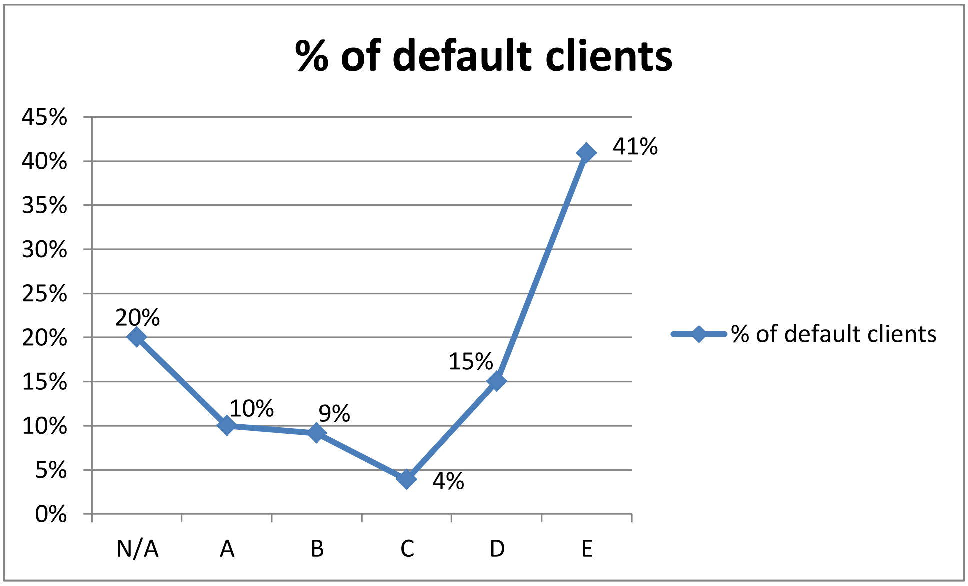

Figure 6 - Share of default clients in each rating within the fourth variable

Figure 6 above shows the share of default clients in each rating of this variable. As we can see, rating E boasts the greatest share of default clients at over 40%.
The 20% share of default clients for the *N/A* rating is due to the fact that only 5 clients received that particular rating.

The overall trend is unsurprising, albeit the *C* rating has lower share of default client than the A and B ratings, but that is again, caused by small number of clients that received either *A* or *B* rating.

5.2.5 Stability and Diversification of the Customers

The next variable is called **Stability and Diversification of the Customers**. Stability is related to the quality of client's relationship with the customers, whereas diversification is related to the quantity of these relationships. In general, the higher the quantity, the better spread of the business risk.

This category is rated on the same scale as the previous variable.

Rating	Total Clients	Default	Non-Default	Information value
N/A	9	2	7	0.0098
A	486	21	465	0.0096
B	1108	68	1040	0.0010
C	1052	67	985	0.0026
D	642	39	603	0.0004
E	930	48	882	0.0031
Total:	**4227**	**245**	**3982**	**0.0266**

Table 6 - Overview of information value of the fifth non-financial variable

As Table 6 indicates, the information value of this variable is negligible. Therefore, this variable will be ignored in the process of creating the risk models later on.

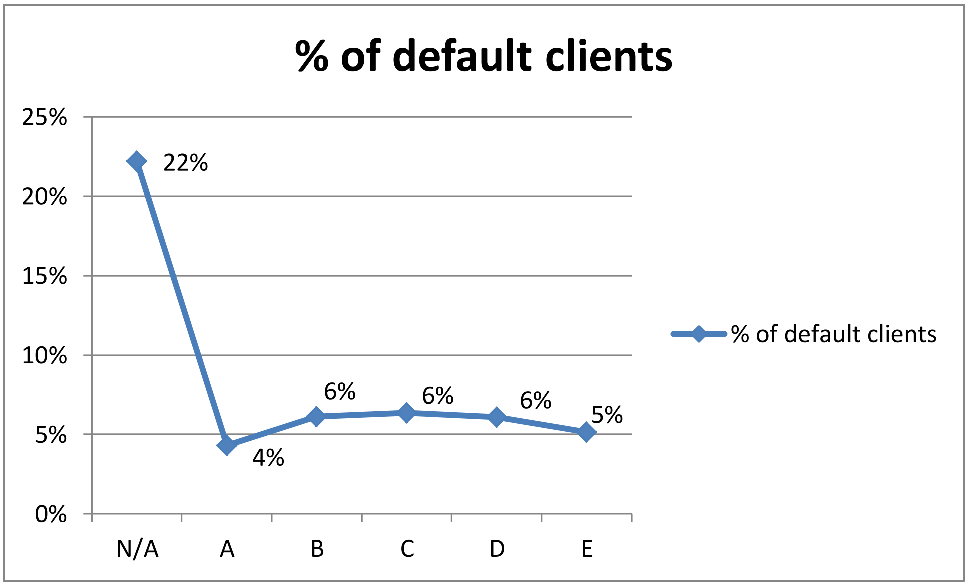

Figure 7 - Share of default clients in each rating within the fifth variable

Figure 7 shows a very interesting development. Notwithstanding the *N/A* rating's share of default clients, which is caused by the very low number of clients who received that rating, the share of default clients is evenly spread across the ratings. The reason for this development might lie in the fact that the total information value of this variable is negligible.

5.2.6 Sensitivity of Input Prices

Another variable is called **Sensitivity of Input Prices**. This variable indicates the degree of sensitivity of input on change of global pricing policy, seasonal fluctuations, pricing policy of the supplier etc.

This category is, again, rated on a scale that ranges from *A* to *E* (best to worst). A Not Available (*N/A*) rating is available as well.

Rating	Total Clients	Default	Non-Default	Information value
N/A	0	0	0	0.0000
A	11	2	9	0.0076
B	137	5	132	0.0062
C	970	52	918	0.0015
D	3039	182	2857	0.0009
E	70	4	66	3.74E-06
Total:	**4227**	**245**	**3982**	**0.0162**

Table 7 - Overview of information value of the sixth non-financial variable

As shown in Table 7 above, the information value of this variable is insignificant. For that reason, this variable will be ignored for the purposes of creation of the risk models.

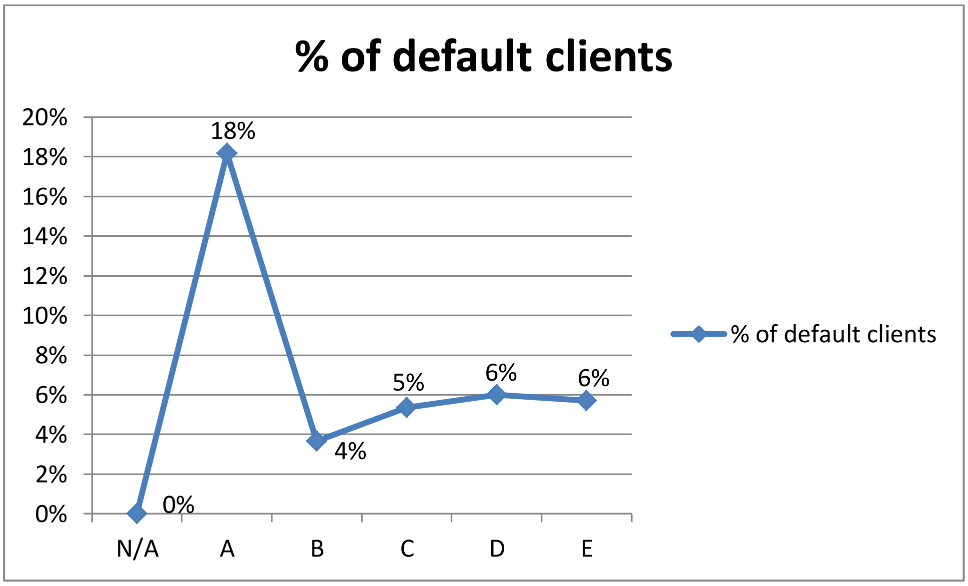

Figure 8 - Share of default clients in each rating within the sixth variable

Figure 8 outlines an interesting trend – clients who received the *A* rating have the highest share of default clients. However, much like in the previous cases of exceptional figures, this fact is caused by very low number of clients in the *A* rating set.

Other than that, the spread of default clients is even considering the number of clients within each rating set. Perhaps the reason for this trend lies in the fact that this variable possesses only a very low predictive power, just as in the case of the previous variable.

5.2.7 Cost of Output

The next variable is called **Cost of Output**. Much like the previous variable, this one depends on the market characteristics. However, SME clients have almost no impact on output prices. Ratings *A* and *B* are generally given to large corporate firms rather than SMEs. This category is rated on a scale from *A* to *E* (best to worst). A Not Available (*N/A*) rating is present as well.

Rating	Total Clients	Default	Non-Default	Information value
N/A	7	2	5	0.0129
A	15	1	14	8.44E-05
B	238	13	225	0.0002
C	2953	163	2790	0.0018
D	981	65	916	0.0050
E	33	1	32	0.0027
Total:	**4227**	**245**	**3982**	**0.0228**

Table 8 - Overview of Information value of the seventh non-financial variable

Much like the two previous variables, this variable has negligible predictive power. For that reason, it will not be used in the risk model. The fact that the rating *C* was given to most clients in the data set implies that the cost of output of a common SME from the data sample is average.

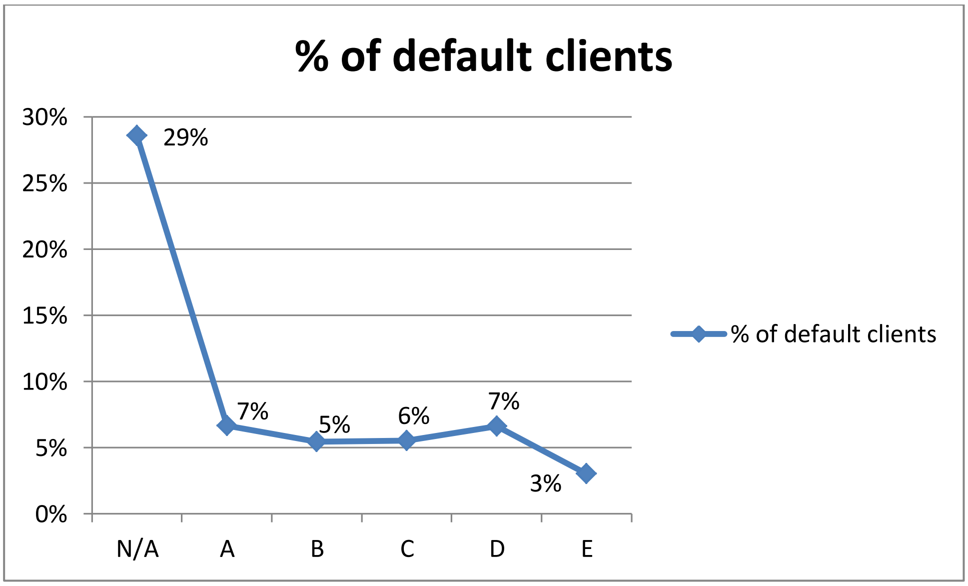

Figure 9 - Share of default clients in each rating within the seventh variable

Figure 9 shows even spread of default clients considering the size of each rating set. This trend is not refuted by ratings *N/A*, *A*, and *E*, all of which contain only a very small number of clients. As was the case with the two previous variables, the reason for this spread lies in the low predictive power of the variable.

5.2.8 Market Entry Barriers

Another variable is called **Market Entry Barriers**. Market entry barriers might be either objective (limited natural resources, high entry cost etc.) or subjective (state regulations, import/export quotas, licenses etc.).

The rating scale differs from the previous seven variables, as there is no *E* rating available for this variable.

Rating	Total Clients	Default	Non-Default	Information value
N/A	0	0	0	0
A	0	0	0	0
B	319	17	302	0.0006
C	3541	208	3333	0.0002
D	367	20	347	0.0004
Total:	**4227**	**245**	**3982**	**0.0011**

Table 9 - Overview of Information value of the eighth non-financial variable

For the fourth time in a row, a non-financial variable boasts only a very low predictive power. Therefore, it will not be used in the risk model. The fact that the rating *C* was given to most clients in the data set implies that the entry barrier to the market of common SME from the data sample is average.

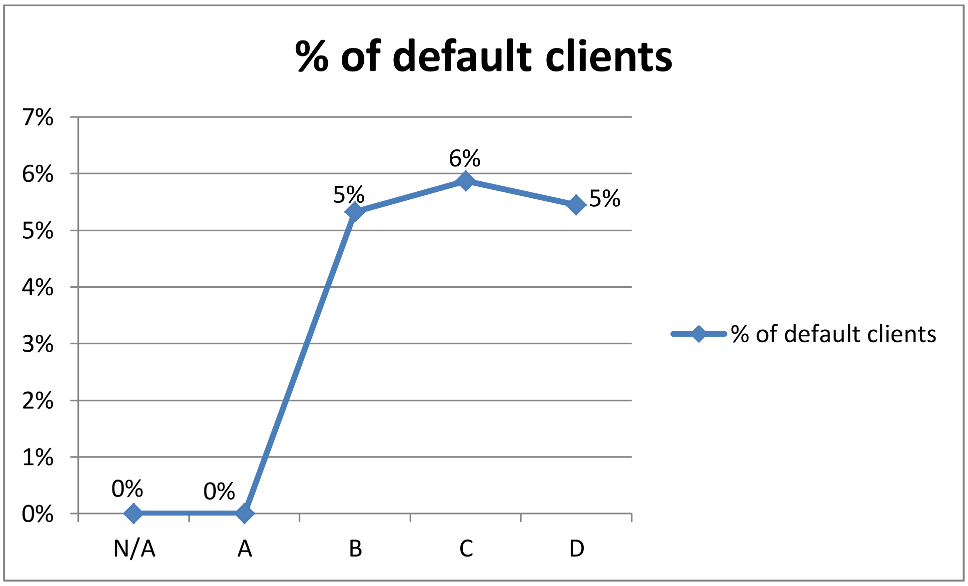

Figure 10 - Share of default clients in each rating within the eighth variable

Figure 10 depicts an even spread of clients for which data was available. No client in the data sample received *N/A* or *A* rating.

5.2.9 Results and Experience of the Management

The next variable is called **Results and Experience of the Management**. This variable refers to both quality and length of the managerial experience.

Key questions include the number of people in top management (i.e. is it a single person or are the responsibilities spread out?), current and past results (i.e. current and past profit levels etc.)
Although this variable is subjective in principle, it can be evaluated objectively, at least to some degree.

This category is rated on a scale from *A* to *E* (best to worst). A *Not Available* rating is present as well.

Rating	Total Clients	Default	Non-Default	Information value
N/A	0	0	0	0
A	0	0	0	0
B	0	0	0	0
C	1763	42	1721	0.2411
D	2432	196	2236	0.0844
E	32	7	25	0.0338
Total:	**4227**	**245**	**3982**	**0.3593**

Table 10 - Overview of Information value of the ninth non-financial variable

This variable has a high predictive power, and thus will be used for creating the risk model. The *C* rating possesses the highest information value at 0.24, which is caused by the fact that only 2% of clients who received this particular rating defaulted later on.

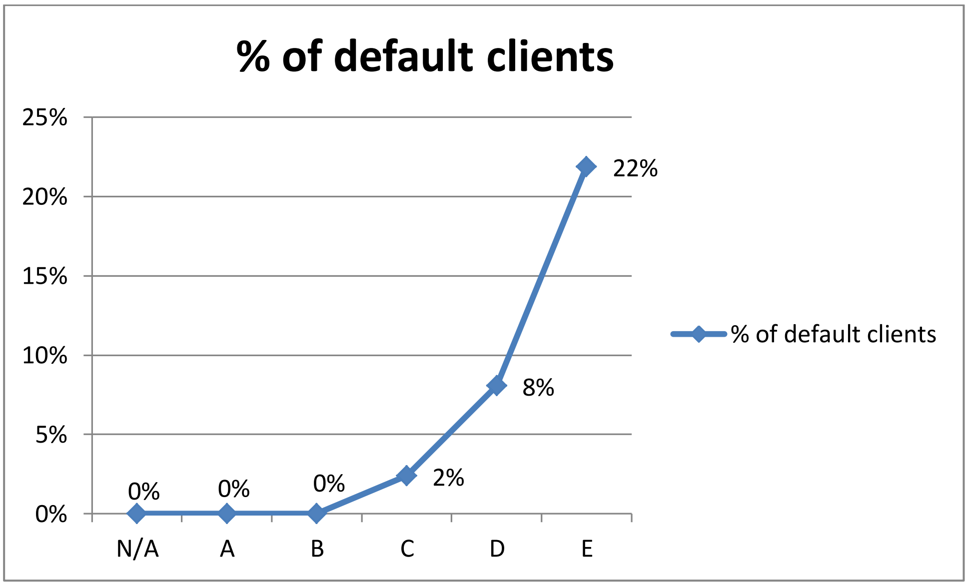

Figure 11 - Share of default clients in each rating within the ninth variable

Figure 11 outlines the share of default clients in each rating of this variable. As we can see, there is no data available for rating *N/A*, *A*, and *B*. The *E* rating boasts the greatest share of default clients at over 20%, which is, however, caused by very low number of clients in this rating set. The *C* rating was received by most default clients, which is not surprising, however, as this rating was given to vast majority of clients in the data set.

5.2.10 Quality of Information from the Client

Another variable is called **Quality of Information from the Client**. This variable rates the quality and accuracy of information that the client forwards to the bank, as well as the willingness and quickness of reaction to bank's requests.

This category is rated on a scale from *A* to *E* (best to worst). A *Not Available* rating is present as well.

Rating	Total Clients	Default	Non-Default	Information value
N/A	0	0	0	0
A	3	2	1	0.0275
B	60	2	58	0.0037
C	2867	124	2743	0.0563
D	1258	104	1154	0.0514
E	39	13	26	0.0975
Total:	**4227**	**245**	**3982**	**0.2365**

Table 11 - Overview of Information value of the tenth non-financial variable

Just like the previous variable, this one has passed the pre-set threshold, and for that reason, it will be used in the risk model. Unlike the previous variable, though, there is no outstanding rating that contributes to the overall information value of the variable in a significant way. The total information value is largely based on information value of ratings *C*, *D*, and *E*.

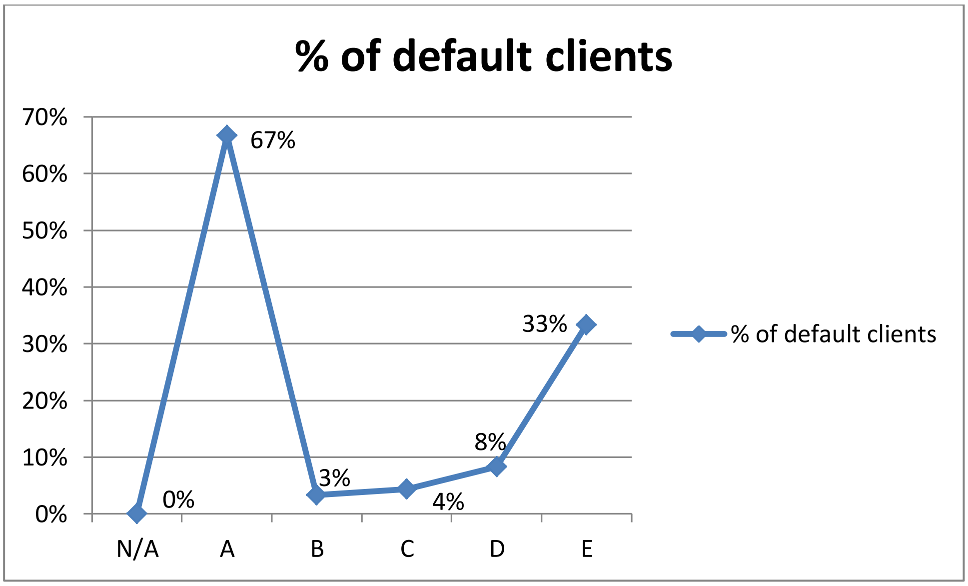

Figure 12 - Share of default clients in each rating within the tenth variable

Figure 12 represents the share of default clients in each rating set of this variable. Interestingly, the *A* rating boasts the greatest share of default clients at over 65%. However, this figure must be put into context – this outstanding number is caused by the fact that only a very small number of clients received the *A* rating. In absolute numbers, ratings *C* and *D* have the most default clients at 124 and 104, respectively.

Notwithstanding the aforementioned exception, the trend depicted in Figure 12 is not surprising – the share of default clients increases as the ratings get worse.

5.2.11 Obligations to the State

The next variable is called **Obligations to the State**. This variable has only three ratings (and a *N/A* rating).

A failure of meeting state obligations indicates poor morale of the company, and also potential inability to repay a loan to the bank as obligations to the state are more important than the other obligations. The reason for failure of meeting state obligations is irrelevant.

Rating	Total Clients	Default	Non-Default	Information value
N/A	0	0	0	0
A	4181	234	3947	0.0013
B	38	9	29	0.0477
C	8	2	6	0.0112
Total:	**4227**	**245**	**3982**	**0.0602**

Table 12 - Overview of Information value of the eleventh non-financial variable

This particular variable turned out to have a very low predictive power. For that reason, it will not be used in the creation process of risk models. However, the reason for low predictive power in this case is very likely the fact that regardless of the state of the client (default or non-default) obligations to the state are generally abided, which is also confirmed by the fact that vast majority (98.9%) of clients in the data sample received the *A* rating in this variable.

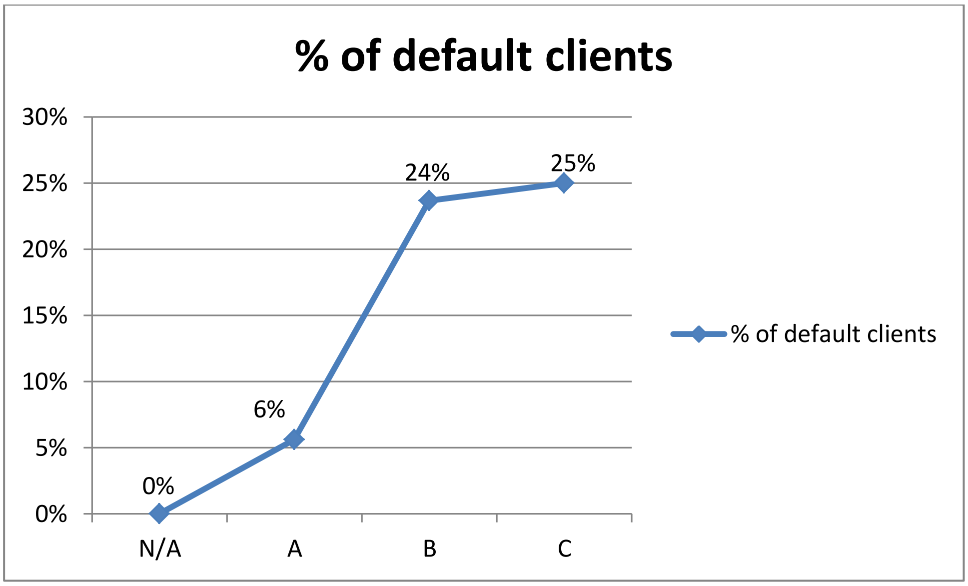

Figure 13 - Share of default clients in each rating within the eleventh variable

Figure 13 represents the share of default clients in each rating set of this variable. As we can see, the share increases as the ratings get worse.

However, all rating set except for the *A* rating set contain none or very small number of clients, which may lead to skewed results.

5.2.12 Turnover Development on Client's Accounts

The next variable is called **Turnover Development on Client's Accounts**. This variable reflects the intensity of relationship between the client and the bank (i.e. how much of the overall turnover of the SME takes place at the bank's accounts).

This category is rated on a scale from *A* to *E* (best to worst), with the addition of the *N/A* (Not Available) rating.

Rating	Total Clients	Default	Non-Default	Information value
N/A	0	0	0	0
A	29	1	28	0.0016
B	3895	192	3703	0.0250
C	281	39	242	0.0948
D	17	10	7	0.1228
E	5	3	2	0.0375
Total:	**4227**	**245**	**3982**	**0.2817**

Table 13 - Overview of Information value of the twelfth non-financial variable

Table 3 shows that this variable does have sufficiently high predictive power to be used in the risk model creation process.

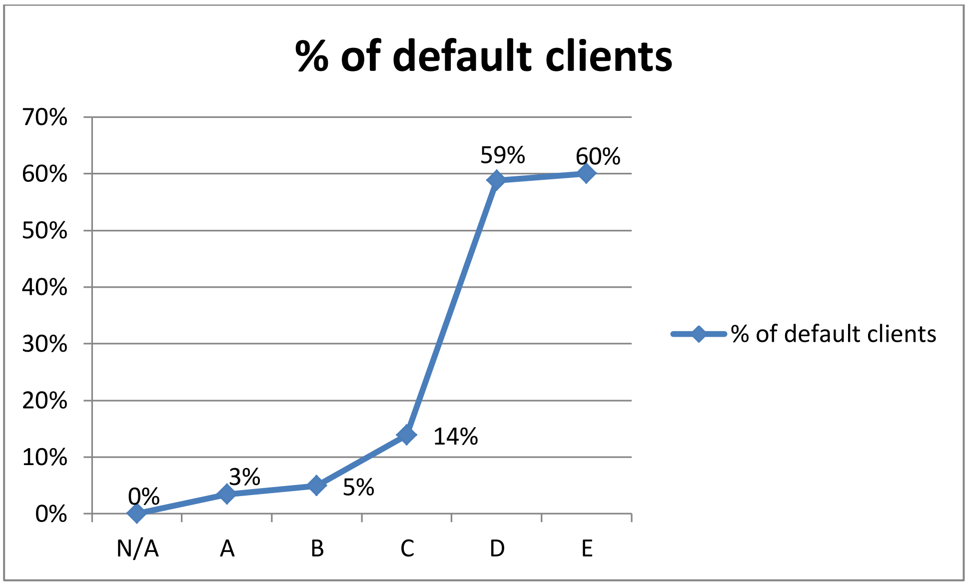

Figure 14 - Share of default clients in each rating within the twelfth variable

Figure 14 outlines that ratings *D* and *E* have the greatest share of default clients – both ratings are at, or near, the 60% share of default clients. However, this is caused by the very low number of clients in both rating sets. Notwithstanding the aforementioned outliners, the trend depicted in Figure 14 is not surprising. It is also worth noting, that majority of clients from the data sample have most of their turnover done at the bank's accounts.

5.2.13 Execution on client's accounts

Another variable is called **Execution on client's accounts**. Execution on the client's account is a serious signal that the financial situation of the client is complicated and needs to be resolved as soon as possible. The reason for execution is irrelevant.

This category is rated on a scale from *A* to *E* (best to worst), with the addition of the *N/A* (Not Available) rating.

Rating	Total Clients	Default	Non-Default	Information value
N/A	9	1	8	0.0015
A	20	1	19	0.0001
B	20	6	14	0.0407
C	79	24	55	0.1648
D	82	9	73	0.0128
E	4017	204	3813	0.0175
Total:	**4227**	**245**	**3982**	**0.2374**

Table 14 - Overview of Information value of the thirteenth non-financial variable

As shown in Table 14, this variable did pass the pre-set predictive power threshold. For that reason, it will be used for the risk model.

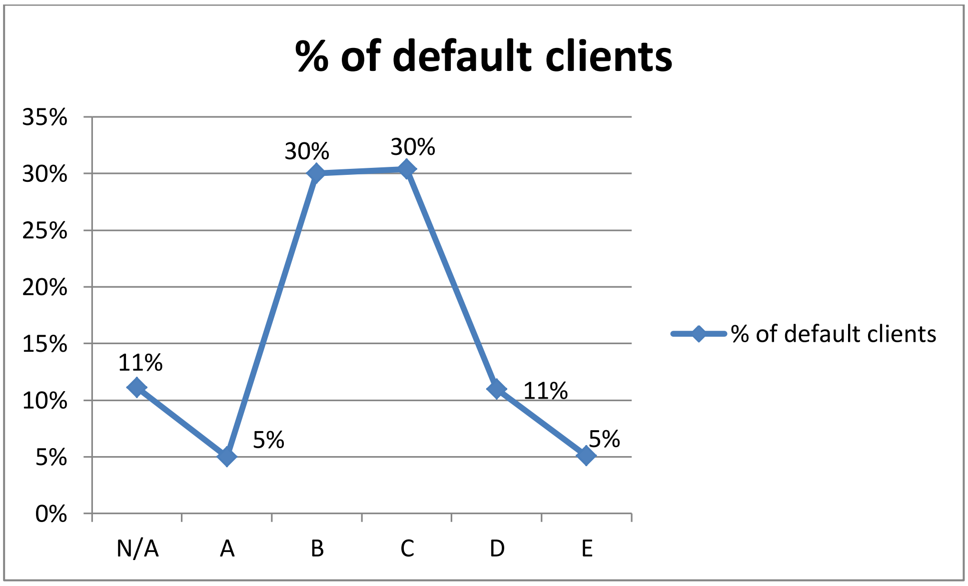

Figure 15 - Share of default clients in each rating within the thirteenth variable

Figure 15 outlines that ratings *B* and *C* have the greatest share of default clients. Both ratings are at 30% share of default clients, the reason for these two outliners can be, once again, found in the very low number of clients in both rating sets.

However, the *E* rating has by far the most default clients (in absolute numbers), which would not be surprising on its own, but the *E* rating was given to vast majority (95%) of the clients in the data sample.

The implication of this finding is clear – almost every client in the data set had an execution on his or her account at the time of creation of this data sample by the bank. This may obviously lead to skewed results later on. The potential gravity of this fact will be examined later in the chapter.

5.2.14 Fulfilment of Contractual Obligations

The next variable is called **Fulfilment of Contractual Obligations**. This variable is one of the basic indicators of economic situation of a client. Failure to fulfil contractual obligations or covenants may point out to potential problems that could de-stabilize the economic situation of a client.

This category is rated on a scale from *A* to *D* (best to worst), with the addition of the *N/A* (Not Available) rating.

Rating	Total Clients	Default	Non-Default	Information value
N/A	48	2	46	0.0012
A	46	1	45	0.0074
B	2463	88	2375	0.1203
C	1533	125	1408	0.0574
D	137	29	108	0.1344
Total:	**4090**	**245**	**3982**	**0.3207**

Table 15 - Overview of Information value of the fourteenth non-financial variable

Table 15 shows that this particular variable did pass the pre-set predictive power threshold rather easily. This makes it an important part of the eventual risk model.

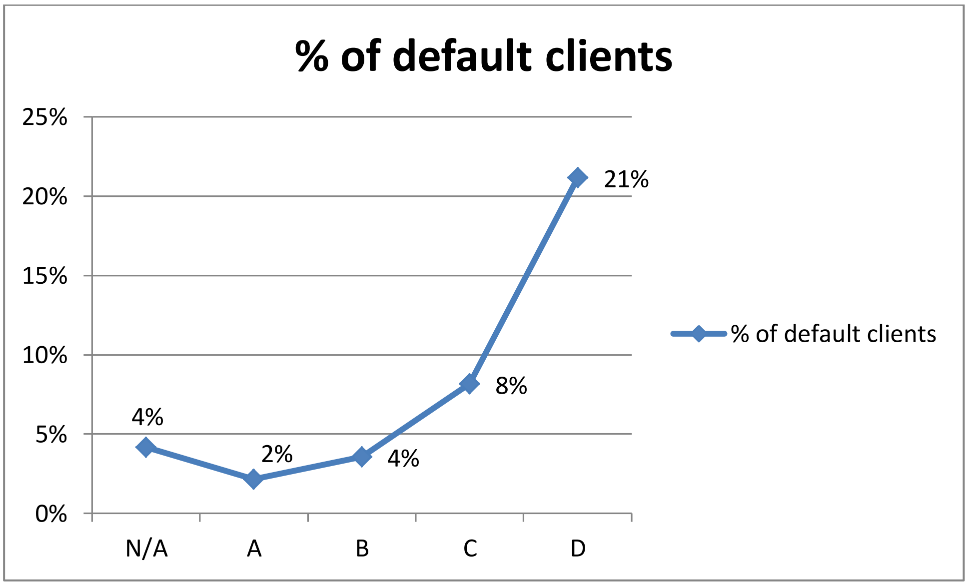

Figure 16 - Share of default clients in each rating within the fourteenth variable

Figure 16 outlines that the rating *D* has the greatest share of default clients, which is the expected development. However, as was the case in some of the previously discussed variables, rating *B* and *C* are the two most commonly given rating, and in absolute numbers, both of these two ratings have been given to more default clients than the *D* rating. This implies that most clients from the data sample are average or above-average when it comes to fulfilling contractual obligations.

5.2.15 Owners and Management

Another variable is called **Owners and Management**. A business or enterprise exists to maximize wealth of the owners. However, the internal relationships between managers, owners, employees, and the relationship between the company and its suppliers, customers, and competition are integral part of this goal.

This category is rated on a scale from *A* to *E* (best to worst), with the addition of the *N/A* (Not Available) rating.

Rating	Total Clients	Default	Non-Default	Information value
N/A	0	0	0	0
A	0	0	0	0
B	5	1	4	0.0043
C	4116	232	3884	0.0008
D	102	11	91	0.0149
E	4	1	3	0.0056
Total:	**4227**	**245**	**3982**	**0.0257**

Table 16 - Overview of Information value of the fifteenth non-financial variable

As shown in Table 16, this variable does not have the sufficient predictive power to be of use in the creation process of the risk model.

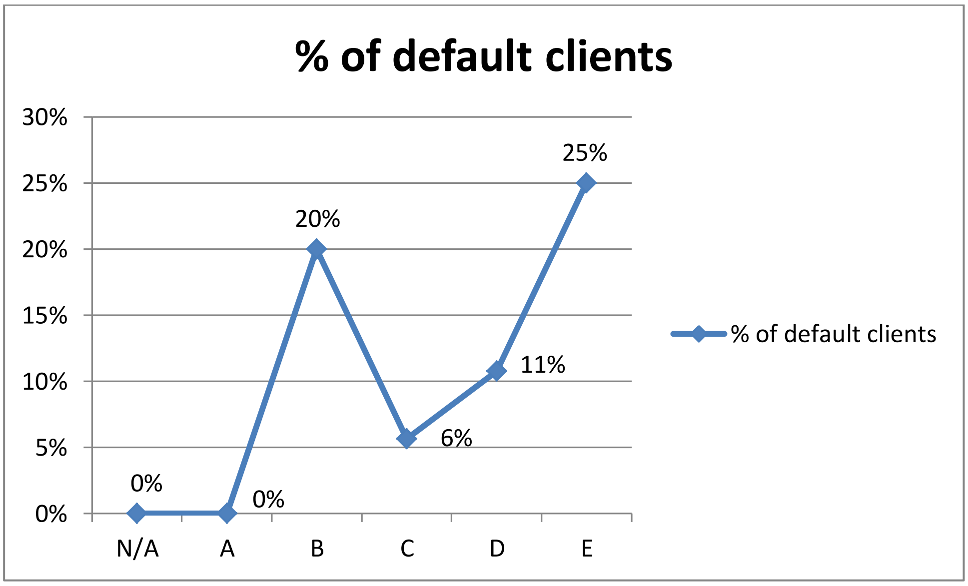

Figure 17 - Share of default clients in each rating within the fifteenth variable

Vast majority (97.3%) of the clients in the data sample were given the *C* rating, which implies average state of relationships within their respective companies. Consequently, most default clients (94.6%) received this rating. This observation is closely related to the poor predictive power of this variable.

5.2.16 Owners' Liability

The last variable is called **Owners' Liability**. This variable evaluates whether the owners themselves are liable, which is preferred in the SME segment, or if a third party is liable for the company's obligations.

This category is rated on a scale from *A* to *B* (best to worst), with the addition of the *N/A* (Not Available) rating.

Rating	Total Clients	Default	Non-Default	Information value
N/A	287	6	281	0.0488
A	3432	220	3212	0.0098
B	508	19	489	0.0208
Total:	**4227**	**245**	**3982**	**0.0794**

Table 17 - Overview of Information value of the sixteenth non-financial variable

Much like the previous variable, this one does not pass the pre-set predictive power threshold. For that reason, it will not be used in the risk model later on.

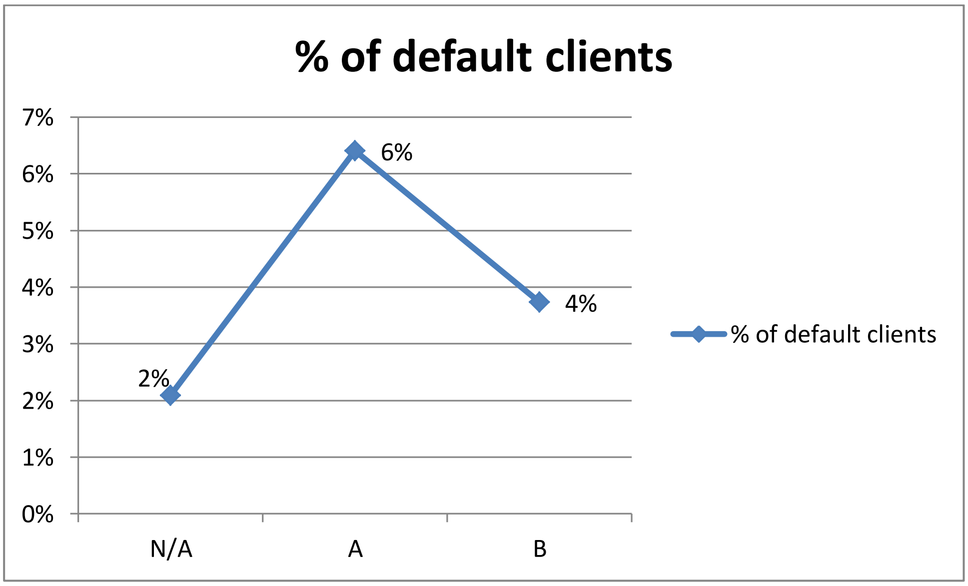

Figure 18 - Share of default clients in each rating within the sixteenth variable

Considering the poor predictive power, it is not surprising that one rating (in this case rating *A*) was given to majority (81.2%) of the clients in the data sample. This also implies that the majority of SME owners are liable for the company. Figure 18 shows relatively even spread of default client share.

5.2.17 Summary of Information value calculation

Now that the Information value was calculated for every single variable, a table that sums up the overall standings may be created:

Variable (category)	Information value
Recent Development of the Financial Situation	**0.5158**
Line of Business	0.0575
Market Position	0.0079
Client's Perspectives	**0.4358**
Stability and Diversification of the Customers	0.0266
Sensitivity of Input Prices	0.0162
Cost of Output	0.0228
Market Entry Barriers	0.0011
Results and Experience of the Management	**0.3593**
Quality of Information from the Client	**0.2365**
Obligations to the State	0.0602
Turnover Development on Client's Accounts	**0.2817**
Execution on client's accounts	**0.2374**
Fulfilment of Contractual Obligations	**0.3207**
Owners and Management	0.0257
Owners' Liability	0.0794
TOTAL	2.6845

Table 18 – Overview of information value for every examined variable

Variables that have passed the Information value threshold of 0.2, and thus have sufficiently high predictive power and will be used in the risk model, are bolded. Only the variables that have passed this pre-set threshold will be used in the next part of the model creation process. The logic behind this decision is as follows – reducing the number of variables from 16 to 7 (i.e. 56% reduction) results in the loss of only 0.2973 of the total information value (i.e. 11% information value loss). In other words, the reduction of the number of relevant variables is easily worth the consequential loss of part of the total information value.

To verify (and potentially enhance) the accuracy of the eventual risk model, the original data sample was also randomly divided into two parts of roughly the same size.
I.e. Sample 1 and Sample 2 are sub-sets of the original data sample provided by the bank.

Sample 1 consists of 126 default clients and 1988 non-default clients.

Sample 2 consists of 119 default clients and 1994 non-default clients.

Variable (category)	**Information value**
Recent Development of the Financial Situation	**0.8225**
Line of Business	0.0993
Market Position	0.0142
Client's Perspectives	**0.6655**
Stability and Diversification of the Customers	0.0623
Sensitivity of Input Prices	0.0134
Cost of Output	0.0344
Market Entry Barriers	0.0246
Results and Experience of the Management	**0.5197**
Quality of Information from the Client	**0.3293**
Obligations to the State	0.0785
Turnover Development on Client's Accounts	**0.3317**
Execution on client's accounts	*0.1932*
Fulfilment of Contractual Obligations	**0.5030**
Owners and Management	0.0623
Owners' Liability	0.1243
TOTAL	3.8782

Table 19 - Overview of information value – based on Sample 1

Information value table for Sample 2:

Variable (category)	**Information value**
Recent Development of the Financial Situation	**0.2683**
Line of Business	0.0238
Market Position	0.0076
Client's Perspectives	**0.2474**
Stability and Diversification of the Customers	0.0287
Sensitivity of Input Prices	0.0232
Cost of Output	0.0198
Market Entry Barriers	0.0167
Results and Experience of the Management	**0.2366**
Quality of Information from the Client	*0.1170*
Obligations to the State	0.0482
Turnover Development on Client's Accounts	0.1969
Execution on client's accounts	**0.2974**
Fulfilment of Contractual Obligations	0.1744
Owners and Management	0.0006
Owners' Liability	0.0314
TOTAL	1.7380

Table 20- Overview of information value – based on Sample 2

As shown in Tables 19 and 20, the same seven categories either passed the pre-set threshold for high predictive power or are close to it. This confirms that the seven categories (variables) have been chosen correctly.

The same two samples will be also used for creating separate risk models, and the models will be evaluated (i.e. model based on the data from Sample 1 will be used to assess risk level of clients in Sample 2, and vice versa) and compared to the model based on the original data sample.

5.3 The Excel Model

Using the information value, it is possible to create a mathematical model for default client detection. As mentioned, only the seven most important categories are used. The remaining variables are ignored as their predictive power is not high enough.

Therefore, the following table is the starting point for creating the mathematical risk model:

Variable (category)	**Information value**
Change of the Overall Financial Situation	0.5158
Client's Perspectives	0.4358
Results and Experience of the Management	0.3593
Quality of Information from the Client	0.2365
Turnover Development on Client's Accounts	0.2817
Execution on client's accounts	0.2374
Fulfilment of Contractual Obligations	0.3207
Total	**2.3872**

Table 21 – Information value of variables with high predictive power

To create the transformation matrix, a numerical value needs to be assigned to each of the sixteen variables. In order to do that, it is imperative that a weight in percentage is calculated for every variable.

Since we know that the total Information value is 2.3872, which is 100% of the overall information value for the remaining seven variables that are being used, and we do know the information value of each variable, we can calculate the weighting of the individual variables.

The weighting of the individual variables is the quotient of the information value of a particular variable and the total information value of all variables combined.

For example, the percentage weighing (PW) of the variable called Recent Development of the Financial Situation (RDFS) is calculated in the following way:

$$PW\ RDFS = \frac{0.5158}{2.3872} = 0.216$$

This value is rounded to three decimal places and then converted to percentage. Thus we get a value of 21.6%.

The percentage weighting of the rest of the variables is calculated in the same way.

The following table shows percentage weighing of all variables.

Variable (category)	**Information value**	**Percentage weighting**
Change of the Overall Financial Situation	0.516	21.61%
Client's Perspectives	0.436	18.26%
Results and Experience of the Management	0.359	15.05%
Quality of Information from the Client	0.236	9.91%
Turnover Development on Client's Accounts	0.282	11.80%
Execution on client's accounts	0.237	9.94%
Fulfilment of Contractual Obligations	0.321	13.43%
Total	**2.387**	**100.00%**

Table 22 – Percentage weighing for every relevant variable

Since the percentage weighting distribution has been established, it is time to decide what the sum of the maximum numerical risk values associated with the variables in the transformation matrix will be.

For the purposes of this thesis, the number 1000 was chosen as it offers sufficient level of detail while maintaining a simple and easy to understand risk model.

Since we already know the relevance in percentage, and we have selected the total maximum value in a numerical way, we can calculate the maximum number of points given to each of the seven categories.

Variable (category)	Information value	Percentage weighting	Max value
Change of the Overall Financial Situation	0,516	21,61%	216
Client's Perspectives	0,436	18,26%	183
Results and Experience of the Management	0,359	15,05%	151
Quality of Information from the Client	0,236	9,91%	99
Turnover Development on Client's Accounts	0,282	11,80%	118
Execution on client's accounts	0,237	9,94%	99
Fulfilment of Contractual Obligations	0,321	13,43%	134
Total	**2,387**	**100,00%**	**1000**

Table 23 – Point distribution for all relevant variables

Table 23 above provides an overview of the maximum number of points for each of the seven categories. Now it is needed to calculate the point spread within each rating.

The variable Recent Development of the Financial Situation will be, once again, used for demonstrating how the calculation is done.

We know that the maximum value given to a rating of this variable is 216. The table below indicates that the rating with the highest predictive power is the *B* rating. Since it has the largest share of default clients, it will be given the maximum - 216 risk points.

Table 24 below provides the overview of the situation (numbers rounded to the nearest integer):

Rating	Default clients	Share of Default clients % within the rating
N/A	2	1%
A	7	3%
B	105	43%
C	55	22%
D	46	19%
E	30	12%
Total:	**245**	**100%**

Table 24 – Default client share in ratings within one variable

The next step is to calculate the share the *B* rating has in the whole set of ratings. Since the *B* rating has identified 42.857% of the default clients, the number 216 presents 42.857% of the total value of risk points given to this variable.

Therefore, it is possible to determine the point value for the whole set of ratings, i.e. 100%.

$$Total\ Points = \frac{216}{0.42857}$$

$$Total\ Points = 504$$

Now that we know the total value for the whole set of ratings, it is possible to calculate numerical value for all ratings in this variable.

To ensure that this approach is correct, we can check by multiplying 504 by 42.857%. The result of this operation is 216, which is equal to the original maximum value given to this variable.

Using this approach, it is possible to calculate the numerical point value for all ratings in this set. The following table shows the results.

Rating	Default clients	Share of Default clients % within the rating	Points given
N/A	2	1%	4
A	7	3%	14
B	105	43%	216
C	55	22%	113
D	46	19%	95
E	30	12%	62
Total:	**245**	**100%**	**504**

Table 25 – Risk points given to individual ratings within one variable

Point distributions of all other variables are done in the same way.

The transformation matrix created by this approach is below (note that the numbers are rounded to the nearest integer, which may lead to slight inaccuracies; please see the attached Excel sheets for 100% accurate display):

Category/Rating	N/A	A	B	C	D	E	Total	Max
Recent Development of the Financial Situation	4	14	216	113	95	62	504	216
Client's Perspectives	1	1	37	183	104	12	339	183
Results and Experience of the Management	0	0	0	32	151	5	188	151
Quality of Information from the Client	0	2	2	99	83	10	196	99
Turnover Development on Client's Accounts	0	1	118	24	6	2	151	118
Execution on client's accounts	0	0	3	12	4	99	119	99
Fulfilment of Contractual Obligations	2	1	95	134	31		263	134

Table 26 – The Transformation Matrix

As an example, a random client instance has been selected. Its ratings in respective variables have been put into the transformation matrix (bolded).

Category/Rating	N/A	A	B	C	D	E	Total	Max
Recent Development of the Financial Situation	4	14	**216**	113	95	62	504	216
Client's Perspectives	1	1	**37**	183	104	12	339	183
Results and Experience of the Management	0	0	**0**	32	151	5	188	151
Quality of Information from the Client	0	2	2	**99**	83	10	196	99
Turnover Development on Client's Accounts	0	1	118	**24**	6	2	151	118
Execution on client's accounts	0	0	3	12	4	**99**	119	99
Fulfilment of Contractual Obligations	2	1	**95**	134	31		263	134

Table 27 – Application of transformation matrix on a randomly selected client

The application of scalar operation gives the following result:

$$\begin{aligned} \mathit{Risk\ points} &= 1 \times 216 + 1 \times 37 + 1 \times 0 + 1 \times 99 + 1 \times 24 + 1 \times 99 + 1 \times 95 \\ &= 384 \end{aligned}$$

All the values calculated for the variables in the transformation matrix were added together. This sum is then divided by the sum of the maximum risk values given to every variable. A percentage figure is the result of the aforementioned operation. However, this result needs to be interpreted in the retransformation matrix later.

The example client mentioned above received 384 points, therefore:

$$100 \times 384 \div 1000 = 38.4\%$$

However, the transformation matrix is inefficient for testing large quantities of data, which is why it needs to be converted into formulas, which can then be used for handling large quantities of data.

The formulas are created using the IF-THEN system of conditions.

For example the formula for the first variable (Recent Development of the Financial Situation) will be (generalized):

IF column 1 = "1" THEN "4" else IF = "2" THEN "14" else IF = "3" THEN "216" else IF = "4" THEN "113" else IF = "5" THEN "95" else IF = "6" THEN "62"

Below is an example of a formula that was actually used:

=KDYŽ(B5=2;"4";KDYŽ(B5=3;"14";KDYŽ(B5=4;"216";KDYŽ(B5=5;"113";KDYŽ(B5=6;"95";KDYŽ(B5=7;"62"))))))

These formulas are applied to the whole data sample.

Full list of formulas that were actually used to evaluate the performance of this model is available in **Appendix 1**.

5.4 Evaluation of the Excel model

The model has been created and used on the data sample. Now it is time to evaluate its performance (i.e. accuracy of risk evaluation). In order to perform an evaluation, a retransformation matrix needs to be created.

As a starting point, the following retransformation matrix will be used:

Percentage of risk points received	**Linguistic variable**
0% to 70%	Non-default
70+%	Default

Table 28 – The basic retransformation matrix

The application of the combination of transformation matrix and the retransformation matrix on all instances in the data sample yielded the following results:

Default client?	**Identified as default based on the Threshold (70%)**	**Result**	**Number of clients**
True	True	OK	170
True	False	Error I	75
False	False	OK	537
False	True	Error II	3445

Table 29 – Results given by the Excel risk model based on the first retransformation matrix

Error I occurs when the model identifies a default client as non-default.
Error II occurs when the model identifies a non-default client as default.

As the Figure 29 shows, the model has successfully detected 170 out of 245 default clients, which is approximately 70% success rate. However, this default detection success rate comes at a cost – the model incorrectly identified an overwhelming number of non-default clients as default – 3445, which is approximately 87% of all non-default clients in the entire data sample.

The total success rate (*TSR*) of the model for a given retransformation matrix may be calculated by adding the correctly identified clients together and dividing it by the total number of clients in the entire data sample.

$$TSR = \frac{170 + 537}{4227} = 0.31$$

The Total Success Rate (TSR) at the default threshold set to 70% is approximately 31%.

However, this retransformation matrix (set default threshold) serves merely as a starting point.
Further testing is needed to find the optimal retransformation matrix (default threshold). The model is tested using thresholds that range from 60% to 95%.

Threshold %	Default clients detected correctly	Error I	Non-Default clients detected correctly	Error II
60%	215	30	154	3828
70%	170	75	537	3445
80%	108	137	1214	2768
85%	85	160	2132	1850
90%	59	186	2783	1199
95%	41	204	3173	89

Table 30 - Overview of the model's performance at various thresholds

It is possible to draw several conclusions from Table 30 above. A number of graphs will be used to better demonstrate the consequences of various default thresholds on model's performance.

Unless stated otherwise, all the following graphs in this sub-chapter display the default threshold value at the X axis, while the success rate is displayed on the Y axis.

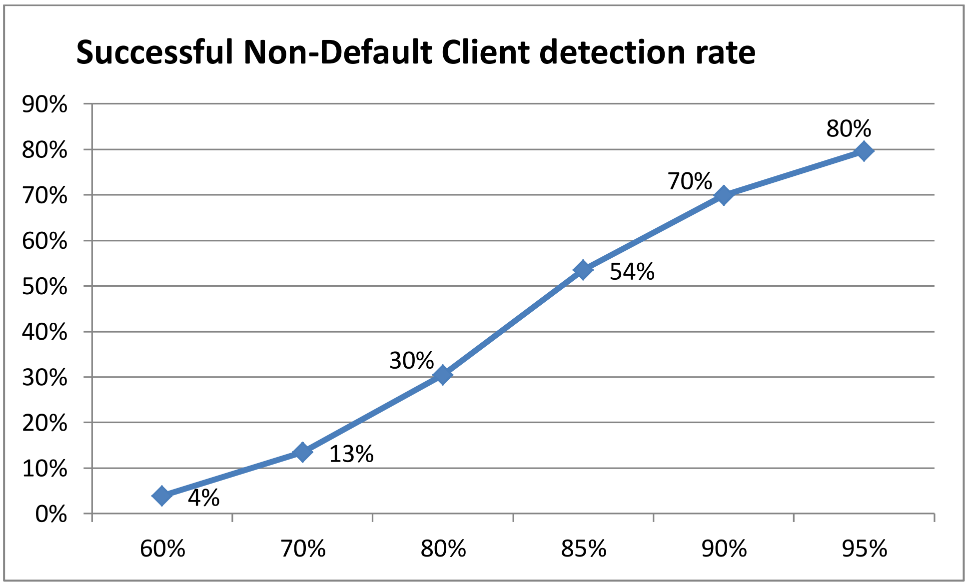

Figure 19 – Overview of correctly identified non-default clients

Figure 19 depicts the fact that as the default threshold (X axis) increases, so does the success rate of non-default client detection (identification), which is displayed on the Y axis. This is a natural consequence of the fact that as the threshold goes up, the sample size remaining in play shrinks.

The curve for Error I rate (Figure 20) is strikingly similar to the curve for successful non-default client detection (Figure 19). This is a logical consequence of the previously described trend – as the threshold goes up, the model identifies more and more clients as non-default.

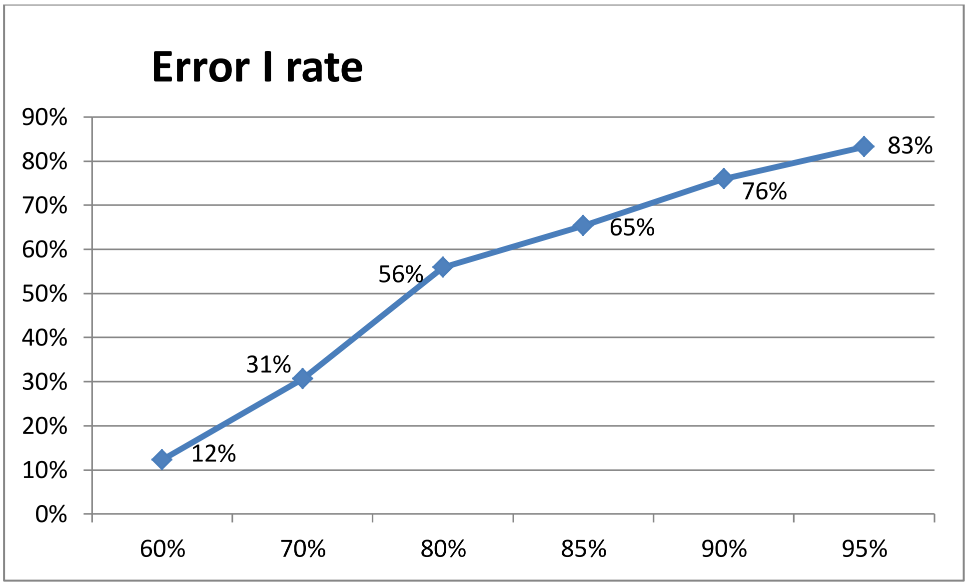

Figure 20 – Error I rate (default identified as non-default) at various thresholds

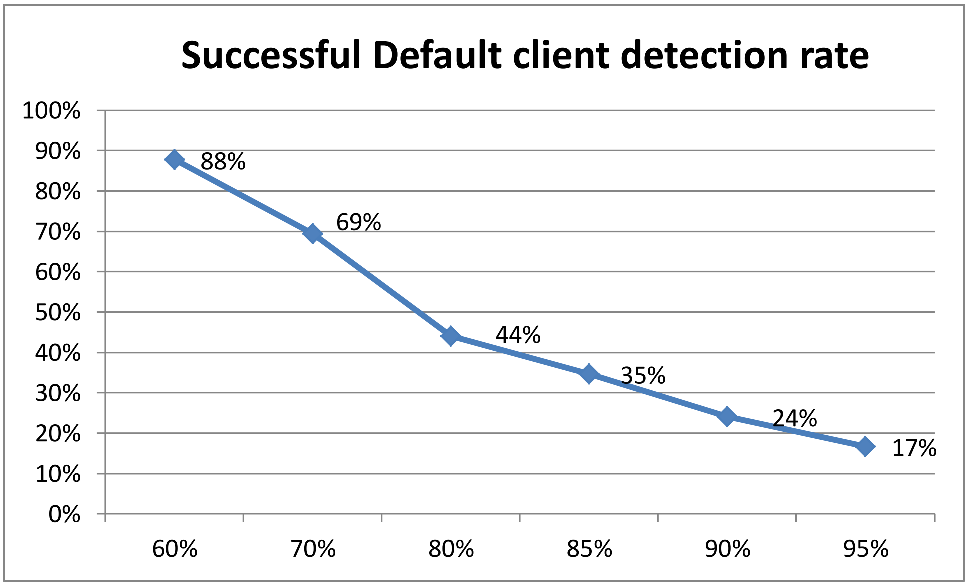

Figure 21 – Rate of successful Default client detection at various thresholds

The two curves outlined in Figure 21 and Figure 22 share the same development. As the threshold increases, the success rate of default client detection **and** Error II rate decrease.

This can be explained by the fact that the share of default clients within the entire data set is spread out from the lowest values to the highest ones. Consequently, increasing threshold shrinks the pool of default clients in play, and as a result, the success rate of default client detection decreases.

By contrast, the Error II rate (Figure 22) decreases because the pool of clients identified as non-default grows as the default threshold increases.

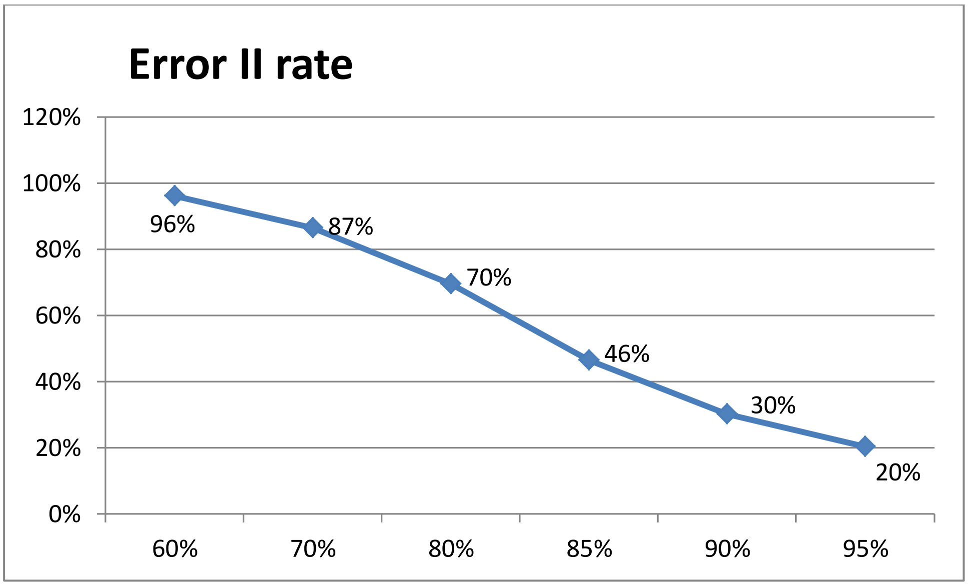

Figure 22 – Error II (non-default identified as default) rate at various thresholds

By combining all the above graphs into a single one, we get the overview of the whole situation:

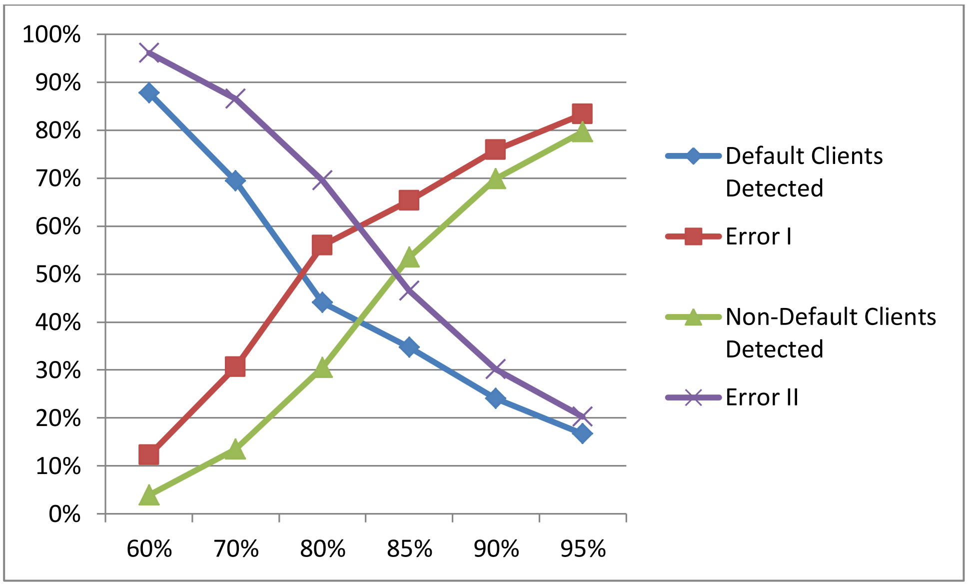

Figure 23 – Overview of the four Success (2) and Error (2) rates

The graph above shows default and non-default detection success rates and Error I and II rates at various thresholds. The success rate is displayed on the Y axis, while the thresholds are on the X axis.

The table below sums up the results including the Total Success Rate of the model at various thresholds.

Threshold %	Default clients detected	Error I	Non-Default clients detected	Error II	Total Success Rate
60%	88%	12%	4%	96%	9%
70%	69%	31%	13%	87%	17%
80%	44%	56%	30%	70%	31%
85%	35%	65%	54%	46%	52%
90%	24%	76%	70%	30%	67%
95%	17%	83%	80%	20%	76%

Table 31 – Complete summary of all error and success rates at various thresholds

The graph below adds the Total Success Rate into the view on the model's performance.

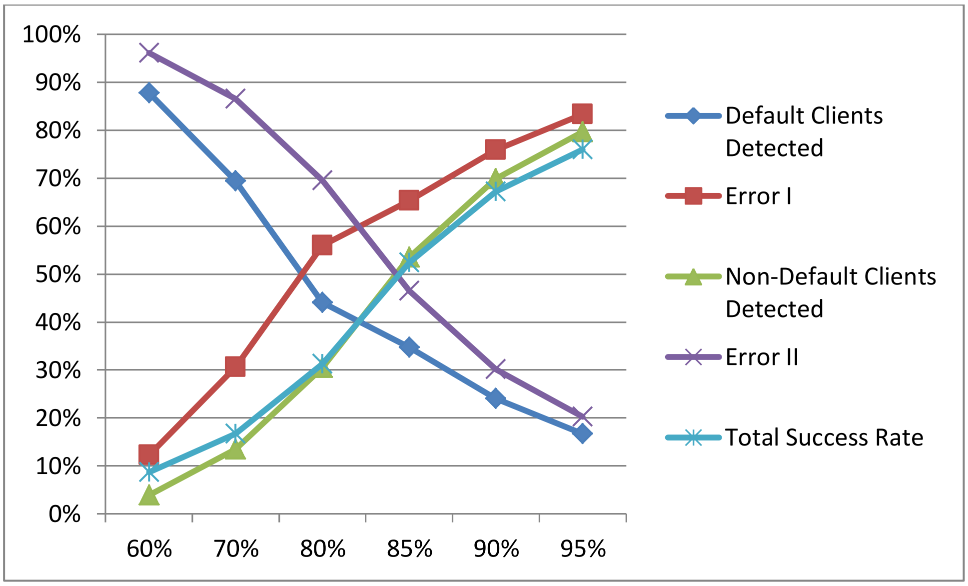

Figure 24 – Overview of all three success rates and both error rates

It is apparent that as the threshold increases, so does the Total Success Rate. However, the success of default client detection rapidly decreases despite the increasing Total Success Rate. For that reason, the bank must use its own preferences to define the retransformation matrix for the fuzzy model.

Should the successful identification of default clients be the most important characteristic, then a low default threshold is required – in this case a threshold of 60% would be appropriate.

A more balanced model requires the usage of the intersection of the default and total success rate curves that occurs between the 80% and 85% default threshold.

At a first glance, it may seem that the risk model's performance is lacking, but it is imperative that further analysis is done.

First of all, the data sample itself is imperfect for this kind of analysis and application. That is mainly due to the fact that the share of default clients is far smaller than share of the non-default clients. Fortunately, the Excel fuzzy risk model can be easily modified when an additional, or even entirely new, sample becomes available.

Second, this model does not account for the financial part of the risk evaluation, on which the bank's internal system is built, and to which this Excel risk model is compared. It is entirely possible that a client with a below average non-financial rating boasts with great financial results. And vice versa, a client with outstanding non-financial rating might go bankrupt without the model ever accounting for that. The Excel model designed in this thesis does not, and can't, account for that.
In addition, even though the bank was asked to provide data sample of client instances with roughly the same financial results, this may not be verified conclusively.

Third, the non-financial variables of a client may change abruptly – without the system noticing the change quickly enough. This naturally leads to unnecessary errors.

Fourth, an overwhelming majority of clients in the data sample have had an execution on their account. This implies that their financial situation was in a bad state, or close to it. This fact negatively influences the overall importance of non-financial factors in this client set.

In light of these facts, the performance of the model is deemed satisfactory. More importantly, the appropriateness of application of fuzzy logic has been confirmed.

The advantages of this fuzzy model are immense – easy maintenance, scalability, and modification, and no special software requirements.

As side products to the main model based on the entire data sample, two more models have been created. The model based data analysis of Sample 1 was applied to data in Sample 2, and vice versa.

The result of application of the model based on data analysis of Sample 1 on the data in Sample 2:

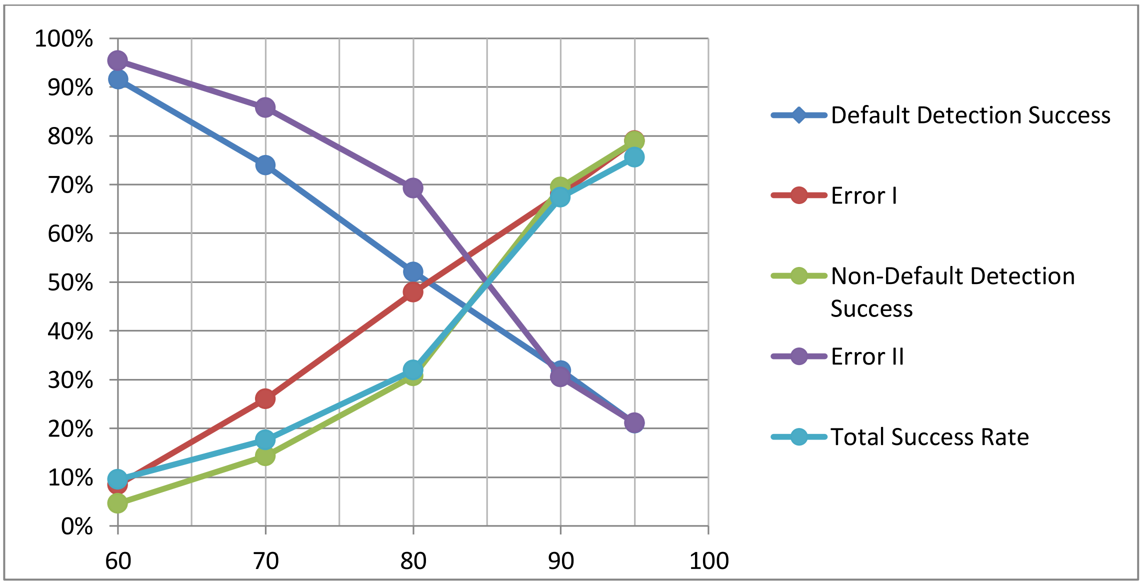

Figure 25 - Overview of all three success rates and both error rates for model based on data from Sample 1

The result of application of the model based on data analysis of Sample 2 on the data in Sample 1:

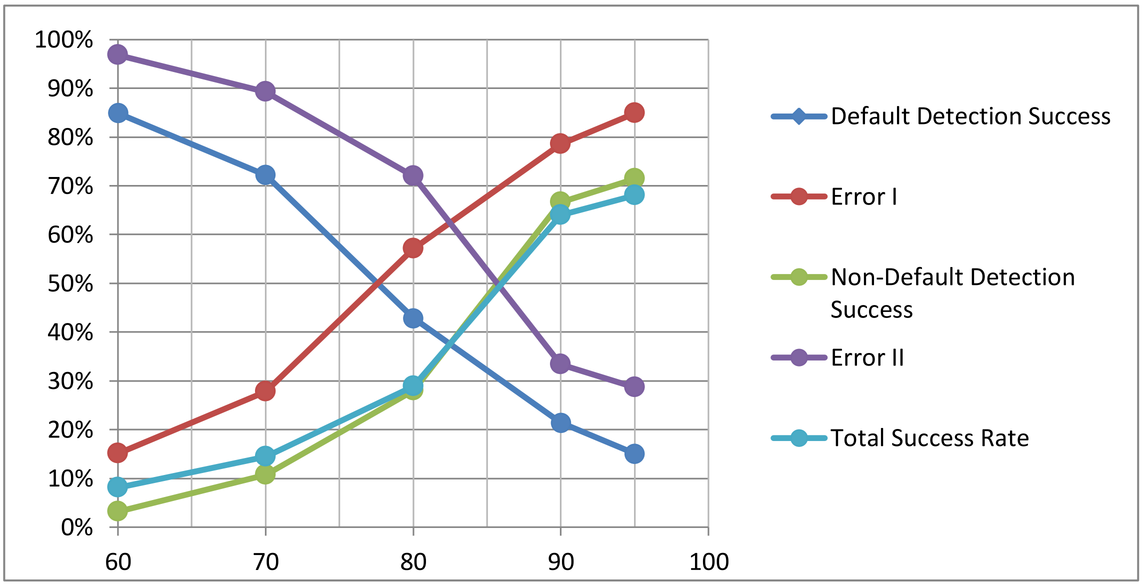

Figure 26 - Overview of all three success rates and both error rates for model based on data from Sample 2

The models based on Sample 1 and 2, respectively, give similar results to the model based on the entire data sample. This confirms that the trends observed in the evaluation of the Excel risk model based on the entire data sample are not flukes.

As outlined in Figures 25 and 26, the very same trends (as in the main risk model evaluation) can be observed – as the threshold increases, so does the Total Success Rate, correct Non-Default detection rate, and Error I rate (i.e. default client identified as non-default). By contrast, the default detection success rate and the Error II rate (i.e. non-default client identified as default) decrease.

5.5 The MATLAB model

Figure 27 below shows the correlation between ratings given to default and non-default clients in each of the 16 non-financial categories.

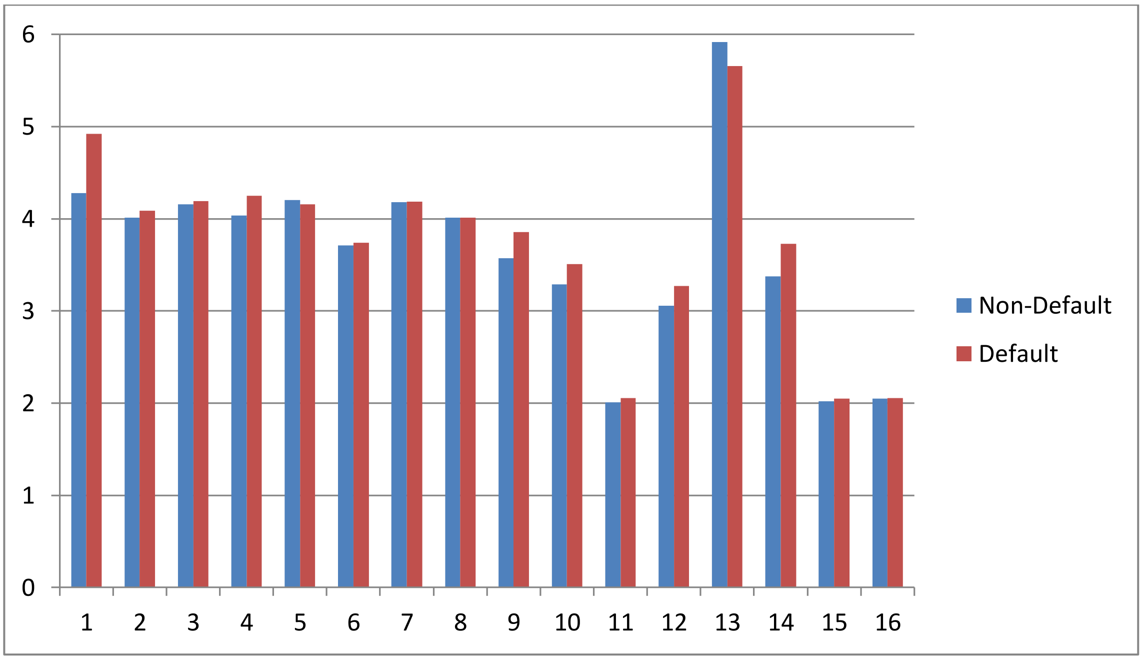

Figure 27 – The difference in rating of default and non-default clients for all non-financial variables

We can see that the differences between the ratings of default and non-default clients are generally very small. X-axis – variable; Y-axis – rating.

The largest differences are apparent in categories 1, 4, 9, 10, 12, 13, and 14. These seven variables happen to have the highest predictive power out of all variables in the data sample, as evidenced by means of information value calculation in sub-chapter 5.2. It is also worth nothing, that Figure 27 provides another view on the fact that an overwhelming majority of clients in the data sample have had an execution on their account (variable 13) at the time of creating the data sample.

Therefore, the MATLAB model will work with the same seven categories that were used in the Excel model.

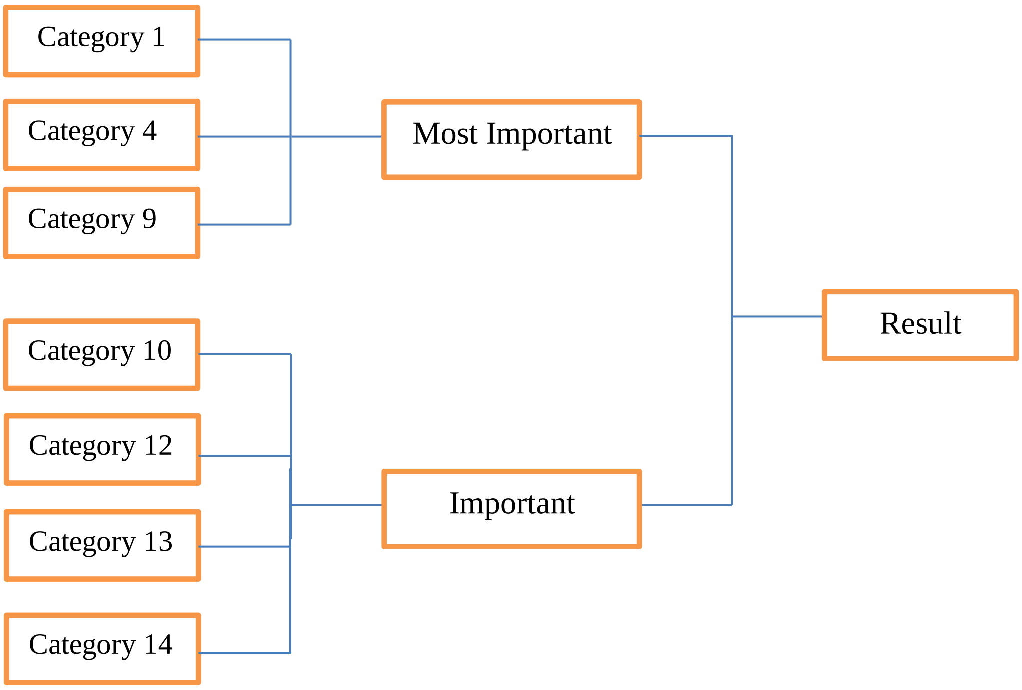

Figure 28 – The MATLAB model scheme

The MATLAB model scheme provides structure overview of the model. The names of the seven used variables have been shortened to keep the diagram simple.

Legend:

Category 1 - Recent Development of the Financial Situation

Category 4 - Client's Perspectives

Category 9 - Results and Experience of the Management

Category 10 - Quality of Information from the Client

Category 12 - Turnover Development on Client's Accounts

Category 13 - Execution on client's accounts

Category 14 - Fulfilment of Contractual Obligations

The variables have been divided into two categories due to the fact that evaluating all seven at once would have led to a huge number of possible rule combinations. The division makes the model simpler and easier to maintain.

The client is evaluated based on the first three variables, all of which have passed the information value threshold of 0.3. The result of this evaluation then serves as an input into the last part of the model.

The remaining four variables, all of which have passed the pre-set threshold of 0.2 are evaluated in the “Important” branch of the model, and the result is then used as the second, and last, input into the last part of the model.

5.5.1 Creation of the MATLAB model

The model consists of three .fis files, and one executable - the .m file. The .m file is shown later in this sub-chapter, and all the files are put on a DVD and added to this thesis.

Fuzzy Logic Toolbox

The model has been created using the Fuzzy Logic Toolbox, which is part of the MATLAB suite (10).

The first editor available in this toolbox is called the Fuzzy inference editor (FIS). FIS displays general information about a fuzzy inference system and allows users to define basic characteristics of the model (the number of input and output variables etc.). In addition, this toolbox enables users to set the way of defuzzification, aggregation etc.

Figure 29 below demonstrates that the “Most Important” branch of the model has three inputs and one output.

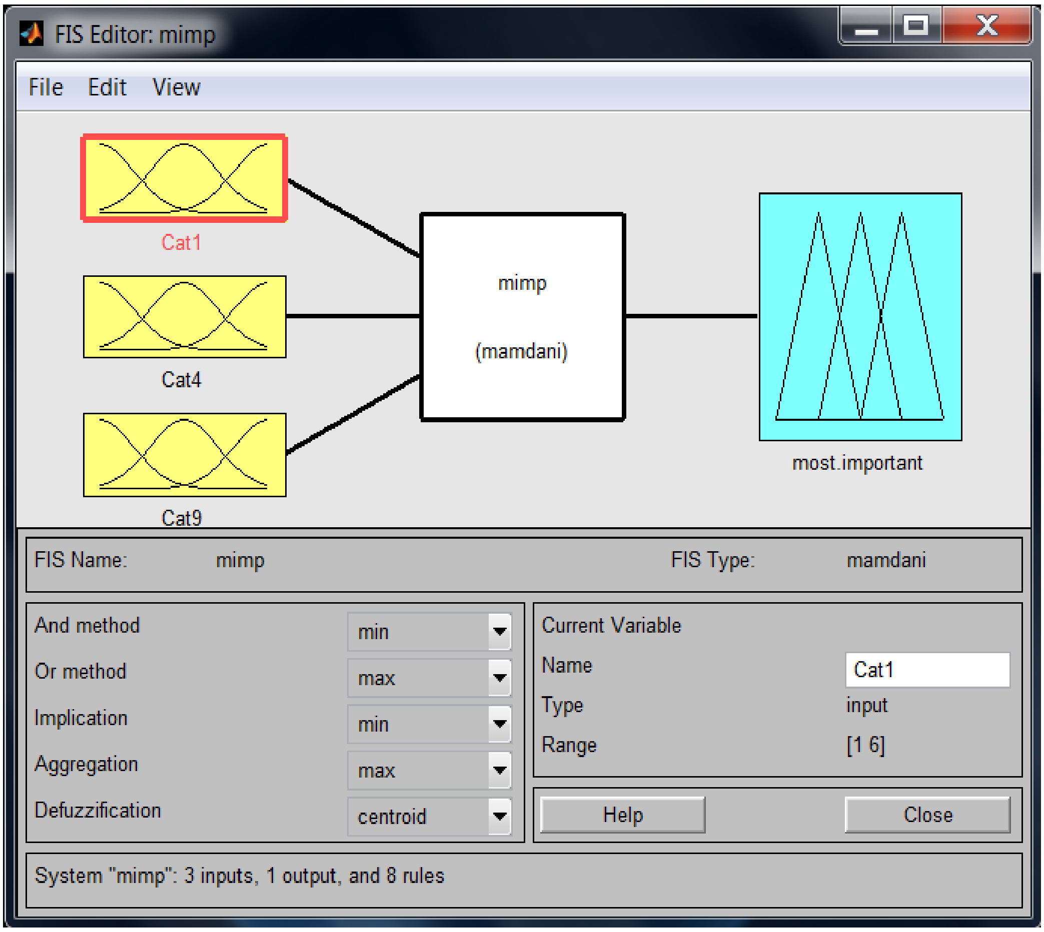

Figure 29 – Fuzzy inference editor

Membership Functions Editor

Membership Functions Editor (MFE) is part of the Fuzzy Logic Toolbox. It can be run directly from the FIS editor. MFE enables users to set characteristics for the inputs and outputs. The number of functions for each output and input is set, and each function has settings such as range, curve type, name, and parameters. To demonstrate, the functions for Category 1 (Recent Development of the Financial Situation) is shown below.

The parameters and range of the functions shown in Figure 30 below are based on data analysis of the data sample. However, the optimal curve was found thanks to the trial-and-error approach, i.e. extensive testing of the MATLAB model.

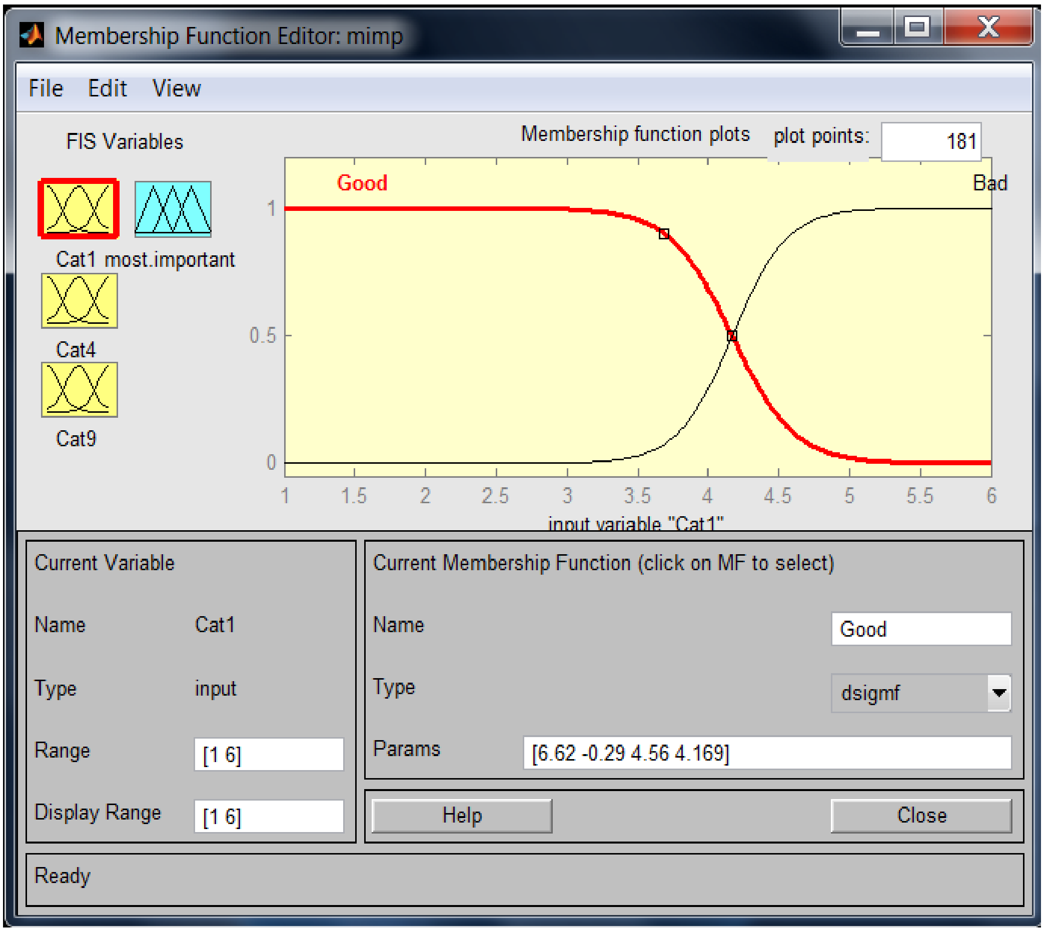

Figure 30 – MFE overview: input variable

The functions for output variables behave in the same way.

Rule Editor

As the name implies, this editor allows users to create fuzzy rules, which govern the whole model. The rules are created by combining individual criterions with AND and OR operands. A weight can be assigned to each and every rule. The weight is 1 by default.

Example of a rule that is actually used in the model:

IF **important** *is Bad* AND **most.important** *is BAD;* THEN **evaluation** *is BAD (default)*

The same rule can be written in the following format: *3 3, 3 (1) : 1*

The figure below shows a screen of Rule Editor, the aforementioned rule is number 7 on the list of rules.

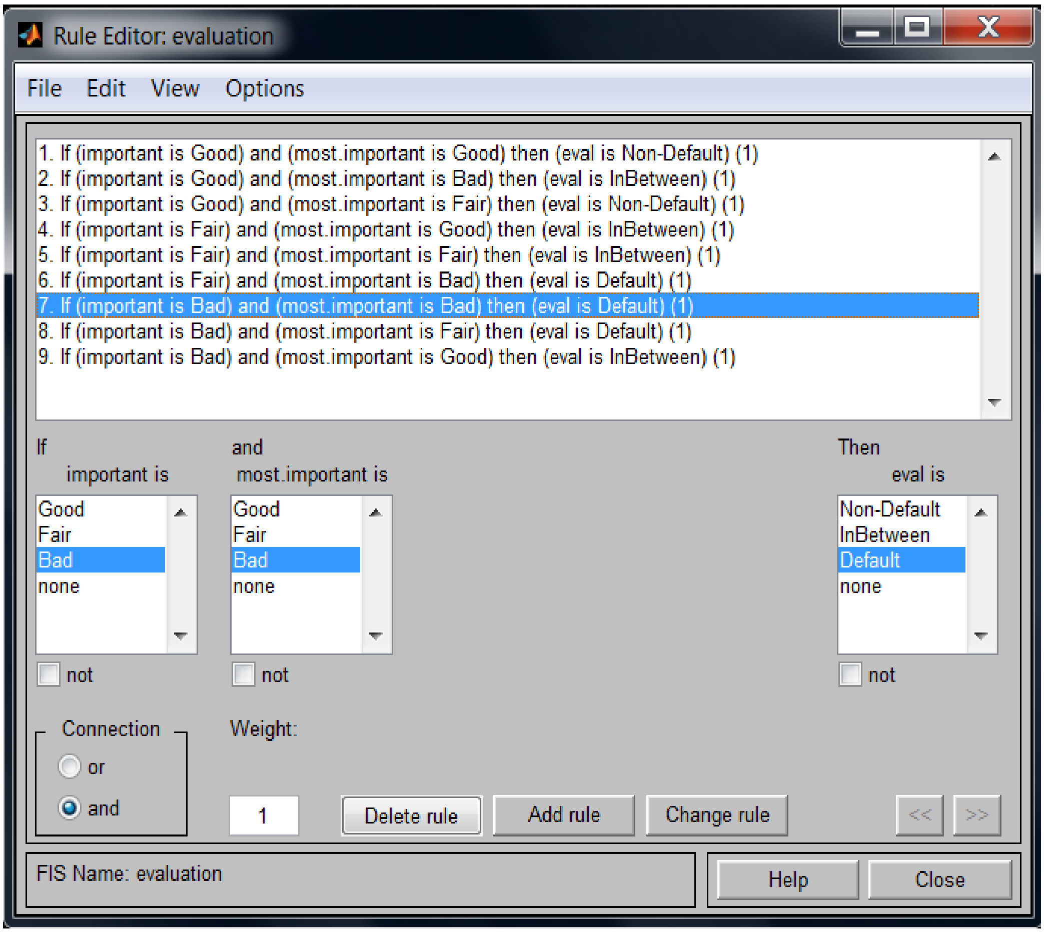

Figure 31 – Rule Editor

Rule Viewer

The Rule Viewer provides easy-to-understand overview of all rules within the model. It contains all rules and all input and output variables. Additionally, this editor is a great tool for debugging and optimization. The vertical red lines indicate the values of the input variables. They may be dragged as needed or it is possible to change the input values by simply typing the values in the box below the list of rules.

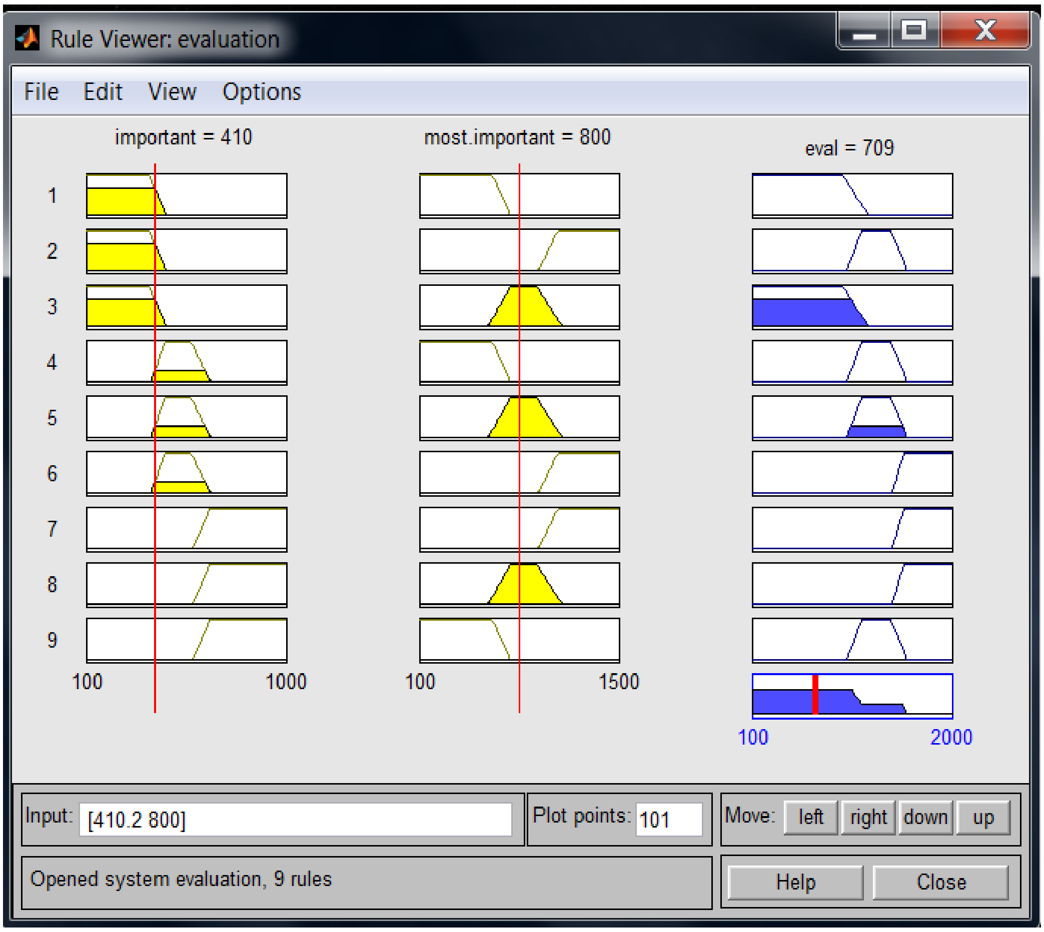

Figure 32 - Rule Viewer

Surface Viewer

Surface Viewer is capable of displaying the final function of two variables in 3D. The two variables (X and Y) can be chosen from the menu. Consequently, it is possible to select any combination of input variables.

Surface Viewer can also be used to check the correctness of the model. The surface area should span across the entire block – both horizontally and vertically.

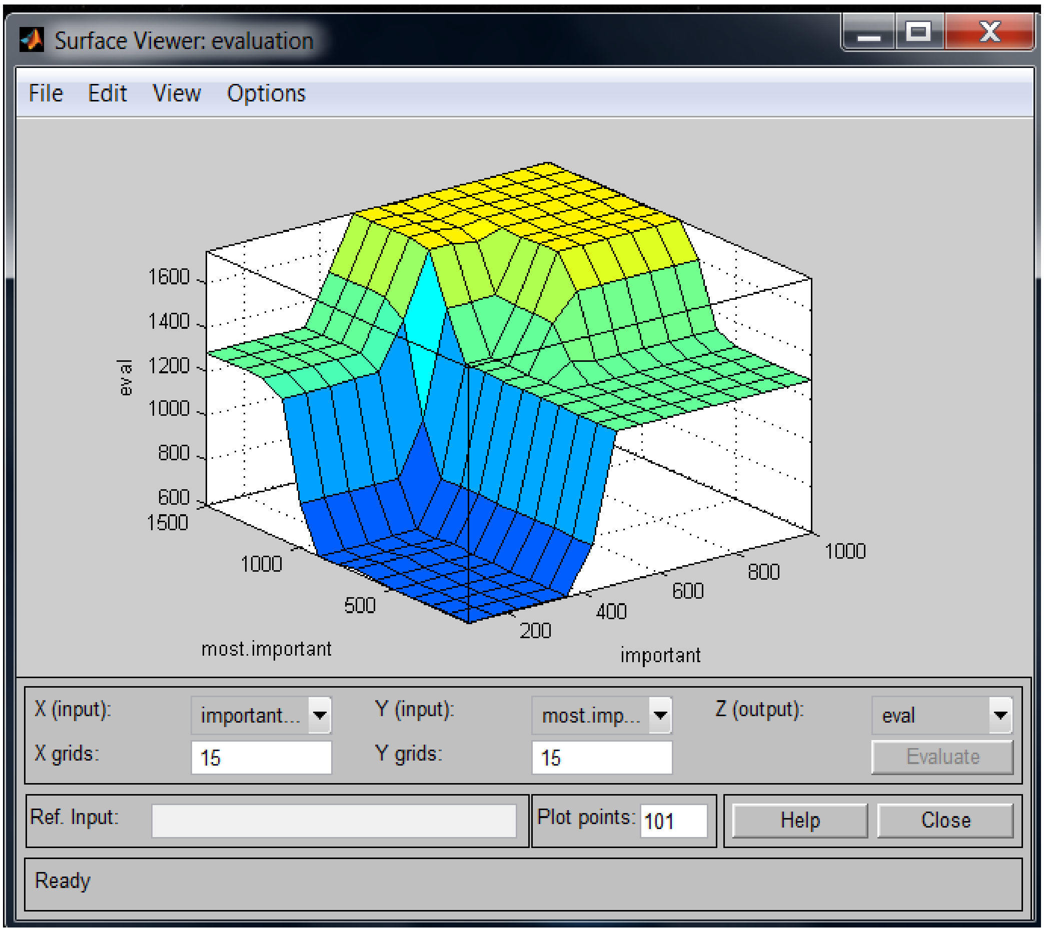

Figure 33 – Surface Viewer

The executable .m file

This file is responsible for processing the input data, i.e. the seven input variables, which are given by the user in two batches. The first batch contains the first three variables, and the second batch contains the remaining four variables.

```
clear all

MIeval=readfis('mimp.fis');
MI=input('Please input Categories 1, 4, and 9 in the following
way (the brackets must be included): [Cat1 Cat4 Cat9] :');
most.important=evalfis(MI,MIeval);

Ieval=readfis('imp.fis');
I=input('Please input Categories 10, 12, 13, and 14 in the
following way (the brackets must be included): [Cat10 Cat12
Cat13 Cat 14] :');
important=evalfis(I,Ieval);
```

```
result=readfis('evaluation.fis');

V=[important most.important];

eval=evalfis(V,result);

'eval = Risk points awarded (out of 1750):'
eval
if eval<1300 'non-default'
else 'default'
end

surfview(result)
ruleview(result)
```

The code implies that the maximum number of risk points that can be given to a client is 1750. The threshold for marking a client as default is set to 1300 points (i.e. 74.3% of the maximum value). The executable .m file is called fuzzyML.

<u>Example of use:</u>

>>fuzzyML

>>[2 2 3]

>>[3 2 2 2]

For a more detailed guide, please consult **Appendix 2**.

5.5.2 Evaluation of the MATLAB model

The MATLAB model has been evaluated on a sample consisting of 162 client instances. The sample contained 102 non-default and 60 default clients.

The result of the model's evaluation is as follows:

Out of 102 non-default clients, the model identified 76 correctly. This gives a success rate of non-default client detection of 74.5%. Consequently, 26 non-default clients have been marked as default, which results in a 25.5% Error II rate.

Out of the 60 default clients, the model correctly detected 54, for a 90% success rate of default client detection. As a result, the Error I rate is 10%.

The Total Success Rate of the MATLAB model is 80.2%, i.e. 130 out of 162 correctly identified clients.

This evaluation has been done on relatively small sample size due to the way data can be sent to input of the model. Despite that, the overall performance of the MATLAB model is considered to be satisfactory. The high total success rate is not impaired by high error rates, as is the case in the Excel model. Although it should be stressed, that the sample size used for evaluating this model is much smaller than the sample size used for the Excel fuzzy model evaluation, which might have skewed the results in MATLAB model's favour.

The other advantages of the MATLAB model stem from the rich features of the MATLAB suite itself. This model offers advanced graphic output capabilities and advanced modelling features. This is at the expense of requirement of the license to the MATLAB suite itself.

6 Conclusion and recommendations

The main aim of this thesis was to create a fuzzy model for detecting default SME clients based on the non-financial variables, and to confirm the appropriateness of applying fuzzy logic in the SME sector of banking.

This thesis has developed two fuzzy models based on fundamental principles of fuzzy logic.

The first model, which is created in Microsoft Excel, is based on a thorough analysis of the data sample provided by the bank. The data analysis was conducted in order to identify variables with high predictive power. The calculation of information value of each variable was used for this purpose. Consequently, 7 (out of 16) non-financial variables were identified, and then processed with the aim of use in the transformation matrix of the fuzzy Excel model.

The transformation matrix was created and used on the entire data sample. The result of this application depends entirely on the retransformation matrix, i.e. the default threshold. The first model achieved default client detection rate of up to 88%, albeit at the expense of a very high (96%) error II rate, i.e. identifying a non-default client as default. Whether that is a problem for the bank or not comes down to preferences of the bank. The bank might prefer a false alarm rather than undetected default client. For that reason, this thesis has put forward a number of retransformation matrices (default thresholds) and the results they yielded.

The high error rate present in the first fuzzy model may be caused by omission of a variable with a very high predictive power from the data sample. Should such a variable be missing, the error rate would be increased significantly. This may be the case with the data sample provided by the bank, albeit it can't be conclusively proven either way.

It is also worth pointing out that the fuzzy Excel model does not, and cannot, account for client exceptions – a personal contact is necessary in such cases. This fact results in the necessity of usage of appropriate data, otherwise unnecessary errors naturally occur.

Another finding worth mentioning is the fact that vast majority of the clients have had an execution on their accounts, which implies poor financial situation, and thus diminishes the overall importance of the non-financial variables.

Considering the relatively limited scope of the data sample, complete omission of financial variables, and unpredictable, exceptional events that might have happened and influenced the final state of the client (i.e. default or non-default state), the overall performance of the Excel model is deemed satisfactory.

The second model, created in MATLAB, is also built on the fundamentals of fuzzy logic, and is reliant on the data analysis, albeit not to the same degree as the Excel model.

The MATLAB model yielded higher default detection rate (90%) and a lower error II rate (24.5%) than the Excel model, but the sample size used for evaluation of this second model does not compare to the size of the data sample used for evaluating the Excel model. This is due to the way the MATLAB model accepts input variables and time constraints. The relatively small sample size might have led to skewed results.

The MATLAB model also holds advantage in the features department, boasting with advanced graphic outputs and modelling. However, this comes at a cost – the MATLAB model is harder to maintain and adjust when needed, at least compared to the Excel model, and while Excel (or its equivalent) is part of any standard Office pack, MATLAB is a highly specialized software suite needed for running the fuzzy MATLAB model presented in this thesis.

In conclusion, even though the MATLAB model offers a better performance and advanced features, the Excel model is recommended for daily use by SME bankers, mainly as a tool for rough estimate of risk level of a SME client. By contrast, the MATLAB model would be of use in a higher-tier department of the bank, serving as a secondary tool for determining and adjusting importance (i.e. weighting distribution) of the non-financial variables.

Based on the performance of both fuzzy models, the appropriateness of application of fuzzy logic in the process of risk evaluation (default detection) of corporate clients can be confirmed.

It is recommended, that both models are kept up to date by processing additional client data (for the Excel model) and adjusting fuzzy rules and functions (for the MATLAB model). Consequently, re-evaluation of both models is in order once new data becomes available.

7 References

[1] JAMIDISHI, M., VADIEE, N., ROSS, T.J. *Fuzzy Logic and Control*. Englewood Cliffs, NJ, 1993. 397 p. ISBN 0-13-334251-4

[2] DOSTÁL, P. *Pokročilé metody analýz a modelování v podnikatelství a veřejné správě*. Brno : AKADEMICKÉ NAKLADATELSTVÍ CERM, 2008. 344 p. ISBN 978-80-7204-605-8.

[3] DOSTÁL, P., SOJKA, Z. *Financial Risk Management*. Zlín : Univerzita Tomáše Bati ve Zlíně, 2008. ISBN 978-80-7318-772-9.

[4] KOČENDA, E., VOJTEK, M. *CESifo Working Paper No. 2862* [online]. 2009. [quoted on 2012-07-18]. Default Predictors and Credit Scoring Models for Retail Banking. Available at: <http://www.cesifo-group.de/DocDL/cesifo1_wp2862.pdf>.

[5] BESSIS, J. *Risk Management in Banking*. 3rd edition. Wiltshare: John Willey & Sons, 2010. 822 p. ISBN 978-0-470-01912-2.

[6] AZIZ, A. M. and DAR, H. A. *Predicting Corporate Bankruptcy: Whither do We Stand?* [online]. Loughborough: Loughborough University, 2004. [quoted on 26-07-2011]. Available at: <https://dspace.lboro.ac.uk/dspacejspui/bitstream/2134/325/3/DepartmentalPaper_AzizandDar_.pdf>.

[7] LEHMANN, B. *Is it worth the while? The relevance of qualitative information in credit rating.* Working Paper presented at the EFMA 2003 Meetings, Helsinki, 2003.

[8] ALTMAN, E. and SABATO, G. *Modeling Credit Risk for SMEs: Evidence from the US Market*. [online] 2005. [quoted on 28-07-2011]. Available at: < http://papers.ssrn.com/sol3/papers.cfm?abstract_id=872336>.

[9] HENDL, J. *Kvalitativní výzkum : Základní metody a aplikace*. 1st edition. Praha: Portál, 2005. 408 p. ISBN 80-7367-040-2.

[10] THE MATHWORKS. MATLAB – Fuzzy Logic Toolbox – User's Guide. The MathWorks, Inc., 2008.

8 List of figures

9 List of tables

10 List of appendices

Appendix I – List of the formulas used in the Excel model

Appendix II – Guide to using the MATLAB model

11 Appendices

11.1 Appendix 1 – List of the formulas used in the Excel model

Category 1 – Recent Development of the Financial Situation

=KDYŽ(B5=2;"4";KDYŽ(B5=3;"14";KDYŽ(B5=4;"216";KDYŽ(B5=5;"113";KDYŽ(B5=6;"95";KDYŽ(B5=7;"62"))))))

Category 4 – Client's Perspectives

=KDYŽ(E5=1;"1";KDYŽ(E5=2;"1";KDYŽ(E5=3;"37";KDYŽ(E5=4;"183";KDYŽ(E5=5;"104";KDYŽ(E5=6;"12"))))))

Category 9 - Results and Experience of the Management

=KDYŽ(J5=3;"32";KDYŽ(J5=4;"151 ";KDYŽ(J5=5;"5")))

Category 10 - Quality of Information from the Client

=KDYŽ(K5=1;"2";KDYŽ(K5=2;"2";KDYŽ(K5=3;"99";KDYŽ(K5=4;"83";KDYŽ(K5=5;"10")))))

Category 12 - Turnover Development on Client's Accounts

=KDYŽ(M5=2;"1";KDYŽ(M5=3;"118";KDYŽ(M5=4;"24";KDYŽ(M5=5;"6";KDYŽ(M5=6;"2")))))

Category 13 - Execution on client's accounts

=KDYŽ(N5=1;"0";KDYŽ(N5=2;"0";KDYŽ(N5=3;"3";KDYŽ(N5=4;"12";KDYŽ(N5=5;"4";KDYŽ(N5=6;"99"))))))

Category 14 - Fulfilment of Contractual Obligations

=KDYŽ(O5=1;"2";KDYŽ(O5=2;"1";KDYŽ(O5=3;"95";KDYŽ(O5=4;"134";KDYŽ(O5=5;"31")))))

11.2 Appendix 2 – Guide for using the MATLAB model

Guide for using the MATLAB model created in this thesis.

The model accepts only numerical values presented in the original data sample, which was provided by the bank.

The program can be executed from the Command Window of MATLAB by simply typing “fuzzyML” (without the quotes). Once the program is running, it will ask for the ratings of the first three variables – variables 1, 4, and 9.

The input for these variables must be entered in the following form (including the brackets and spaces):
[Category1 Category4 Category9]
-hit ENTER

In the next step, the program will ask for input of the four remaining variables – variables 10, 12, 13, and 14.

The input for these variables must be entered in the following form (including the brackets and spaces):
[Category10 Category12 Category13 Category14]
-hit ENTER

The program will display the result of the risk evaluation.

Since this model runs in the MATLAB’s internal environment, and thus possesses no risk to the operating system of the user, the model is not concerned with invalid inputs and other potential user-created problems.