Vysoká škola ekonomická v Praze

Fakulta informatiky a statistiky

Katedra ekonometrie

Data envelopment analysis as an alternative approach to managing risks in banking

DOKTORSKÁ DISERTAČNÍ PRÁCE

Doktorand : Zuzana Fialová

Školitel : prof. Ing. Josef Jablonský, CSc.

Obor : Ekonometrie a operační výzkum

Praha, 2014

Prohlášení

Prohlašuji, že jsem disertační práci na téma „Data envelopment analysis as an alternative approach to managing risks in banking“ zpracovala samostatně a že jsem uvedla všechny použité prameny a literaturu, ze které jsem čerpala.

V Praze dne

…….

podpis

ABSTRACT

Title:	Data envelopment analysis as an alternative approach to managing risks in banking
Author:	Zuzana Fialova
Department:	Department of Econometrics
Supervisor:	prof. Ing. Josef Jablonský, CSc.

The implementation of the Basel II capital adequacy framework promoted internally modelled risk parameters and allowed banks to build their own models. The recent crisis pointed at the gaps in the Basel II Accord, seeing banks having trouble to deal with lack of liquidity and higher default rates. The minimum regulatory capital held by the banks turned out to be insufficient and banks started looking for other techniques to better quantify the risks they are exposed to. Model accuracy is a key objective to meet the capital adequacy requirements while facing severe economic conditions. The purpose of this thesis is to suggest a new approach to credit modelling. Data envelopment analysis (DEA) can overcome some the difficulties that the banks deal with. The key opportunity in using DEA and its modifications is in the fact that this method does not require prior information about the classification between good and bad units and only requires financial and other data about the client in question. This thesis analyses the performance of DEA applied on a real world portfolio of corporate loans compared to the two standard methods used in the banking sector. Logistic regression is the most popular method, having few restrictions and providing output in the form of a probability of default. The second method is the discriminant analysis giving similar results to the logistic regression but being based on more assumptions. The model is validated by comparing the model output with the actual status and its predictive power evaluated.

Key words: Credit modelling, risk management, data envelopment analysis

ABSTRAKT

Název:	Alternativní přístup k řízení rizik v bankovnictví za použití analýzy obalu dat
Autor:	Zuzana Fialova
Katedra:	Ekonometrie
Vedoucí práce:	prof. Ing. Josef Jablonský, CSc.

Zavedení pravidel kapitálové přiměřenosti, v podobě druhé z Basilejských dohod, umožnilo bankám odhadovat míru vlastních rizik, a spolu s tím podpořilo tvorbu vlastních interních modelů rizikových parametrů. Nedávná krize ukázala na mezery této druhé dohody: banky potýkající se s nedostatkem likvidity a vyšším podílem nesplacených pohledávek. Minimální výše regulatorního kapitálu banky se ukázala být nedostatečná a banky začaly hledat jiné techniky pro lepší kvantifikaci rizik, kterým jsou vystaveny. Právě přesnost těchto modelů je základem pro správný odhad úrovně kapitálu a zároveň čelí vážným ekonomickým podmínkám. Cílem této práce je navrhnout nový přístup k modelování úvěrového rizika a to pomocí metody analýzy obalu dat (z angl.. Data Envelopment Analysis, dále jen DEA), která nabízí nové možnosti, jak se vyrovnat s problémy bank při odhadování svých rizik. Zásadní předností používání DEA a jejích modifikací spočívá v tom, že tato metoda nevyžaduje předchozí informace o rozdělení portfolia na dobré a špatné jednotky a vyžaduje pouze finanční a jiné údaje o dotyčném žadateli o úvěr. Tato práce analyzuje výsledky použití metody DEA na konkrétním portfoliu podnikových úvěrů a srovnává je s dvěma obecně používanými metodami v bankovním sektoru. Logistická regrese je nejvíce populární metodou s jen malým počtem omezení a poskytuje výstup ve formě pravděpodobnost selhání. Druhou metodou je diskriminační analýza, která se svými výsledky podobá logistické regresi, ale je založena na více předpokladech a podmínkách. Schopnost modelu předpovídat rizika je vždy ověřeno nejen statisticky, ale i porovnáním výstupů modelu se skutečným stavem.

Klíčová slova: Modelování kreditního rizika, řízení rizika, analýza obalu dat

CONTENTS

INTRODUCTION

The world has encountered a deep dive into the economic recession, seeing banking system falling, when people stopped repaying their debts; investments losing their value and lack of activity in the credit market. The difficulties in the banking system forced the banking authorities, such as the Bank of International Settlements (BIS), the World Bank or the International Monetary Fund (IMF), to review the banking regulations.

To enhance supervision of the banking sector, in 1988, the Bank of International Settlements (BIS), located in Switzerland, published an agreement to regulate the minimum capital to be held by banks. This is also known as the first of the Basel Accords. The Basel I Accord requested institutions to hold eight % of their exposure, whatever the riskiness of the customer was. The works on Basel II started in 1999 and the transition from Basel I to Basel II took place in 2006 with mutual runs. A full implementation was required from 1st January 2007.

The introduction of the Basel II Capital Accord has encouraged financial institutions to build their own internal rating systems assessing the credit risk of their various credit portfolios. Compared to the first one, Basel II looks at the economic conditions, the riskiness of the borrowers and it allows banks to tailor their models to their specific portfolios. Despite that fact, the experience with recent crisis pointed at the gaps in the second Framework. Some people started to question the usefulness of the Basel Framework and even made it responsible for the size of the crisis itself (Cannata & Quagliariello, 2009). Banks were forced to adopt stronger risk management practices and strategies and restructure them to be able to survive.

Credit scoring models were identified as one of the major elements to be revised. This thesis provides an insight into the development of a credit scoring model, looking at the various stages of the development and showing how alternative approaches can bring the extra performance required. Three methods will be applied on a real world portfolio of a European bank. The portfolio contains different sizes of companies, small and medium enterprises (SME), as well as public and large enterprises (PLE).

The logistic regression, sometimes called the logit model, is used for prediction of the probability of occurrence of an event by fitting data to a logistic curve. This method will be applied to a portfolio of 6759 counterparties. The model will be finally validated through comparison to the actual status and the performance of the model assessed. Same dataset will be used to assess the performance of the second most popular method, discriminant analysis. An alternative approach will finally be presented as on opportunity for banks and other credit providers to identify and correctly quantify the risks they face.

This thesis is limited by the data quality issues. The banks all over the world fight with poor data quality. The reason is simple. The poor data quality results in losses. With databases counting hundreds of thousands of records or millions, even a small percentage of data quality issues mean thousands of records being wrong. There is an escalation of inaccuracy. The reports of many departments are based on the incorrect data. The reports of other departments use inaccurate reports as a base for their reports. Departments handling the wrong data are unable to produce accurate reports that the higher management of the bank could draw accurate conclusions from. There are a number of different sources of the poor data quality. In every step of the data handling, there is space for data quality issues. It can be in the collection of data, transfers, modifications, storage. Given the huge number of records involved in the experiment, statistical software handling big datasets is needed for the purpose of data storage, manipulation, calculations and building models. These software products usually come with reporting tools that can be used to present the results of any analysis.

This thesis is divided into several chapters. The first part of the thesis talks about various types of risk that the banking system is exposed to and the importance of managing those risks. No strategic decisions can be done without having assessed all types of risks the bank has to deal with. Badly estimated risks lead to inaccurate credit models and in an extreme case to insolvency of the bank. The second part presents the domain of credit scoring and the methodologies used. The third part of the thesis describes the data collection, sanity check, modifications, selection, statistics and summary. No analysis can be done without a prior look at the data set to be used for applications of any methods. Two standard methods are applied on the selected portfolio in order to get a benchmark of the model performance. The new approach is applied on the same data set and results compared to the benchmark given by the results of the standard credit scoring methods.

The purpose of this thesis is to identify the areas where the use of the DEA models could contribute to process improvement and help the risk managers to minimize risks the banks face. The idea is to suggest an alternative way of managing risks and show that using the standard methods is a question of comfort and innovation is always a benefit. The Data envelopment analysis can offer new and innovative ways of managing risks. The key opportunity in using the DEA and its modifications is in the fact that this method does not require prior information about the classification between good and bad units and only requires financial and other data about the client in question. The final part of this thesis assesses the possibility of application of the two stage network DEA to the credit scoring models. Network DEA is a fairly new term and the aim is to show its potential use and advantages compared not only to the standard methods but as well to the basic DEA models.

The demonstration of the potential use of the DEA as a new approach to the credit risk management practices is shown through:

- Application of the DEA methodologies on a corporate portfolio of 6759 counterparties and assessing the performance of the model,

- Validation of the results through comparison with the observed data,
- Comparison of the performance of the credit scoring model based on DEA to the benchmark given by the results of the standard methods of logistic regression and discriminant analysis,
- Development of all models in SAS to ensure automation and flexibility. The data contains thousands of units whose credit risk needs to be assessed on a regular basis and with a number of model versions. SAS has no limitations on the number of variables or DMUs and the macro can be easily modified by simple code change. Furthermore, the macros can be run automatically by setting the schedule in SAS. The results can feed directly into a spread sheet and get updated without the need to open SAS Interface.

1 LITERATURE REVIEW

Bernstein covered the whole history of risk, covering five periods from before 1200 up to present. His findings show that risk is not a term unknown to our predecessors (Bernstein, 1996). The first book on probability and games dates back to 16th century. Over time, names, such as Cardano, Pascal, Fermat, Graunt, the Bernoullis, De Moivre, Bayes, Laplace, Galton, Keynes, von Neumann, Baumol, Knight, Markowitz, Leland, Rubinstein or Thaler, contributed to the risk theory and have become well known authors in the risk field. It is a mystery why it took so long to develop the mathematics of chance. The first workings on this subject dates back to 16th century, when Girolamo Cardano published his book Liber de Ludo Alaea.

Mathematical modelling of finance theory started with the works of Markowitz in the early 1950s (Markowitz, 1952). The efforts of the decade after were directed towards evaluation of firms for the purpose of acquisitions and mergers, optimization of a company's finance mix and using investment portfolios to manage risks.

Beaver was a pioneer in business failure prediction research. He conducted an analysis of likelihood ratios based on a Bayesian approach. He argued that the default prediction problem could be regarded as a problem of evaluating the probability of financial distress conditional upon the value of a specific financial ratio (Beaver, 1966). Financial ratios are the simplest tools for evaluating and predicting the financial performance of firms. They have been used in the literature for many decades. The main advantage was clearly their simplicity. For each ratio, one simply compares the firm's value against a set cut-off point and decides on the classification accordingly. In response to Beaver, (Tamari, 1966) suggested that an analyst cannot merely rely on one ratio and made an attempt to weight ratios arbitrarily. He introduced the so called risk index where points are assigned to the firm based on the value of its ratios. The approach of the financial ratios (also called univariate statistical approach), gave rise to the methods for business failure prediction based on the multivariate statistical analysis. Shortly afterwards (Altman, 1968) proposed to use the discriminant analysis based on five financial ratios for business failure prediction. His model took the name of the Z-score where Z is the index created by a linear combination of the 5 ratios.

The work of (Altman, 1968) has encouraged others to come up with new methods dealing with limitations of the discriminant analysis. Among these, Ohlson presented empirical results of his study predicting corporate failure (Ohlson, 1980). He used data from the seventies of about 2 thousand firms and concluded that 4 basic factors affect the probability of failure: the size of the company, a measure of the financial structure, a measure of performance and a measure of current liquidity. Another work focused on the limitations of discriminant analysis was presented by (Eisenbeis, 1978). In his paper, Eisenbeis reviews different techniques and models described in various journals and points out that the statistical scoring models discussed in the literature have focused primarily on the minimization of default rates leaving behind the objective of a scoring model, which is the lender profit maximization or cost. The

paper also shows that, even ignoring these shortcomings, the models used typically suffer from statistical deficiencies. After 1980, the use of the multiple discriminant analysis has decreased, but still is by far the dominant classical statistical method, followed by logit analysis (Altman & Saunders, 1998). Linear probability and multivariate conditional probability models (Logit and Probit) were introduced to the business failure prediction literature in late 1970s.

The extensive research in this area is due to the consequences of business failures. It does not only influence individuals, but it has an impact on the whole society. This fact encouraged the G-10 central banks to apply common minimum capital standards to their banking industries, under the name of the 1988 Basel Accord (Basel Committee of Banking Supervision, 1988). The standards are almost entirely addressed to credit risk, the main risk incurred by banks. The bank's assets are grouped into 5 categories and assigned a certain risk weight based on the category. Revision of the Basel I regulations was supported by the analysis of (Altman & Saunders, 2001). Their critique pointed at the lack of sufficient degree of granularity when using external agency ratings.

Managing credit risk involves assessment of the client's ability to repay the debt in terms of probabilities. Credit scoring brought what managers needed to measure the risk.
The first traces of credit scoring date back to 1930s in USA. In 1940 (Plummer & Young, 1940) published a work on credit practices used at that time. Still only two elements were taken into account, past experience and intuition. David Durand (Durand, 1941) was the first to publish a study on how to differentiate between good and bad loans. He was working on a research project for the US National Bureau of Economic Research and based his ideas on Fisher's discriminant analysis (Fisher, 1936) who introduced the idea of discriminating between groups in a population (of plants). Durand's study included 7200 loans with 37 banks and finance companies. As part of his study was a survey on the credit factors indicative of good risk. He identified two major factors, the applicant's moral character, which is judged by his past payment record as well as by his general reputation and the stability of his employment, which is the criterion of the performance of his earning power.

In 1956, Bill Fair and Earl Isaac founded a credit scoring and business consulting firm named Fair, Isaac and Company with the intention to help financial services companies make complex and accurate business decisions. Their credit scoring model is known as FICO after their initials, named after the FICO score that represents the creditworthiness of a client. The FICO credit score is calculated statistically, measuring the risk of default by taking into account various factors based on the information from an applicant's credit files. As there is more than one factor, these need to be weighted. The factors included in the FICO score are the following: payment history, credit utilization, length of credit history, types of credit used, recent searches for credit. The higher is the score, the higher is the chance of the consumer to pay back.
Myers & Forgy confirmed that in consumer lending, statistical approach to credit scoring represents an improvement over the judgmental-intuitive evaluation of credit risk. As well, their results showed that equal weights for all significantly predictive items were as effective

as weights from the more sophisticated discriminant analysis or stepwise multiple regression (Myers & Forgy, 1963). Several other studies have been done in this area and it can be concluded that by the mid-1960s, the conceptual framework of modern consumer credit scoring was developed. At that time, credit scoring was the exclusive domain of consumer credit. Corporate lending followed (with the work of (Beaver, 1966), see above) when academics and practitioners started to consider statistical techniques to replace or enhance the traditional ratio analysis in evaluating the health of the companies.

Over the years, a lot of research has been done in consumer and corporate lending credit scoring. Despite the modern credit practices and the Basel II agreement that has been designed to avoid collapse of banking industry, the crisis hitting the world in 2008 has shown the opposite. The second Basel Accord has been reviewed and seen further changes in the form of a third version and banks themselves reviewed their risk management approaches to find gaps and weaknesses.

Ian Brown, Analytics Specialist at SAS[1], highlighted five key challenges in credit modelling (Brown, 2012) that can lead to performing models and therefore to a lower risk. These are:

a) A clear and consistent database that is crucial for model development and validation,
b) Low default portfolio making accurate predictions are difficult, and methods such as under sampling or oversampling need to be considered,
c) Accurate models – the more accurate model, the lower the risk,
d) Reject inference methodologies applied on a portfolio of rejected applicants[2],
e) Forward-looking indicators, such as the deviations from a trend for the ratio of domestic credit to GDP, are still an area which requires more development for rating models.

Brown suggests that with the right amount of knowledge and openness to try new ideas, financial institutions could potentially take the benefits of applying novel analytical techniques; such as neural networks (Angelini, Tollo, & Roli, 2008) (Lee & Chen, 2005), support vector machines (Huang, Chen, & Wang, 2007), genetic programming (Ong, Huang, & Tzeng, 2005), decision trees (Yobas, Crook, & Ross, 2000) and recently the Data envelopment analysis that is being discussed in this paper.

Data Envelopment Analysis was first developed for use in evaluating activities of not-for-profit entities participating in public programs (Charnes, Cooper, & Rhodes, 1978). Over the years, it has been applied not only on public but as well as on private entities, such as hospitals, universities, manufactures and banks. The two basic models of the DEA, the CCR

[1] SAS is a powerful business analytics software that finds its use in many organizations, especially financial institutions, telecommunications and other, where there is need for storing huge amount of records, data mining, analysis, statistics, calculations, graphics and much more (SAS Institute Inc.)
[2] More details on Reject Inference and the methodologies can be found in (Crook & Banasik, 2004).

model and the BCC model, have been adjusted and enlarged to better fit the real life problems.

First study attempting to use risk-adjusted efficiency measures was published by (Berg, Førsund, & Jansen, 1992). They studied the productivity growth during the deregulation of the Norwegian banking industry within the framework of Data Envelopment Analysis. Similar studies were made in the following years (Hughes, Lang, Mester, & Moon, 1996). In the late 1990s, data envelopment analysis (DEA) was introduced to the analysis of credit scoring (Troutt, Rai, & Zhang, 1996).

DEA has been suggested as an alternative tool to measure risks in various forms. Min & Lee proposed a DEA-based approach to credit scoring (Min & Lee, 2004). The empirical results were also validated by supporting analyses (regression analysis and discriminant analysis) and by testing the model's discriminatory power using actual bankruptcy cases of 103 firms. A similar study was carried out by (Feruś, 2008) on 100 construction companies that obtained a credit loan from a Polish bank in the years 2001–2003. Another DEA model applied on credit risk evaluation and bringing impressive results is the concept of worst practice DEA introduced by (Paradi, Asmild, & Simak, 2004). The idea is to identify worst performers by placing them on the frontier. And therefore recognize the largest improvement potential. Similar model was presented by (Shuai & Li, 2005). Olson & Wu recently suggested a DEA VaR model as a new tool to conduct risk management in enterprises and provide means to quantitatively improve decision making with respect to risk (Olson & Wu, 2010).

2 CREDIT RISK ASSESSMENT

2.1 Risk definition and types of risk

Before discussing the many faces of risk, we first need to understand what we mean by risk. Risk can be defined as the chance or possibility of suffering misfortune, damage, loss or other adverse consequence. By understanding why something may go wrong, we can try to prevent or reduce the possibility of it occurring, or lessen the effect should it occur. Risks are present in our everyday lives. Driving might be a good example of a life threatening risk. But there has been taken actions to reduce the likelihood of accidents occurring, such as vehicle safety features, legal speed limits, pedestrians crossing areas, etc.

Businesses are no different. In every activity that they undertake, there are risks involved, such as the possibility of incurring a loss of some sort. This may not be monetary, but it can still damage the business. Risk in general can be of any type – it can have positive or negative consequences.

Banks are exposed to a number of types of risk across five broad categories, as follows:

a) Market risk is the risk of loss due to the changes in market value of financial assets. Market risk is an institution's sensitivity to a change in economic value of its assets and liabilities (including off-balance sheet) due to market price movements which may arise in the future. Bank is exposed to market risk because of positions held in its trading portfolios and its non-trading business including the bank's treasury operations.

b) Regulatory risk arises from failure to comply with the requirements of regulators or laws in any of the countries in which the bank operates.

c) Reputational risk is the risk of damage to our reputation that may arise from an apparent failure to properly manage risks within the bank.

d) Operational risk is the risk of loss resulting from inadequate or failed internal processes, people and systems or from external events including legal risk.

e) Credit risk is the risk of loss caused by the failure of a customer to pay their obligations.

Identification, assessment, regulation and mitigation of those risks represent the aim of risk management. It can be applied throughout the life of an organization, and to a wide range of activities, including strategies and decisions, operations, processes, functions, projects, products, services and assets.

The guidelines also recommend to design and implement risk management plans and frameworks while taking into account the varying needs of a specific organization, its particular objectives, context, structure, operations, processes, functions, projects, products, services, or assets and specific practices employed (International Organization for Standardization, 2009).

The strategies to manage threats (uncertainties with negative consequences) typically include transferring the threat to another party, avoiding the threat, reducing the negative effect or probability of the threat, or even accepting some or all of the potential or actual consequences of a particular threat, and the opposites for opportunities (uncertain future states with benefits).

The major three risks banks deal with are Credit, Market and Operational Risk. The biggest portion of the total risk arises from credit.

Risk is the potentiality that both the expected and unexpected events may have an adverse impact on the bank's capital or earnings. The expected loss is on the account of the borrower through risk premium and reserves created out of the earnings, whereas the unexpected loss has to be covered by the capital. Each type of risks is measured to determine both the expected and unexpected losses, as explained in chapter 2.2.

This thesis is focused on the evaluation of the biggest of the risks in the industry of banking, credit risk.

2.2 Credit Risk Regulations

The main purpose of banking is borrowing finances in return for an interest rate. The risk that comes with borrowing is the possibility of not getting the borrowed amounts back.
Assessing the credit risk is necessary for the bank to exclude fraudulent, insolvent and other customers that will deprive the bank not only from its profit from the borrowing but the loan amount as well as other fees related to the service provided.
Credit risk having an impact not only on banks shareholders, managers, staff and the clients, but as well on suppliers, clients, competitors, and regulatory bodies, among others, has been in the centre of attention of the banking authorities, such as the Bank of International Settlements (BIS), the World Bank, the IMF and the Federal Reserve. Having followed its mission to serve central banks in their pursuit of monetary and financial stability and following on the 1970s crisis that brought the issue of regulatory supervision of internationally active banks to the fore, the BIS actions resulted in the 1988 Basel Capital Accord and its "Basel II " revision of 2001-06.

The introduction of the Basel II Capital Accord has encouraged financial institutions to build internal rating systems assessing the credit risk of their various credit portfolios. To determine capital adequacy, the bank's assets are risk weighted.

Subject to certain minimum conditions and disclosure requirements, banks that have received supervisory approval to use the IRB[3] approach may rely on their own internal estimates of risk components in determining the capital requirement for a given exposure. The risk components include measures of the probability of default (PD), loss given default (LGD), the exposure at default (EAD), and effective maturity (M). Figure 1 illustrates how variation in realized losses over time leads to a distribution of losses for a bank.

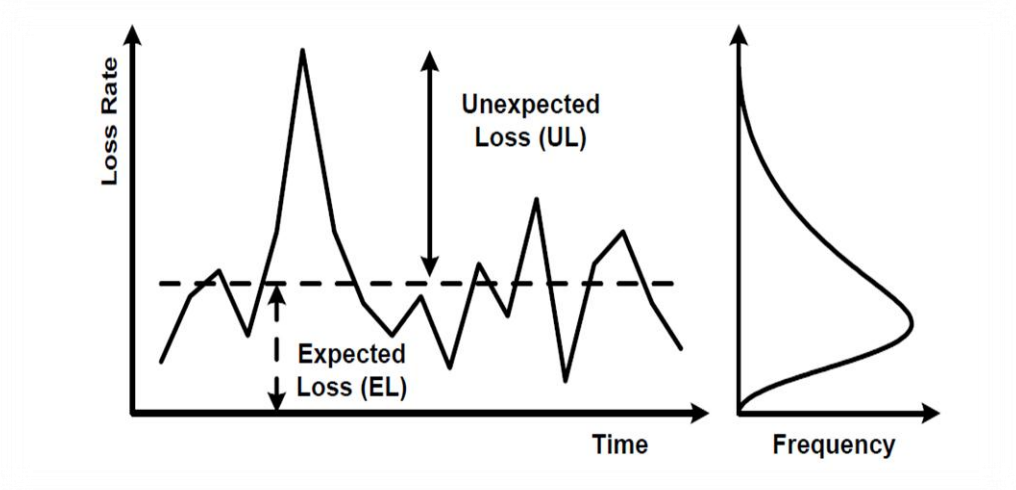

FIGURE 1: LOSS OVER TIME (SOURCE: (BASEL COMMITEE ON BANKING SUPERVISION, 2005)

While it is never possible to know in advance the losses a bank will suffer in a particular year, a bank can forecast the average level of credit losses it can reasonably expect to experience. These losses are referred to as Expected Losses (EL) and are shown in Figure 1 by the dashed line. Financial institutions view Expected Losses as a cost component of doing business, and manage them by a number of means, including through the pricing of credit exposures and through provisioning (Basel Commitee on Banking Supervision, 2005).

As mentioned above, the main credit risk parameters are PD, LGD, EAD and M and are commonly used by the risk managers to assess riskiness of the deal to be booked. Probability of Default (PD) is an estimate of the likelihood, expressed as a percentage that a customer will default during the next twelve-month period. Exposure at Default (EAD) is an estimate of the credit exposure, expressed in monetary terms that the bank would have to the customer in the event of default within the next twelve months. Loss Given Default (LGD) is an estimate, expressed as a percentage of EAD, of the amount that will be lost by the bank in the event that the customer defaults. The difference between the EAD and the net amount of the expected recovery represents the LGD. It must be based on Economic Loss which includes recovery costs and expenses and discount effects. For Basel Agreement purposes the LGD is required to be representative of losses in a downturn. Maturity (*M*) is the residual maturity of an exposure; it is the number of years remaining during which the borrower is permitted to fully repay what they have borrowed for a particular facility.

[3] IRB is an abbreviation of Internal Ratings Based - Basel II accords allowed for internally modelled ratings

Every banking institution needs to know and evaluate their losses that might occur as a negative but certain element of every banking business. The Expected Loss is an estimate measure of loss that the Bank's portfolio will encounter. Over time, cumulative EL should (roughly) equal cumulative losses. The provisions should be taken to cover the expected loss (the essence of any provision is to save money for losses you expect in the future). The curve in Figure 3 describes the likelihood of losses of a certain magnitude. The area under the entire curve is equal to 100% (i.e. it is the graph of a probability density).

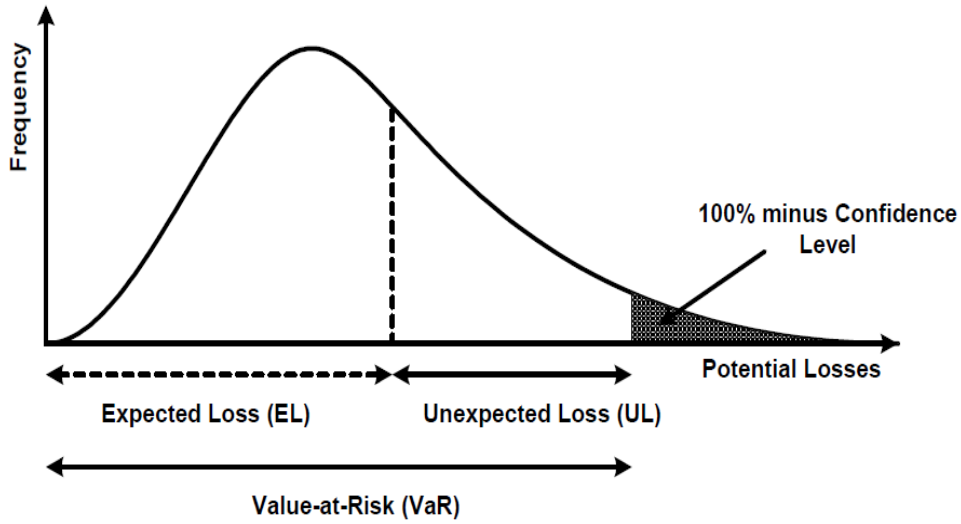

FIGURE 2: LOSS DISTRIBUTION (SOURCE: (BASEL COMMITEE ON BANKING SUPERVISION, 2004)

The curve shows that small losses around or slightly below the Expected Loss occur more frequently than large losses. The likelihood that losses will exceed the sum of Expected Loss (EL) and Unexpected Loss (UL) - i.e. the likelihood that a bank will not be able to meet its own credit obligations by its profits and capital - equals the hatched area under the right hand side of the curve. 100% minus this likelihood is called the confidence level and the corresponding threshold is called Value-at-Risk (VaR).

2.3 Credit scoring

Every legal entity is required to have a credit grade, represented with the probability of default (PD). These grades are internally calculated through built models. These models are based on an algorithm that predicts the future classification of the applicant as a good or bad credit risk using the known profile of the subject. The algorithm is derived using a multivariate analysis technique that allows identifying characteristics of the profile and respective weights of recent or current borrowers whose status as good or bad risks is known. A scoring formula, a scorecard, a model – all these are algorithms and the essence of credit

scoring. Credit scoring is a tool designed to help the bank quantify and manage the financial risk involved in providing credit.

The following diagram represents the model use, starting from the data input, the model itself, the output giving information to calculate the capital requirements and business decisions based on it.

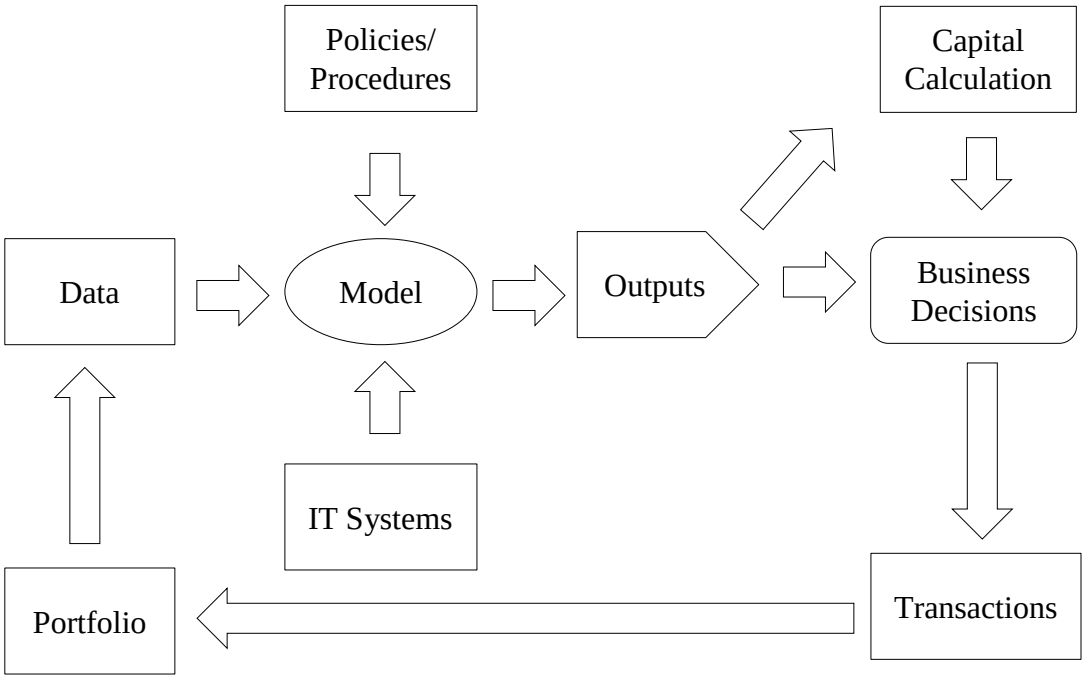

FIGURE 3: MODEL USE (SOURCE: OWN CHART)

The aim of credit scoring is not only useful to quantify the probability of default of an obligor. It can also be a tool to identify accounts that have the potential for fraud. It can provide an overall evaluation of the receivables portfolio by identifying its quality as well as the potential for bad debt write-offs and corresponding reserve requirements.

In general, we have two types of credit models. The first type is called **Application Model** and these models are built to provide credit check on applications. In other words, this is the first credit decision and grade that is assigned to the customer. These models are based on mostly financial variables that come from a credit bureau or the application forms. Credit Bureaus are currently the easiest way how to get financial and other data on customers. It allows producing quick credit checks that wouldn't slow the sale process, but still provide enough security for the vendor.

The second type is called **Behavioural Models** that are used to track the payment behaviour of the customer. The output of behaviour models is the probability that an active account will be delinquent and/or written-off and/or experience bankruptcy and/or sent to a collection agency and/or exhibit some other type of derogatory payment behaviour over a specified period of time. If the customer's behaviour is bad, the bank can lower its credit limit and

increase the PD and therefore reduce its exposure to the client or put the customer of a watch list. These models include the borrower's own payment history with the bank.

The model can be structured into a financial scorecard, which assesses the strength of an entity's financial statements, and a non-financial scorecard, which assesses the entity's account behaviour and qualitative factors. Every level of the model variables gets a score. It can have the form of a binned scorecard, where every value of the variable falls into a range. Each range then has a score assigned. The scores are then mapped to grades and PDs. Table 1 represents an example of a binned scorecard.

Variable name	Attribute	Score
Current ratio	Less than 1	-10
	1-2	10
	Other	0

TABLE 1: EXAMPLE OF A SCORECARD (SOURCE: OWN DESIGNED TABLE)

Other form of a scorecard is a simple equation that leads to the score. Each variable is multiplied by the modelled coefficient and the total results into a score.

Bad rates are usually calculated for the various ranges of scores and cut off score is chosen with an acceptable bad rate.

Some of the strategies for high risk customers are (Siddiqi, 2006):

- Declining credit
- Assigning a lower credit limit
- Asking the applicant to put a higher deposit
- Charging a higher interest rate on loan or premium on insurance
- Putting the applicant on a watch list

2.4 Model development process

Successful modelling of a complex data set is a mix of science, statistical methods and experience and common sense. A high importance is put on the development and the performance of the model to meet regulatory requirements as well as the banks strategic plan. The model development process consists of the four stages as in Figure 4. The usual practice is that a specialized models team only looks after the model development, monitoring and analysis.

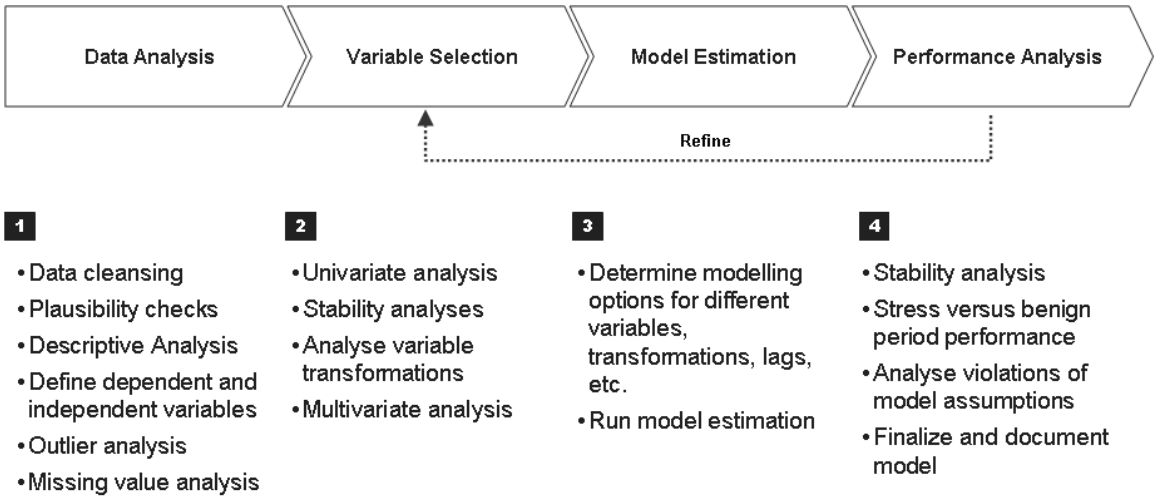

FIGURE 4: MODEL DEVELOPMENT PROCESS (SOURCE: OWN CHART)

Conceptually, the model is designed to predict the probability of a default event over a 12 month time-horizon. The explanatory data cover a range of different variables collected as part of the credit loan applications. The preparatory work includes apart from collecting relevant data, cleaning and manipulating it to ensure it is suitable for model-building. The factor analysis selects the variables that add the highest value to the model. The calibration component of the model governs how the model scores are mapped to PDs. The model is finally validated as described in Figure 5.

Different PD models are built to provide adequate credit grading depending upon the nature of the deal and customer. The traditional counterparties would be Large Corporates and Mid Corporates. Other models can be designed for industries with specific characteristics, such as Investment Property.

2.5 Model performance

The model performance is first dependent on the data and second on the business insight. It is obvious that the better model the bank develops the less capital it will need to hold. But there are several conditions, such as accurate data entering the model or policies and regulations. Even a good scoring system won't predict with certainty any individual loan's performance, but it should give a fairly accurate prediction of the likelihood that a loan applicant with certain characteristics will default.

The assumption is that future borrowers will have credit behaviour alike past borrowers with similar profiles. Statistical significance and representativeness have to be respected. Due to the fact that past borrowers had been screened by loan officers during their approval process, the population of clients with known credit risk status is biased. The profiles of rejected applicants have to be confronted with the profiles of recent good and bad clients and results considered. In the effort to reduce bias, the algorithm is developed using a sample and tested

on a hold-out sample. The discrimination power of the algorithm is measured and tested if statistically and economically significant. If acceptable, the credit scoring system is implemented and used in the process of screening loan applicants. Results of the use of the credit scoring system are regularly verified using monitoring techniques and reports.

The performance of the model is not only measured through accuracy. Another criterion is the speed of the classification, the flexibility of the model for any amendments or future changes, the ease of understanding of the classification method and why it has reached its conclusion. Classification methods which are easy to understand (such as regression or tree-based methods) are much more appealing to the users than are methods which are essentially black boxes (neural networks). Neural networks though are well suited to situations where we have a poor understanding of the data structure.

Specifically, a major impediment to model validation (or "back testing" as it is popularly known) is the small number of forecasts available with which to evaluate a model's forecast accuracy. That is, while VaR models for daily, market risk calculations generate about 250 forecasts in one year, credit risk models can generally produce only one forecast per year due to their longer planning horizons. Obviously, it would take a very long time to produce sufficient observations for reasonable tests of forecast accuracy for these models. In addition, due to the nature of credit risk data, only a limited amount of historical data on credit losses is available and certainly not enough to span several macroeconomic or credit cycles. These data limitations create a serious difficulty for users' own validation of credit risk models and for validation by third-parties, such as external auditors or bank regulators (Lopez, 2000). (Lopez, 2000) recognized the need of evaluating the performance of credit risk models and proposed evaluation methods based on cross-sectional simulation.

Diversification between the bads and goods is an important measure and indicator of model performance.

3 STANDARD METHODS USED IN CREDIT SCORING

A range of techniques have been developed for analysing firm's performance and to define the probability of its failure. From the simplest of the methods, financial ratios, through the most used methods such as linear discriminant, logistic regression and neural networks to Kernel methods using support vector machine. The credit scoring area found development opportunities in statistics, in operations research such as mathematical programming, nonlinear fuzzy mathematics, such as the neural network method or decision making approaches.

Since the PD modelling problem basically boils down to a discrimination problem (defaulter or not), one may rely on the various classification techniques, but keeping it as transparent and easy to understand as possible, since the credit risk models will be subject to supervisory review and evaluation, Hence, techniques such as neural networks or support vector machines are less suitable due to their black box nature.

Two main types of statistical models for modelling defaults are duration models and classification models. In duration models, the focus is on the time to default. Disadvantages of the duration models are that the data sets are often too limited and that the model does not provide an estimate of the PD directly, which is required by Basel II. The other main approach in modelling the probability of default is through classification models. The most popular models in this category are discriminant analysis and probability models (Duffie and Singleton, 2003). Historically, discriminant analysis and regression have been the most widely used techniques for building scorecards. Each of the methods has its own advantages and limitations.

3.1 Financial ratios

Financial ratios are the simplest method for evaluating and predicting financial performance. It is a straight forward and transparent method, but its limitations are significant: it is difficult to develop a meaningful set of industry comparatives; a firm's balance sheet doesn't always reflect the real financial situation of the firm (e.g. Inflation); seasonal factors can distort a ratio analysis; some ratios might be difficult to identify as bad or good; a firm can have bad and good ratios. (Beaver, 1966) (Altman, 1968)

3.2 Linear regression

Regression analysis aims at modelling and understanding relationships between data. Having observations $X_i (i = 1, \dots, N)$ from a so-called independent variable X, observations $Y_i (i = 1, \dots, N)$ from a so-called dependent variable Y and a parameterized function $f(\beta_0, \beta_1, \dots, \beta_d)$ $(d \geq 1)$ and supposing that a functional relationship

$$Y = F(X, \beta_0, \beta_1, \dots, \beta_d) \tag{3.2-1}$$

holds, one tries to determine the value of the parameters $\beta_0, \beta_1, \dots, \beta_d$ such that f fits the data best. The choice of the functional form f (and thus the number of parameters), as well as the optimality criterion, are quite arbitrary.

Without doubt, linear regression methods are best known and most widely used. In this case f is a linear function, and both dependent and independent variables are real-valued. Least-square estimators of the involved parameters can be easily derived in explicit form. Four principal assumptions justify the use of linear regression models, one of which is the important normality assumption. Further to normality of the error distribution, independence and homoscedasticity is required. The aim of the regression is to find that relationship while minimizing the sum of squared residuals. In reality, the linear relationship might not always be the best fit. The model can use numeric variables only.

3.3 Logistic regression

For the purpose of the present thesis, however, linear regressions are not appropriate. Our dependent variables are default events, which occur with certain frequencies/ probabilities. A probability takes values on the unit interval only, while our independent variables may be assumed to be real valued. Thus, a linear choice for f cannot be used. The simple approach taken by the logistic regression maps a real number in a monotonic way onto the unit interval, thus obtaining a variable $\pi \in [0, 1]$ with the possible interpretation as a probability.

We define the logistic function

$$F(z) = \frac{1}{1+\exp(-z)} \tag{3.3-1}$$

which is a strictly increasing function (Figure 5).

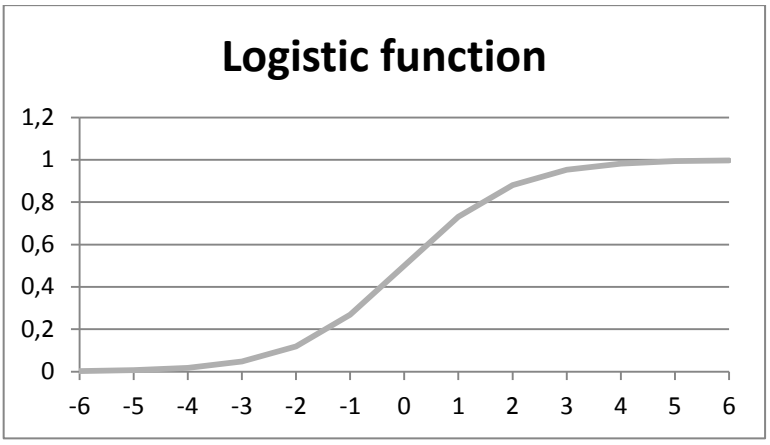

FIGURE 5: LOGISTIC FUNCTION (SOURCE: OWN CALCULATIONS)

We assume a *d*-dimensional vector of independent variables $(X_1, \dots, X_d)$ and we abbreviate $X = (X_1, \dots, X_d)$, where we artificially set X_0 = 1. For $\beta = (\beta_0, \beta_1, \dots, \beta_d) \in \mathbb{R}^{d+1}$ we introduce the notation

$$\beta.X = \beta_0 + \sum_{i=1}^{d} \beta_i X_i \tag{3.3-2}$$

The regression formula reads

$$\pi = F(X;\beta) = \frac{1}{1+\exp(-\beta.X)} \tag{3.3-3}$$

Alternatively, we can write, using the logit transform $\log(\pi/(1-\pi)$

$$\log\left(\frac{\pi}{1-\pi}\right) = \sum_{k=0}^{d} \beta_k X_k \tag{3.3-4}$$

3.3.1 ML Estimation

We assume now we have $n_j (j = 1, \dots, N)$ realizations of *Y*, and that we have *N* realization of the *d*-dimensional random variable $X_1, \dots, X_d$. We denote the j_{th} realization by

$$y_{j,} \quad x_j : (1, x_{j,1} \dots, x_{j,d}) \tag{3.3-5}$$

where y_j counts the number of defaults in the j_{th} sample of size n_j. We further introduce the dependent variable π_j associated with the j_{th} sample. The standard approach of linear regression, to use least-squares-method to derive parameter estimates $\beta_i \; of \; \beta_j (i = 0, \dots, d)$ for which the residual

$$R(\beta) = \sum_{j=1}^{N} (\pi_j - F(x_j ; \beta))^2 \rightarrow min! \tag{3.3-6}$$

does not work for the present setting, as it does not lead to a non-biased, minimum-variance estimator.

We model default as a Bernoulli event with parameter π. It is well known that for a independent and identically distributed sequence of Bernoulli distributed random variables $\varepsilon_1, \varepsilon_{2,} \dots, \varepsilon_n$ with parameter π, we have that $Y = \sum_{i=1}^{n} \varepsilon_i$ is Binomially distributed with parameters (n, π). Therefore the Likelihood function, which is just the density of the discrete random variable *Y*, is given by

$$\varphi(y|\beta) = \prod_{j=1}^{N} \binom{n_j}{y_j} \pi_j^{y_j} (1-\pi_j)^{n_j - y_j} = \frac{n_j!}{y_j!(n_j - y_{j)!}} \pi_j^{y_j} (1-\pi_j)^{n_j - y_j} \quad (3.3\text{-}7)$$

for any possible realization $y \epsilon \{0,1, \dots., n\}\ of\ Y$. The Maximum-Likelihood method consists in finding the parameter $\hat{\beta}$ which minimizes $\varphi(Y|\beta)$. The constant factor in (3.3-7) is irrelevant for minimization. Furthermore, we may use (3.3-3) for π. Hence, we can define the Likelihood function as (Czepiel, 2011)

$$l(y|\beta) = \prod_{j=1}^{N} e^{(\beta.x)y_j} \left(1 - \frac{e^{\beta.x_j}}{1+e^{\beta.x_j}}\right)^{n_j} = \prod_{j=1}^{N} e^{(\beta.x_j)y_j} (1 + e^{\beta.x_j})^{-n_j} \quad (3.3\text{-}8)$$

In view of the monotonicity of the natural logarithm, minimizing this likelihood is equivalent to minimization of the log-likelihood function

$$l(y|\beta) = \sum_{j=1}^{N} \left(y_j (\beta.x_j) - n_j \log(1 + e^{\beta.x_j})\right) \quad (3.3\text{-}9)$$

A value $\hat{\beta} \in \mathbb{R}^{d+1}$ is called a critical point of $l(y|.)$, if

$$\frac{\delta l}{\delta \beta}|_{\beta=\hat{\beta}} = 0 \quad (3.3\text{-}10)$$

Using elementary differentiation rules, we obtain formulas for the first and second order derivatives of the log likelihood function l (cf. (Czepiel, 2011), equations (11) and (12))

For $k, k' \in \{0,1, \dots, d\}$ we have

$$\frac{\delta l}{\delta \beta_k} = \sum_{j=1}^{N} y x_{j,k} - n_j \pi_j x_{j,k} = \sum_{j=1}^{N} x_{j,k} \left(y_j - n_j \frac{e^{\beta.x_j}}{1+\exp(\beta.x_j)}\right) \quad (3.3\text{-}11)$$

$$Hl_{kk'} = \frac{\delta^2 l}{\delta \beta_k} = \sum_{j=1}^{N} n_j \pi_j x_{j,k} x_{j,k'} (1 - \pi_j) = -\sum_{j=1}^{N} n_j x_{j,k} x_{j,k'} \frac{e^{\beta.x_j}}{(1+e^{\beta.x_j})^2} \quad (3.3\text{-}12)$$

The matrix Hl with entries defined in (3.3-12) is the Hessian[4] of l. We formulate now two basic facts which lead to a recipe for finding the ML estimator $\hat{\beta}$.

Theorem 1.1. Consider equation (3.3-7). The following holds:

[4] In mathematics, the Hessian matrix or Hessian is a square matrix of second-order partial derivatives of a function.

(1) A point $\hat{\beta} \in \mathbb{R}^{d+1}$ is a critical point of (3.3-7), if and only if β satisfies the non-linear system

$$\sum_{j=1}^{N} x_{j,k}(y_j - n_j\left(\frac{e^{\beta.x_j}}{1+\exp(\beta.x_j)}\right) = 0 \qquad (3.3\text{-}13)$$

(2) Assume that Hl is negative definite, for all β. Then, if is a critical point of (3.3-11), it is the unique global maximum.

Proof. Part (1): Since $l(y,.)$ is differentiable everywhere, any critical point is characterized by equality in the non-linear system (3.3-13) derived from setting (3.3-11) equals zero.

Part (2): Due to equation (3.3-12), the Hessian of l(y,.) is negative definite. Hence any critical point is a local maximum. There can only be one global maximum, because l is strictly concave.

According to Proposition (3.3-3) we need to find a solution $\hat{\beta}$ to (3.3-13), and then check negative definiteness of the Hessian matrix Hl, for all β. Both tasks are non-trivial in the present setting.

In fact, it is well known that even a globally strictly concave function need not have a maximum (e.g. the $g(x) = -\exp(x)$), whereas the literature gives sufficient conditions for the existence of global maxima (see, e.g. (Soriano, 1993) and note that their sufficient criterion is not fulfilled in the present setting).

3.3.2 The logit and odds theory

The logit transformation $\log(\pi/(1-\pi)$, used in (3.3-4), represents the many of the desirable properties of a linear regression model. The logit is linear in its parameters, may be continuous, and may range from minus infinity to plus infinity depending on the range of x. (Hosmer & Stanley, 1989) .

Suppose that the binary variable takes two values, success and fail. Then, the probability of success is π and the odds of success are $(\pi/(1-\pi)$. The log of the odds of success is opposite to the log of the odds of fail. Probability ranges from 0 and 1 and odds range from 0 and positive infinity. The higher is the probability, the higher are the odds.

The graphic representation below shows the inverse shape of odds and log of the odds.

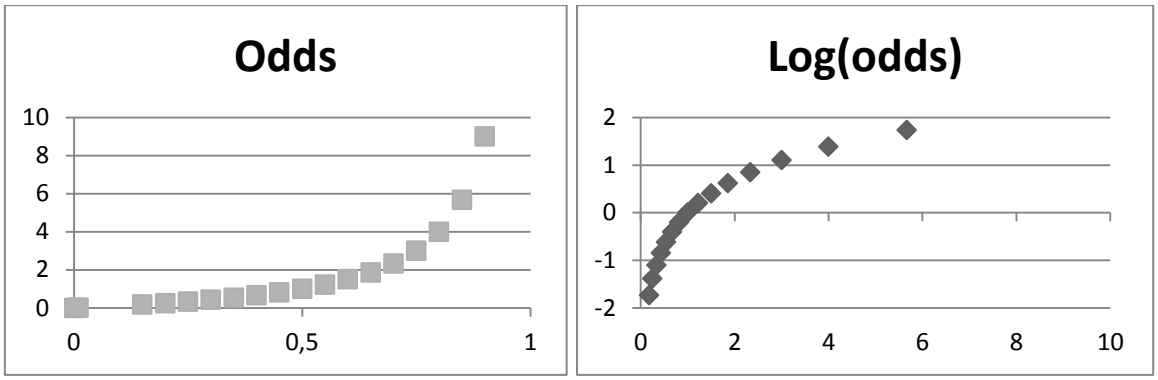

FIGURE 6: ODDS AND LOG OF ODDS (SOURCE: OWN CALCULATIONS)

The transformation from probability to log odds is an attempt to deal with the difficulty of modelling a variable with restricted range. It maps probability ranging between 0 and 1 to log odds ranging from negative infinity to positive infinity. Another reason is that among all of the infinitely many choices of transformation, the log of odds is one of the easiest to understand and interpret (Statistical Consulting Group).

3.4 Discriminant analysis

Discriminant analysis is one of the most used methods of credit scoring around the world. The major purpose of the discriminant analysis is to find a set of features that can best determine group membership of the object and find a classification rule or model to best separate those groups.

This method gives a solution for problems with use of categorical variables, although requiring them to be continuous. There are also the other inevitable shortcomings: the model assumes that the distributions of independent variables are normally distributed, but in practice the data are often not completely normal distribution, resulting in the unreliability of statistical results. Despite its strict restrictions on data, it still has value when it comes to multiple group classification. Unlike binary Logistic regression, Discriminant analysis as an earlier alternative to logistic regression, can handle more than 2 groups. A disadvantage of this approach is that it does not yield estimated PDs.

3.4.1 Discriminant analysis theory

When two or more populations have been measured in several characters, special interest attaches to certain linear functions of the measurements by which the populations are best discriminated (Fisher, 1936). In 1936, Fisher introduced the idea of discriminating between groups in a population (of plants). In 1941, Durand (Durand, 1941), who was working on a research project for the US National Bureau of Economic Research, realized that Fisher's discriminant analysis could be used to differentiate between good and bad loans. He has made a study on 7200 loans with 37 banks and finance companies, for the purpose of identifying credit standards.

The discriminant model generates a classification equation:

$$c_g = c_1 x_1 + c_2 x_2 + \cdots + c_n x_n \quad (3.4\text{-}1)$$

where:
c_g is the sample classification score for each group (g),
c_0 is a constant,
c_i, $i = 1,2, \ldots, n,$ are classification coefficients, where n is the number of variables,
x_i, $i = 1,2, \ldots, n,$ are the unstandardized variable values.

The discriminant function, also known as a classification criterion, is determined by a measure of generalized squared distance (Rao, 1973). Each observation is placed into the group from which it has the smallest generalized distance. The generalized squared distance from a multivariate vector $X = (X_1, \ldots, X_d)$ from a vector of values with mean $\bar{X} = (\bar{X}_1, \ldots, \bar{X}_d)$ and covariance matrix S is defined as:

$$D_M(x) = \sqrt{(x-\mu)^T S^{-1} (x-\mu)} \quad (3.4\text{-}2)$$

This statistic is sometimes called the Mahalanobis distance. It is a unitless measure introduced by P. C. Mahalanobis in 1936 (Mahalanobis, 1936). It differs from Euclidean distance in that it takes into account the correlations of the data set and is scale-invariant. This generalized squared distance is then converted into a score of similarity to each group, and the case is classified into the group it is most similar to.

The discriminant analysis might seem similar to the cluster analysis. But the aims of the two methods are completely different. While the purpose of the cluster analysis is to construct a classification, the discriminant analysis requires prior knowledge of the classes. Unlike logistic regression, the discriminant analysis can be used with small sample size datasets. On the other hand, the discriminant analysis is limited by more assumptions and restrictions than the logistic regression.

These are assumptions of normal distribution for the response variables. Not only that each of the dependent variables is normally distributed within groups, but that any linear combination of the dependent variables is normally distributed. Another important assumption is that each group must have a sufficiently large number of cases.

3.4.1 Prior probabilities

As said before, the linear discriminant function predicts group membership based on the squared Mahalanobis distance from each observation to the centroid of the group plus a function of the prior probability of membership in that group. The prior probability is the

probability of an observation coming from a particular group in a simple random sample with replacement.

If the prior probabilities are the same for both groups (also known as equal priors) then the function is only based on the squared Mahalanobis distance. The prior probabilities are 50% for each group. If the prior for group A is larger than for group B, then the function makes it more likely that an observation will be classified as group A.

The default in software programs is usually equal priors, as the function is the simplest, and therefore the most computationally efficient. Alternatives are proportional priors (using priors that are the proportion of observations from each group in the same input data set) and user-specified priors.

3.4.2 Error rates and cross validation

To evaluate the performance of a discriminant criterion is commonly done by estimating error rates (probabilities of misclassification) in the classification of future observations.
These error-rate estimates include error-count estimates and posterior probability error rate estimates. The error rate can also be estimated by cross validation.

The data that is available for model building is not always sufficient to create a sizable training set and a validation set that represent the predictive population well. This is the case of a low default portfolio. The frequency of the default events is so low that one can't allow for a division of the dataset to satisfy both modelling and validation needs.

Cross validation is an attractive alternative for estimating prediction error. Cross-validation will not solve the problem of lack of a validation data set, but will increase the robustness of the model built on the training data set.

It consists of partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set). This fitted model is used to compute the predicted residual sum of squares on the omitted part, and this process is repeated for each of *k* parts. Every data point gets to be in a test set exactly once, and gets to be in a training set *k*-1 times. The variance of the resulting estimate is reduced as *k* is increased. The disadvantage of this method is that the training algorithm has to be rerun from scratch k times, which means it takes k times as much computation to make an evaluation. The sum of the predicted residual sum of squares so obtained is the estimate of the prediction error.

Cross validation is a model evaluation method that is better than residuals. The problem with residual evaluations is that they do not give an indication of how well the machine learning algorithm will do when it is asked to make new predictions for data it has not already seen. A way to overcome this is to remove data before training begins. Then when training is done,

the data that was removed can be used to test the performance of the learned model on "new" data. The issue here could be if there isn't sufficient data in each group to allow for reducing the frequency.

There are different types of cross validation:

- In one-at-a-time cross validation, the first observation is held out as a single-element test set, with all other observations as the training set. This is repeated for all observations.
- In blocked validation, the dataset is divided into several blocks of test sets.
- A similar method is split-sample cross validation, in which successive groups of widely separated observations are held out as the test set.
- Random sample cross validation where test sets can be selected from the observed data randomly

3.5 Decision tree

Decision tree model specifies the interaction of characteristics in a control hierarchical order and determines the probability of attributes to be chosen. It makes use of the ''if then else'' rules to base the next question from the answer to the preceding question. A probability value is assigned to each attributed option. Although decision tree models are argued to be more predictive than scorecard models, they can, however, abruptly become complex and unstable when updated information leads to an alteration in the first question and results in a dramatic change of their decision structure (Cheng, 2007).

3.6 Neural network

Neural network models are even more flexible as they allow the characteristics to be interacted in a variety of ways. They consist of a group or groups of connected characteristics. A single characteristic can be connected to many other characteristics, which make up the whole complicated network structure. They outweigh decision trees and standard scoring methods, as they do not assume uncorrelated relations between characteristics. They also do not suffer from structural instability in the same way as decision trees because they may not rely on a single first question for constructing the whole network. Yet, the network is difficult to build up (Cheng, 2007), (West, 2000).

3.7 Kernel methods

Kernel methods (support vector machine) is an example of an alternative method, survival analytic models are mainly employed to determine if a loan will result in a default or be paid-off before the mature date. Such a model traces the loan repayment performance of each borrower in a certain period (say the first year of a 5-year loan). The model predicts the

likelihood of a default for the remaining period is predicted. Such models, therefore, are not appropriate for loan application exercises unless loan repayment performance can be determined prior to the commencement of the project (Cheng, 2007).

3.8 Method selection in literature

Various studies have been produced in order to assess performance and accuracy of the methods applied. The results and conclusion on the superiority of any method were not consistent. (Desai, Crook, & Overstreet, 1996) investigated neural networks, linear discriminant analysis and logistic regression for scoring credit decision. They concluded that neural networks outperform linear discriminant analysis in classifying loan applicants into good and bad credits, and logistic regression is comparable to neural networks. The study of Yobas et al. compared all four techniques: traditional, neural networks, genetic algorithms and decision trees using the same credit applicant datasets and using a realistic division of cases between the 'good' and 'bad' groups. a. They found that LDA was superior to Gas like many other studies, but neural networks were inferior to LDA (Yobas, Crook, & Ross, 2000).

Investigation of Neural Networks for the purpose of credit scoring is the aim of another study done by (West, 2000). It concludes that from the traditional methods, logistic regression is found to be the most accurate of the traditional methods. Both the mixture-of-experts and radial basis function neural network models should be considered for credit scoring applications.

4 DATA ENVELOPMENT ANALYSIS AS AN ALTERNATIVE APPROACH TO MANAGING CREDIT RISK

Data envelopment analysis (DEA) can overcome some of the difficulties the banks deal with when modelling the default prediction of their customers.

Data envelopment analysis is an effective tool for multi-criteria decision-making used across various sectors to evaluate efficiencies[5] of decision making units (DMU[6]). It was first developed for use in evaluating activities of not-for-profit entities participating in public programs (Charnes, Cooper, & Rhodes, 1978). Over the years, it has been applied not only on public but as well as on private entities, such as hospitals, universities, manufactures and banks.

The stages of DEA must all be related to the aims and values of the organisation:

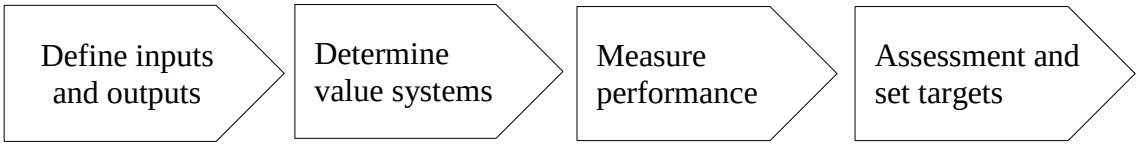

FIGURE 7: STAGES OF THE DEA (SOURCE: OWN CHART)

As opposed to the above mentioned methods, such as logistic regression, discriminant analysis or neural network, the DEA does not require ex ante information of good/bad classification. This method only needs ex post information of the observed inputs[7] and outputs[8] to calculate the credit scores (Min & Lee, 2004).

The theory of the Data envelopment analysis is based on measuring an efficiency score using a weighted sum of inputs and outputs. This efficiency score is calculated for each unit and compared.

Assume a set of n decision making units with m inputs and r outputs. The efficiency θ is determined as the weighted sum of outputs to the weighted sum of inputs.

[5] Pareto-Koopmans (Koopmans , 1951) efficiency is achieved if and only if none of the inputs or outputs can be improved without worsening other inputs or outputs.
[6] DMU are units of organizations such as banks, universities or hospitals that usually perform the same function. Use of this term redirects emphasis of the analysis from profit-making businesses to decision-making entities. The analysis can be applied to any unit-based enterprise that controls its mix of inputs and decides on what outputs to produce.
[7] Any resource used by a unit to produce its outputs (products or services); can include resources that are not a product but are an attribute of the environment in which the units operate, and they can be controlled or uncontrolled.
[8] The products (goods, services, or other outcomes) that result from the processing and consumption of inputs (resources); may be physical goods or services or a measure of how effectively a unit has achieved its goals, and may include profits where applicable.

$$\theta = \frac{\sum_{k=1}^{r} u_k y_{kj}}{\sum_{i=1}^{m} v_i x_{ij}} \quad (5\text{-}1)$$

where v_i is the assigned weight of the i-th input x_{ij} for unit j, $i = 1,2,\dots,m$, $j = 1,2,\dots,n$ and u_k is the assigned weight of the k-th output y_{kj} for unit j, $k = 1,2,\dots,r$, $j = 1,2,\dots,n$.

The inputs and outputs are the actual values observed, but the weights need to be determined. The difficulty comes in obtaining an agreed common set of weights. Units might do this in a different way and this way doesn't have to give the best results. Therefore, the DEA itself assigns a unique set of weights for each unit. The weights are determined using mathematical programming, while maximising the efficiency of the unit.

The results of a DEA analysis can be pictured in a chart, as shown in Figure 8. Given a fixed input x and two outputs $y_1$and y_2, the efficient frontier[9] on the chart is defined by points *D, E, F, G, H*. Units providing greater amounts of outputs with the fixed input are called efficient and form the efficient frontier here.

Applying the DEA approach to this set of units will identify units *D, E, F, G* and *H* as efficient forming an envelope round the entire data set. Units *A, B* and *C* are located within this envelope therefore inefficient. To become efficient, they need to look for a target. These are points *A'* for unit *A, B'* for unit *B* and *C'* for unit *C*. The target points all lie on the intersection of the efficient frontier and the straight line traced from the origin.

By projecting each unit onto the frontier, it is possible to determine the level of inefficiency by comparing them to a single reference unit or a convex combination of other reference units. The projection refers to a virtual DMU which is a convex combination of one or more efficient DMUs. Thus, the projected point may itself not be an actual DMU.

[9] The efficient frontier is a concept introduced by (Markowitz, 1952). He applied linear programming to portfolio theory. The efficient frontier was formed by portfolios with best possible return for its level of risk.

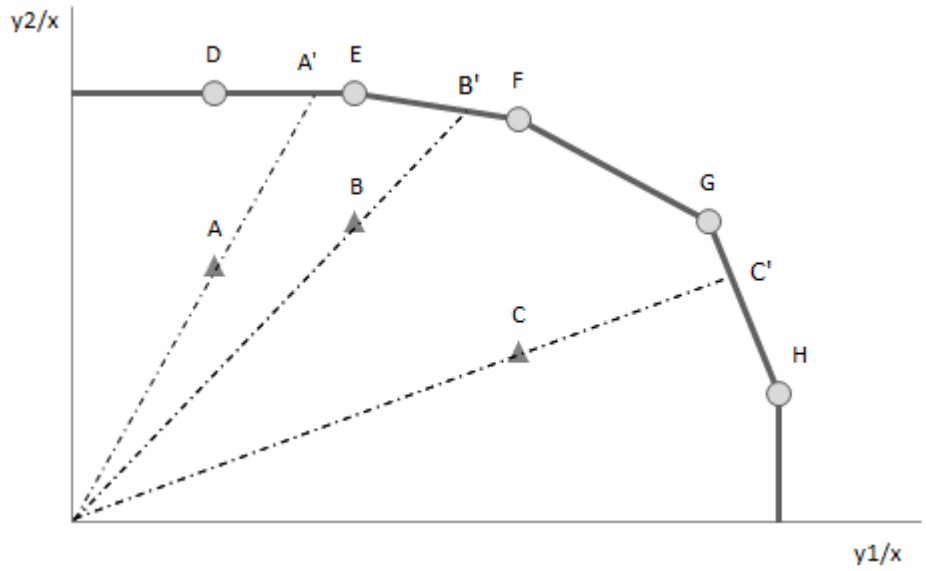

FIGURE 8: EFFICIENT FRONTIER (SOURCE: OWN CHART)

The efficiency score that is the result of the DEA model is in fact a relative efficiency. The outcome of the DEA model is not an absolute efficiency of the unit, but efficiency compared to the rest of the bunch and therefore we call this a relative efficiency.

4.1 Basic DEA models

There are 2 basic models of the DEA, varying in the nature of the returns to scale, the CCR model and the BCC model.

4.1.1 CCR model formulation

The CCR model was developed by Charnes, Cooper and Rhodes (Charnes, Cooper, & Rhodes, 1978). It took more than 20 years to design mathematical framework for the work of M.J. Farrell on the Frontier analysis (Farrell, 1957). His article intended to deal with shortcomings in the usual index number approach to productivity measurement.
The proposed measure of the efficiency of any DMU is obtained as the maximum of a ratio of weighted outputs to weighted inputs subject to the condition that the similar ratios for every DMU be less than or equal to unity (Charnes, Cooper, & Rhodes, 1978).

Having *n* units, *r* outputs and *m* inputs, the efficiency of the *DMUj* is defined as follows:

$$max \ \frac{\sum_{k=1}^{r} u_k y_{kj}}{\sum_{i=1}^{m} v_i x_{ij}} \qquad (4.1\text{-}1)$$

where $u_k > 0$, $k = 1,2, \dots, r$ for all units $j, j = 1,2, \dots, n$
$v_i > 0$, $i = 1,2, \dots, m$ for all units $j, j = 1,2, \dots, n$

Here x_{ij}, y_{kj} are the known outputs and inputs of the, j_{th} DMU and $u_k, v_i$are the weights to be determined by the solution of this problem. To prevent zero weights given to factors that manage poorly and to avoid false technical efficiency, the non-archimedean infinitesimal[10] ε was later introduced:

$$u_k, v_i \geq \varepsilon \tag{4.1-2}$$

The model is looking for optimum weights that would lead to the maximum possible efficiency score where both pure technical efficiency and scale efficiency are aggregated into one value.

Once converted into a linear form, the following model will be possible to solve with the techniques of linear programming. The denominator must be removed from the objective function to eliminate non-linearity and put equal to 1. The outputs can't exceed the inputs and therefore the weighted outputs minus weighted inputs.
The final linear model has the following form:

$$\sum_{k=1}^{r} u_k y_{kj}$$

$$\sum_{i=1}^{m} v_i x_{ij} = 1 \tag{4.1-3}$$

$$\sum_{k=1}^{r} u_k y_{kj} - \sum_{i=1}^{m} v_i x_{ij} \leq 0 \ , i = 1,2,\dots,m \ , k = 1,2,\dots,r \ , \ j = 1,2,\dots,n$$

$$u_k, v_i \geq \varepsilon$$

The CCR model assumed constant returns to scale and was built on the basic findings in economics, such as production functions and related concepts. Constant returns to scale may be assumed if an increase in a unit's inputs leads to a proportionate increase in its outputs i.e. there is a one-to-one, linear relationship between inputs and outputs. For example, if a *k*-% increase in inputs yields a *k*-% increase in outputs, the unit is operating at constant returns to scale. This means that no matter what scale the unit operates at, its efficiency will, assuming its current operating practices, remain unchanged.

4.1.2 BCC model description

The CCR model was further developed by Banker et al. (Banker, Charnes, & Cooper, 1984) allowing for variable returns to scale[11]. It was called after their initials, the BCC model. The

[10] Extremely small positive value that is impossible to measure. Its use ensures that all inputs and outputs are accorded some positive value. (Arnold, Bardhan, Copper, & Gallegos, 1998)

[11] Returns to scale refer to the production function and a proportional change in inputs resulting in changes in output. If input and output change by the same factor, we talk about constant returns to scale (CRS). If output increases more than the change in input, we talk about the increasing returns to scale (IRS). If output increases less than the proportional change in input, we talk about decreasing returns to scale (DRS). IRS and DRS are two types of the variable returns to scale (VRS). More details about returns to scale in DEA models can be found in (Banker, Cooper, Seiford, Thrall, & Zhu, 2004)

main distinction between the BCC and the CCR model is the introduction of the parameter *w*, not defining the envelopment surface to go through the origin. The technical and scale efficiency are separated. And a new variable is introduced to determine whether operations were conducted in regions of increasing, constant or decreasing returns to scale (Banker, Charnes, & Cooper, 1984).

The BCC model has the property of convexity. The convexity constraint, which forms part of the formulation of the BCC model, ensures that each composite unit is a convex combination of its reference units. Under the CRS assumption, the DEA efficient frontier[12] would be represented with a line, causing less units to become efficient than on the convex curve in the case of VRS. Figure 9 demonstrates the impact of the constant returns to scale compared to the variable returns to scale. In the first case, only unit C is efficient, compared to 3 units being found efficient under VRS.

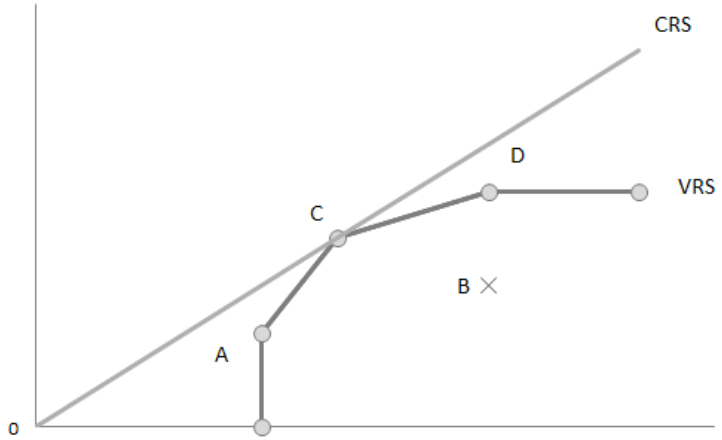

FIGURE 9: CRS AND VRS MODELS (SOURCE: OWN CHART)

Every linear program can be formulated in a different way without any change to the results. This is called duality. The CCR model shown above is a primal model. Its dual model is formulated below, by assigning a dual variable to each constraint in the primal model. The dual models contain new variables, the so called slacks[13].

s^+ is added to a *less than or equal to* constraint to convert inequality to an equality.
s^- is a surplus variable subtracted from a greater than or equal to constraint to convert it to equality.

[12] DEA efficient frontier is made up of the best performing DMUs. These are 100% efficient. Any unit not on the frontier has an efficiency of less than 1.
[13] The underproduction of outputs or the overuse of inputs; represents the improvements (in the form of an increase/decrease in inputs or outputs) needed to make an inefficient unit become efficient.

Input-oriented (4.1-4)	**Output-oriented** (4.1-5)
$max\ \sum_{k=1}^{r} u_k y_{k0} + \mu$ subject to $\sum_{i=1}^{m} v_i x_{i0} = 1$ $\sum_{k=1}^{r} u_k y_{kj} - \sum_{i=1}^{m} v_i x_{ij} + \mu \leq 0$ $u_{k,} v_i > 0,\ i = 1,2,\dots,m\ ,\ k = 1,2,\dots,r\ ,$ $j = 1,2,\dots,n$	$min\ \sum_{i=1}^{m} v_i x_{i0} + v$ subject to $\sum_{k=1}^{r} u_k y_{k0} = 1$ $\sum_{i=1}^{m} v_i x_{ij} - \sum_{k=1}^{r} u_k y_{kj} + v \geq 0$ $u_{k,} v_i > 0,\quad i = 1,2,\dots,m,\ \ k = 1,2,\dots,r,$ $j = 1,2,\dots,n$
The value of the μ defines the nature of the returns to scale. If μ = 0 then the returns to scale are presumed to be constant. If free then the returns to scale are presumed to be variable.	The value of the *v* defines the nature of the returns to scale. If *v* = 0 then the returns to scale are presumed to be constant. If free then the returns to scale are presumed to be variable.

TABLE 2: PRIMAL DEA MODELS (SOURCE: (ZHU, 2009))

The above models, (5.1-4) and (5.1-5), are sometimes called multiplier models. In an input oriented model, DMU is not efficient if it is possible to decrease any input without decreasing any output or increasing another input. Inversely, in the case of an output oriented model, a DMU is not efficient if it is possible to increase any output without increasing any input or decreasing another output.

The primal model has ***n + s + m + 1*** constraints whilst the dual model has ***m + s*** constraints.

As *n*, the number of units, is usually considerably larger than *t + m*, the number of inputs and outputs, it can be seen that the primal model will have many more constraints than the dual model. For linear programs in general the more constraints the more difficult a problem is to solve. Hence for this reason it is usual to solve the dual DEA model rather than the primal (Emrouznejad, DEA Zone).

Dual models are formulated as follows.

Input-oriented (4.1-6)	**Output-oriented** (4.1-7)
$\min \theta_0 - \varepsilon(\sum_{i=1}^{m} s_i^- + \sum_{k=1}^{r} s_k^+)$ $\sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = x_{i0}\theta_0$ $\sum_{j=1}^{n} \lambda_j y_{kj} - s_r^+ = y_{k0}$ $\lambda_j \geq 0\ , j = 1,2,\dots,n$ $s_i^-, s_k^+ \geq 0\ , i = 1,2,\dots,m,\ k = 1,2,\dots,r$	$\max \phi_0 - \varepsilon(\sum_{i=1}^{m} s_i^- + \sum_{k=1}^{r} s_k^+)$ $\sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = x_{i0}$ $\sum_{j=1}^{n} \lambda_j y_{kj} - s_k^+ = y_{k0}\phi_0$ $\lambda_j \geq 0\ , j = 1,2,\dots,n$ $s_i^-, s_k^+ \geq 0,\ i = 1,2,\dots,m,\ k = 1,2,\dots,r$
The above model assumes constant returns to scale. For variable returns to scale, we need to add an extra constraint of $\sum_{j=1}^{n} \lambda_j = 1$.	The above model assumes constant returns to scale. For variable returns to scale, we need to add an extra constraint of $\sum_{j=1}^{n} \lambda_j = 1$.

TABLE 3: DUAL DEA MODELS (SOURCE: (ZHU, 2009))

The linear programming technique looks for values of λ_j to form a unit with outputs $\sum_j \lambda_j y_{kj}$ and inputs $\sum_j \lambda_j x_{i0}$ than would be more efficient than unit under evaluation.

The two models are symmetric and the results of the dual model are the so called shadow prices. They are very useful for economic interpretations. The shadow prices give us directly the marginal worth of an additional unit of any of the resources (constraints).

The aim of the DEA is not only to calculate the efficiencies of the DMUs, but to provide additional information how to reach the efficient frontier for the inefficient units. The values that the inefficient units should aim for are called target values.

For unit *j*, $j = 1,2,\dots,n$, the target value is calculated:

a) by means of productive unit vectors

$$x_i^* = \sum \lambda_j^* x_{ij}\ , i = 1,2,\dots,m \quad (4.1\text{-}8)$$
$$y_k^* = \sum \lambda_j^* y_{kj}\ , k = 1,2,\dots,r \quad (4.1\text{-}9)$$

where λ^* is the vector of optimal variable values for each unit *j, j=1,2,…,n*.

b) by means of the efficiency rate and values of the slack variables s^- and s^+

Input-oriented CCR model

$$x_i^* = \sum \theta x_{ij} - s_i^- \ , i = 1,2, \dots, m \quad (4.1\text{-}10)$$
$$y_k^* = \sum y_{kj} + \ s_k^+ \ , k = 1,2, \dots, r \quad (4.1\text{-}11)$$

Output-oriented CCR model

$$x_i^* = \sum x_{ij} - s_i^- \ , i = 1,2, \dots, m \quad (4.1\text{-}12)$$
$$y_k^* = \sum \emptyset y_{kj} + \ s_k^+ \ , k = 1,2, \dots, r \quad (4.1\text{-}13)$$

where θ is the efficiency rate in the input-oriented model and $\emptyset$ is the efficiency rate in the output-oriented model.

4.2 DEA modifications and developments

Two basic models of the DEA, the CCR model and the BCC model, have undergone many modifications and developments to better fit the real life problems. In general these models differ in their Orientation (Input or Output orientation), Diversification and Returns to Scale (CRS –constant returns to scale, VRS – variable returns to scale, NIRS – non-increasing returns to scale, NDRS – non-decreasing returns to scale, etc.) or types of measure (Radial measure, Non-radial measure, Hyperbolic measure, etc.).

The following section looks in more detail on various modifications and model enhancements that have been discussed in the literature.

4.2.1 Weights restriction

An example of such modification is weight restriction. The basic CCR model imposes no constraints on factor weighs and allows for flexibility in the choice of weights. This can be strength and a weakness at the same time. No a priori values are assigned to the various weights. Thus, in the basic CCR model, the only constraint on factor weights in a positivity requirement (apart from the output-input relationships, providing the relative nature of the analysis). The same factor may be assigned different weights, when viewing the relative efficiency of different DMUs. Imposing bounds on factor weights is aimed at controlling both kinds of flexibility (Golany & Roll, 1993)

4.2.2 Malmquist index

The DEA based Malmquist productivity index can be decomposed into technical change (whether the production frontier is moving outwards over time) and efficiency change (whether firms are getting closer to the production frontier over time) using the Malmquist

index proposed by Färe et al. (Färe R. , Grosskopf, Roos, & Lindgren, 1992). But in fact, the index to measure productivity change was introduced already 10 years earlier by (Caves, Christensen, & Diewert, 1982) and they named the index after Malmquist[14]. However, Färe et al. merged efficiency theory as developed by (Farrell, 1957) with the Malmquist index of Caves et al. to propose a Malmquist index of productivity change that is now commonly used in the literature.

This index is composed of distance functions, and is therefore superior to alternative indexes of productivity growth (such as the Törnqvist index and the Fisher Ideal index) because it is based only on quantity data and makes no assumptions regarding the firm's behaviour (Grifell-Tatjé & Lovell, 1996).

4.2.1 Super-efficiency

A weakness of DEA is that, typically, more than one unit exists that can be evaluated as efficient when the number of DMUs is not enough relative to the number of inputs and outputs. Super-efficiency data envelopment analysis model, developed by (Andersen & Petersen, 1993), can be used in ranking the performance of efficient decision making units (DMUs). Super-efficiency data envelopment analysis model is identical to the standard model except that the DMU under evaluation is excluded from the reference set – λ of the DMU is set to 0.

Under the assumption of variable returns to scale (VRS, NIRS, NDRS), the super-efficiency model may be infeasible for some efficient DMUs. (Chen, 2005) and other authors have come with models overcoming the infeasibility of a super-efficiency model.

Super-efficiency models can be used not only to rank efficient units, but as well for outlier identification as shown in (Banker & H.Chang, 2006) or measuring productivity changes (Färe R. , Grosskopf, Roos, & Lindgren, 1992), (Berg, Førsund, & Jansen, 1992) or sensitivity and stability analysis (Zhu, 2001).

4.2.2 Inputs and outputs control

Knowing the efficiency for a certain DMU might not be sufficient and one can look for more information on the impact of a certain input or output on the level of performance. In these cases, a subset of inputs are reduced in the same proportion while keeping outputs at their current level or a subset of outputs are increased in the same proportion while keeping inputs at their current level. Measure specific models (Zhu, 2009) provide that information and can be used to model uncontrollable inputs and outputs (Banker & Morey, 1986).

[14] Professor Sten Malmquist was a Swedish economist and statistician. His ideas were behind the construction of the index named after him, the Malmquist Index (Malmquist, 1953).

The management can prefer different targets along the efficient frontier than the ones provided by the development and non-radial DEA models. (Zhu, 1996) has developed a set of weighted non-radial DEA models where various efficient targets along with the frontier can be obtained. The input oriented preference structure model is actually a DEA model with fixed input multipliers. As shown in (Chen, 2005), the DEA/preference structure models can be derived by traditional MOLP techniques.

4.2.3 Stratification DEA method

The stratification method allows for an algorithm to remove the best practice frontier to allow the remaining inefficient DMUs to form a new second level best practice frontier. This can be done until there are no more inefficient units. The basic idea of this method is partitioning of the set of DMU into several levels of best practice frontiers. Each best practice frontier provides an evaluation context for measuring the relative attractiveness (Zhu, 2009).

4.3 Network DEA

Network DEA (NDEA) models were introduced by (Färe & Whittaker, 1995). They investigated the underlying performance information in a firm's interacting divisions or sub-processes that would otherwise remain unknown to management. There is usually no information about what happens inside the sub-processes.

The efficiency estimates that are produced by NDEA and account for divisional interactions are more representative of a dynamic business than static measures reporting overall performance without opening the so-called "black box" of production. The combination of sub-technologies into networks provides a method of analysing problems that the traditional DEA models cannot address. The specification of the sub-technologies enables the explicit examination of input allocation and intermediate products that together form the production process. (Färe, Grosskopf, & Whittaker, 2007).

Two-stage Network DEA model is the simplest model in the NDEA framework. (Färe & Grosskopf, 2000) develop a general formulation of the network DEA which attempts to provide deeper structure to the 'black box' transformation of the conventional DEA. They have applied the methodology on 3 examples: a static production technology with intermediate products, a dynamic production technology, and technology adoption (or embodied technological change). In the two-stage DEA, all the outputs from the first stage are the only inputs to the second stage, in addition to the inputs to the first stage and the outputs from the second stage. The outputs from the first stage to the second stage are called intermediate measures. Two-stage model developments can be found in (Sexton & Lewis, 2003). Their model is similar to the one presented by (Färe & Whittaker, 1995) and (Färe & Grosskopf, 1996) and (Färe & Grosskopf, 2000), except for 2 aspects: their approach explicitly computes the efficiencies of the sub-DMUs and establishes separate efficient frontiers for Stage 1 and Stage 2.

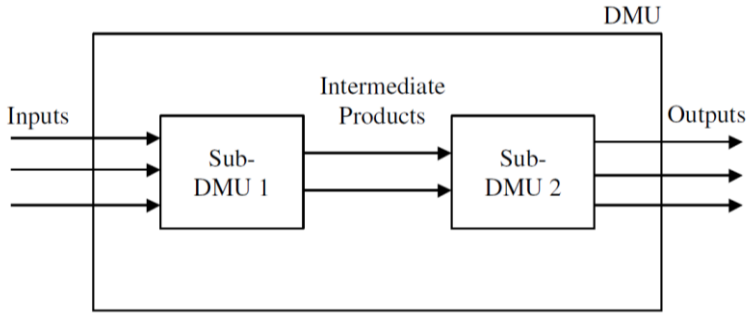

FIGURE 10: TWO-STAGE NETWORK DEA MODEL (SEXTON & LEWIS, 2003)

We will assume output orientation and constant returns to scale as for the standard DEA model presented above. We formulate and solve 3 DEA problems. Assume x_{ij} is the level of input i consumed by DMU j, for $i = 1,2,\dots,m$ and y_{kj} is the level of output produced and consumed by DMU j, for $k = 1,2,\dots,r$ and z_{pj} the level of output produced by DMU_j, for $p = 1,2,\dots,s$. Then the efficiency for Stage 1, 2 and for the organization itself is.

Stage 1 for DMUj:

$$\max \phi_{10}$$
$$\sum_{j=1}^{n} \lambda_j x_{ij} \leq x_{i0}$$
$$\sum_{j=1}^{n} \lambda_j y_{kj} \geq y_{k0}\phi_{10} \quad (4.3\text{-}1)$$
$$\lambda_j \geq 0 \ , j = 1,2,\dots,n$$
$$i = 1,2,\dots,m,\ k = 1,2,\dots,r$$

Stage 2 for DMUj:

$$\max \phi_{20}$$
$$\sum_{j=1}^{n} \mu_j y_{kj} \leq y_k^*$$
$$\sum_{p=1}^{s} \mu_j z_{pj} \geq z_p \phi_{20} \quad (4.3\text{-}2)$$
$$\mu_j \geq 0$$

To obtain DMU_j's organizational efficiency, we need to solve

$$\max \phi_j$$
$$\sum_{j=1}^{m} \pi_j x_{ij} \leq x_i$$
$$\sum_{p=1}^{s} \mu_j z_{pj} \geq z_p \phi_j \quad (4.3\text{-}3)$$
$$\pi_j \geq 0$$

where $y_k^* = \sum_{k=1}^{r} \lambda_j^* y_k$ is the level of intermediate product r that the Stage 1 would have produced if it were efficient. The asterisks on y_k^* indicate the optimal values of these variables (Sexton & Lewis, 2003).

Other examples of two-stage Network DEA model can be found in the following studies. (Chen & Zhu, 2004) developed an efficiency model that identifies the efficient frontier of a two-stage production process linked by intermediate measures. An application to non-life insurance companies in Taiwan was delivered by (Kao & Hwang, 2008). This paper modifies the conventional DEA model by taking into account the series relationship of the two sub-processes within the whole process.

Under this framework, the efficiency of the whole process can be decomposed into the product of the efficiencies of the two sub-processes. Their overall efficiency is defined as the product of efficiencies of the two stages, their models assume constant returns to scale (CRS), and that the weights (or multipliers) on the intermediate measures are the same for the two stages. The paper of (Chen, Liang, & Zhu, 2009) examines relations and equivalence between two existing DEA approaches that address measuring the performance of two-stage processes. The majority of the studies above utilized the radial models (CCR and BCC) that stand on the assumption that inputs and outputs undergo proportional changes (Tone & Tsutsui, 2008).

The following steps involve development of a network DEA model. For an **output-oriented** model:

- first a general DEA model is solved for the upstream node at the 1^{st} stage to obtain the optimal solution of outputs,
- at the next stage, a part of (or all of) optimal outputs obtained at the upstream node are applied as intermediate inputs to the next node,
- After solving DEA models for all nodes in turn, a final optimal output is obtained at the last node.
- The firm-level efficiency score is measured as the final optimal output divided by an observed output (Tone & Tsutsui, 2008).

First attempts to apply the Network DEA in risk management appeared few years after its introduction. Matthews evaluated bank performance in risk management practices using a Network DEA approach where an index of risk management practice and an index of risk management organisation are used as intermediate inputs in the production process (Matthews, 2011).

5 DATA DESCRIPTION

This section outlines data collection, treatment, modifications and descriptive analysis of the variables included in the dataset studied.

5.1 Data collection

Reliable historical data represents a key element in successful model development. The original set of data available for analysis represents a bank portfolio of corporate customers' applications of a total of 7194 from January 2008 to December 2012 for 6759 unique entities.

The raw dataset consists of 335 variables in total, including

- general information about the company, such as industrial sector or country,
- historical balance sheet and other financial information at the time of application,
- behavioural fields looking at the payment history,
- dates,
- default indicators.

5.2 Data cleansing

The data need to be analysed prior to estimation. The data analysis primarily aims at identifying a manageable set of significant independent variables that can be further used in model building (shortlist of candidate variables). In the course of this analysis, preliminary quality checks, identification of potential structural breaks, and the correlation structure between the independent variables are considered.

A quality check needs to be performed in order to ensure that the data is accurate and reliable, as a model itself can only be as reliable as the data used for its construction. The following tests should typically be carried out:

- Plausibility checks
- Analysis of distributions
- Analysis of missing values
- Analysis of outliers

In order to assess the impact of these potential outliers, the final models were estimated on a sample in which they were removed.
Special care must be taken when editing data so that you do not alter or throw out responses in such a way as to bias your results.

5.3 Data exclusions and modifications

Starting with data quality issues, we need to check for duplicate records, missing values, misspelling of names or other character values, incorrect values such as negatives for exposures. Data quality is a concern in every bank or institution. The data set used for this analysis is not an exception. There are several reasons. The loan applicant's financial accounts don't necessarily reflect the reality and might be subject to data quality issues. The input data process within the bank is usually a complex system of IT and manual procedures. Despite audits and internal controls, it is exposed to a certain level of data inaccuracy.

5.3.1 Duplicates

The first step is to check for duplicate records. It can happen that with all the merging, linking and setting tables together, the record can be duplicated. These duplicates have to be identified and deleted from the dataset. SAS allows for easy identification of duplicates even within such a big set of data. The duplicates are then removed so that only one record is kept for analysis. In this case, no duplicates were found. All 7194 observations were found to be unique.

5.3.1 Extreme values

The data quality does not concern only missing values, but as well extreme observations or outliers[15]. An outlier may indicate a sample peculiarity or may indicate a data entry error or other problem. To track those errors through complex input data process is a difficult and expensive task; therefore in most banks a certain level of inaccurate data level is accepted, if immaterial and in accordance with regulations.

However in regression, outliers need to be identified and sometimes removed as they usually have a higher influence on the regression than the rest of the data. In ordinary least squares regression, measures of influence (such as Cook's D[16]) help to determine whether individual cases have undue impact on the fitted regression model and the coefficients of the predictors. The best strategy is to look at the distribution of Cook's D values and see whether there are any conspicuously large values relative to the others.

The 1st and 99th percentile was calculated to identify extreme values. Any extreme values will be removed from the modelling dataset, using a cap level on the selected variables.

[15] an outlier is an observation that is numerically distant from the rest of the data; typically points further than three or four standard deviations from the mean are considered as "outliers"
[16] See (Cook, 1977) for details

5.3.2 Missing values

The problem comes up with missing values – the density of their occurrence does not allow to simply omitting the entries with missing values. There are several options to deal with missing values:

a) Dropping cases with missing values, but this dramatically reduce the number of records and quality of observation data
b) Excluding indicators with missing values, if the percentage of the missing values is more than a certain percentage
c) Coding missing values as an additional attribute
d) Substituting missing values with estimated values, depending on the nature of the values

For the purpose of this thesis, two of the above methods have been applied.

a) the financials showing more than 30% of missing values, a total of 221 variables have been excluded from the dataset
b) blank entries of the other financials, showing less than 30% of missing values, have been substituted with an estimate

Financials with a high portion of blank entries indicate limited availability of the data and will not be a reliable predictor for the models to be built. Therefore, these variables need to be removed from the data set.

Percentage of missing values	Number of variables
Higher than 30%	221
Lower than 30%	116

TABLE 4: MISSING VALUES (SOURCE: OWN CALCULATIONS)

The remaining fields do have more than 70 % of values populated, but we still need to replace the missing values. Methodologies that will be applied do not handle missing values and would remove all records with a missing value. These can be replaced with:

a) Zero – assuming that missing value is equal to zero value,
b) Mean - the presence of extreme values causes that the mean is not a suitable measure for the estimate to replace missing values. The further are located the extreme values, the more influence they have on the calculation of the mean,
c) Median, inversely to the mean, is not influenced by the outliers and represents a measure of the location separating the data into 2 halves of 50 %.

Table 5 shows the number of customers and records excluded from the original dataset. Outliers will only be excluded if shown that it improves the quality of the model.

Exclusions	Customers	Records	Fields
Raw data	6759	7194	337
Duplicates	0	0	0
Exclusions due to missing data	0	0	221
Data after exclusions	**6759**	**7194**	**116**

TABLE 5: EXCLUSIONS (SOURCE: OWN CALCULATIONS)

5.4 Selection of variables

Depending on the nature of the analytic problem, the selection of variables may involve anything between a simple choice of predictors to elaborate more complex analyses using a wide variety of statistical and graphical methods in order to identify the most relevant variables.
A number of approaches to selecting characteristics are commonly used:

a) expert knowledge, experience and feeling for the data and characteristics provides a good complement to the formal statistical manipulations,
b) factor analysis,
c) principal component analysis,
d) multi-collinearity,
e) cluster analysis,
f) stepwise statistical procedures (for ex. Forward stepwise methods sequentially add variables, at each step adding that variable which leads to the greatest improvement in predictive accuracy
g) selecting individual characteristics by using a measure of the difference between the distributions of the good and bad risks on that characteristic. One common such measure is the information value, defined as

$$IV_i = \sum_j (p_{ij} - q_{ij}) \ln\left(\frac{p_{ij}}{q_{ij}}\right), = 1,2,\dots,m\ , j = 1,2,\dots,n \qquad (5.4\text{-}1)$$

where p_{ij} is the number of good risks in attribute *j* of a characteristic *i* divided by the total number of good risks and q_{ij} is the number of bad risks in attribute *j* of a characteristic *i* divided by the total number of bad risks. Typically any characteristic with an information value of over 0.1 will be considered for inclusion in the scorecard. Another common measure is the χ^2statistic derived from a cross tabulation of class (good or bad) by the attributes of the characteristic in question. From the perspective of multivariate[17] statistics, such an approach has obvious shortcomings. (Hand & Henley, 1997)

[17] Multivariate analysis includes all sort of statistical methods that simultaneously analyse multiple measurements on individual objects under investigation. The objective is to predict the changes in the dependent variable using the changes in the independent variables.

For the purpose of this thesis, the expert opinion and the cluster analysis were used to facilitate the first level of selection of the modelling variables. Further selection will be performed to identify the most powerful variables for each of the methodology: stepwise selection for the logistic regression and linear discriminant, information value ranking for DEA.

5.4.1 Expert opinion selection

The expert knowledge selection is the first method to identify the appropriate variables to be included into the models. All variables are considered in terms of their predictive power, intuitiveness, whether they provided new information, whether they are applicable to the whole portfolio and other criteria. In some cases, it is also beneficial to further transform the variables by elementary mathematical function, such as logarithm, exponential function, polynomial functions (e.g. x^2), inverse function ($1/x$), etc. These transformations should be considered when the independent and dependent variable exhibit a non-linear relationship that can be remediated by one of these functions. This step is part of the performance analysis of the final models. Based on an expert knowledge, 42 fields were identified as irrelevant and will be excluded from the modelling variable set.

Exclusions	**Fields**
Dataset after exclusions	116
Irrelevant data	74
Final Modelling Variable Set	**42**

TABLE 6: EXPERT SELECTION (SOURCE: OWN CALCULATIONS)

5.4.2 Testing for Multi-collinearity

Independent variables can be mutually highly correlated and therefore contain the same information. This phenomenon is known as multi-collinearity. A model including combinations of multi-collinear variables may exhibit non-robust statistical properties where the sensitivity of the estimates are highly volatile and less precise than if estimated on uncorrelated independent variables.

These problems include the following characteristics (Greene, 2003):

- Small changes in the data produce wide swings in the parameter estimates
- Coefficients may have very high standard errors and low significance levels even though they are jointly significant and the Coefficient of Determination R^2 for the regression is quite high
- Coefficients may have incorrect signs or implausible magnitudes

In order to avoid such problems, highly correlated variables must be identified and should not enter the final model in combination. Table 7 shows partial results from the multi-collinearity analysis. High correlation can be defined as exceeding 80% (positive or negative).

Pearson Correlation Coefficients	Delivered Risk Indicator	Number of Directors Resigned Last Year	Number of Current Directors
Delivered Risk Indicator	1	-0.1763	-0.3522
		<.0001	<.0001
Number of Directors Resigned Last Year	-0.1763	1	0.4035
	<.0001		<.0001
Number of Current Directors/Owners	-0.3522	0.4035	1
	<.0001	<.0001	

TABLE 7: CORRELATIONS BETWEEN VARIABLES (SOURCE: OWN CALCULATIONS)

We can support our results from the multi-collinearity test by cluster analysis.

5.4.1 Cluster analysis

Cluster analysis is a collection of statistical methods used to identify groups of samples that behave similarly or show similar characteristics while the groups or clusters are dissimilar to each other. Cluster analysis differs fundamentally from classification analysis. In classification analysis, we allocate the observations to a known number of predefined groups or populations. In cluster analysis, neither the number of groups nor the groups themselves are known in advance. Two common approaches to clustering the observation vectors are hierarchical clustering and partitioning. In hierarchical clustering we typically start with n clusters, one for each observation, and end with a single cluster containing all n observations. At each step, an observation or a cluster of observations is absorbed into another cluster. We can also reverse this process, that is, start with a single cluster containing all n observations and end with n clusters of a single item each. In partitioning, we simply divide the observations into g clusters. This can be done by starting with an initial partitioning or with cluster centres and then reallocating the observations according to some optimality criterion. (Rencher, 2002)

SAS can be used to facilitate clustering of variables. The VARCLUS procedure divides a set of numeric variables into either disjoint or hierarchical clusters. Associated with each cluster is a linear combination of the variables in the cluster, which may be either the first principal component[18] or the centroid component[19]. If the cluster components are centroid components of the covariance matrix, each subtest score is simply the sum of the item scores for that cluster. (SAS Institute Inc., 1999)

The VARCLUS procedure was run on a set of 42 selected variables and 26 clusters were found with the following structure. Better results were achieved using the centroid option.

[18] The first principal component is a weighted average of the variables that explains as much variance as possible.
[19] Centroid components are non-weighted averages of the variables

This method seeks for a hierarchy of clusters by starting with one cluster and splitting it into a hierarchy until each cluster has only a single eigenvalue greater than one, thus satisfying the most popular criterion for determining the sufficiency of a single underlying factor dimension. This is called the divisive approach.

Table 8 displays the R2 value of each variable with its own cluster and the R^2 value with its nearest cluster. It gives the squared correlation of the variable with its own cluster. The larger the value, the better it fits into the cluster. The R^2 value for a variable with the nearest cluster should be low if the clusters are well separated. The last column displays the ratio of 1 - R^2 own /1 - R^2 nearest for each variable. Small values of this ratio indicate good clustering. It can be seen that the cluster analysis supports the results of the multi collinearity analysis carried out before.

26 Clusters		**R-squared with**		**1-R**2**
Cluster	Variable	Own Cluster	Next Closest	Ratio
Cluster 1	Total Receivables	0.8877	0.2686	0.1535
	Cash in bank and hand	0.6137	0.3128	0.5621
	Current Assets	0.9146	0.1707	0.103
	Total Assets	0.9236	0.3824	0.1236
	Creditors Short Term	0.8355	0.1412	0.1916
	Total Liabilities	0.9242	0.381	0.1224
	Working Capital	0.3805	0.1361	0.7171
Cluster 2	Delivered Risk Indicator	1	0.1319	0
Cluster 3	Number of Accounts Placed for Collections in L12M	0.7447	0.0278	0.2625
	Number of Accounts Placed for Collections in L36M	0.8851	0.0186	0.1171
	Number of Accounts Placed for Collections in L60M	0.8951	0.0315	0.1083
	Number of Accounts Placed for Collections in L84M	0.8593	0.0305	0.1452
Cluster 4	Calculated Raw Score	0.9597	0.1985	0.0503
	Calculated Risk Indicator	0.8691	0.1568	0.1552
	Percentile Failure Score	0.7048	0.4587	0.5454
Cluster 5	Debt to Worth Ratio	1	0.044	0
Cluster 6	Sic Code	1	0.1754	0
Cluster 7	Percentage Trade Paid 60	1	0.0141	0
Cluster 8	Current Ratio	1	0.0016	0
Cluster 9	Number of Directors/Owners Resigned Last Year	1	0.2133	0
Cluster 10	Tangible Assets	1	0.1333	0
Cluster 11	Number of Directors/Owners Holding Shares	1	0.0504	0
Cluster 12	Financial Strength Indicator	1	0.0509	0
Cluster 13	Financial Assets	0.7828	0.1574	0.2577
	Fixed Assets	0.7264	0.1975	0.3409
	Net Worth	0.8155	0.3155	0.2695
	Tangible Net Worth	0.8264	0.3093	0.2514
	Cash Flow	0.6775	0.392	0.5304
Cluster 14	Quick ratio	1	0.0017	0
Cluster 15	Net Profit Loss	1	0.049	0
Cluster 16	Number of subsidiaries	1	0.0915	0

Cluster 17	Number of Current Directors/Owners	1	0.2424	0
Cluster 18	Issued Capital	0.9942	0.1282	0.0067
	Paid up Capital	0.9942	0.1287	0.0067
Cluster 19	Time since start-up in days	1	0.1008	0
Cluster 20	Solvency Ratio	1	0.044	0
Cluster 21	Legal form code	1	0.0889	0
Cluster 22	Variance in Tangible Net Worth	1	0.2423	0
Cluster 23	Number of Employees	1	0.1333	0
Cluster 24	Current Paydex Score	1	0.2318	0
Cluster 25	Number of Current Directors Appointments L12M	1	0.2133	0
Cluster 26	Number of bank accounts	1	0.2424	0

TABLE 8: CLUSTER STRUCTURE (SOURCE: OWN CALCULATIONS)

As said before, when creating the groups of variables with similar characteristics, the groups themselves have to be dissimilar. The inter cluster correlations show how big relationship is between the clusters. The highest value is between cluster 1 and cluster 13, having a value of 0.5657.

Cluster	**1**	**2**	**3**
1	1	-0.0563	0.0893
2	-0.0563	1	0.06686
3	0.0893	0.0669	1
4	0.0384	-0.2861	0.0105
5	0.0200	-0.0187	0.0022
6	-0.0239	0.0889	-0.0338
7	0.0315	0.0172	0.0771
8	-0.0007	0.0150	-0.0015
9	0.0549	-0.1763	0.0204
10	0.2119	-0.0733	0.1688
11	-0.0183	-0.0839	0.0023
12	-0.0061	0.1647	0.0482
13	**0.5657**	-0.0795	0.0821

TABLE 9: INTER-CLUSTER CORRELATIONS (SOURCE: OWN CALCULATIONS)

The total variation explained (38.09) gives the sum of the explained variation over all clusters. The final proportion (0.907) represents the total explained variation divided by the sum of cluster variation. This value, 0.907, indicates that about 91% of the total variation in the data can be accounted for by the 26 clusters.

Number of Clusters	Total Variation Explained by Clusters	Proportion of Variation Explained by Clusters	Minimum Proportion Explained by a Cluster	Minimum R-squared for a Variable	Maximum 1-R**2 Ratio for a Variable
1	4.9880	0.1188	0.1188	0.0004	
2	9.5893	0.2283	0.0893	0.0007	1.0198
3	12.5005	0.2976	0.0845	-	1.0139
4	14.8894	0.3545	0.0937	0.0340	1.1231
5	16.2837	0.3877	0.1302	0.0945	1.0553
6	17.4570	0.4156	0.1675	0.1059	0.9845
7	18.5463	0.4416	0.2491	0.1059	0.9463
8	19.5581	0.4657	0.2751	0.1059	0.9463
9	21.3435	0.5082	0.2852	0.1182	0.9090
10	22.6456	0.5392	0.4121	0.1457	0.8599
11	23.6704	0.5636	0.4613	0.1457	0.8599
12	24.5847	0.5853	0.4785	0.1457	0.8599
13	27.1490	0.6464	0.5002	0.3076	0.7844
14	28.1486	0.6702	0.5039	0.3076	0.7844
15	29.1409	0.6938	0.5077	0.3076	0.7844
16	29.9828	0.7139	0.5213	0.3076	0.7844
17	30.8517	0.7346	0.5456	0.3076	0.7844
18	32.5457	0.7749	0.5489	0.3805	0.7152
19	33.4480	0.7964	0.6049	0.3805	0.7152
20	34.2382	0.8152	0.6491	0.3805	0.7152
21	34.9401	0.8319	0.6784	0.3805	0.7152
22	35.6932	0.8498	0.6826	0.3805	0.7171
23	36.3281	0.8650	0.7012	0.3805	0.7171
24	37.0467	0.8821	0.7309	0.3805	0.7171
25	37.5848	0.8949	0.7462	0.3805	0.7171
26	38.0924	0.9070	0.7647	0.3805	0.7171

TABLE 10: CLUSTER ANALYSIS SUMMARY (SOURCE: OWN CALCULATIONS)

The structure of the clusters and the proportions of variance explained by the clusters can be pictured in a tree graph.

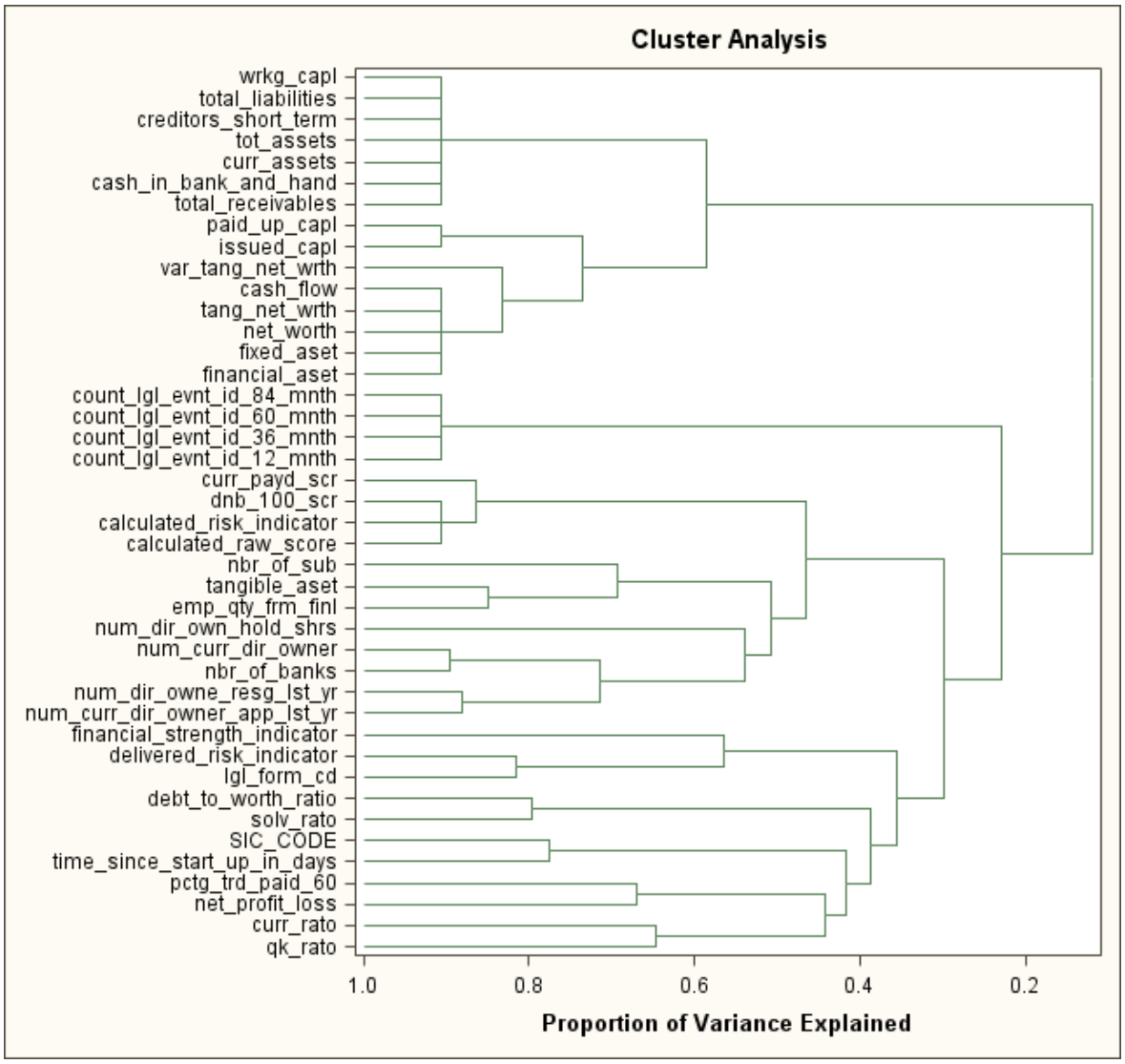

FIGURE 11: CLUSTER DENDOGRAM (SOURCE: OWN CALCULATIONS)

Looking from left to right in the diagram, objects and clusters are progressively joined until a single, all-encompassing cluster is formed at the right (or root) of the diagram. Clusters exist at each level of the diagram, and every vertical line connects leaves and branches into progressively larger clusters.

5.4.2 Final selection of variables

Exclusions involved variables having too high level of missing values or incorrect values (more than 30%), irrelevant or those being dependent or used for a calculation of another variable. The final data set that will be considered for model development includes 42 application model variables, 10 behavioural variables and other general fields and dates.

Further selection of the variables will be performed using the stepwise statistical approach and information values method at a later stage of model estimation. Outliers have been identified for several variables and special care will be taken when building the models. A robustness check will be performed by estimating the model in which the outliers are removed.

5.5 Data summary

We have two sets of data that can be used to build both types of credit scoring models. Application models will be based on the selected list of 42 model variables. These will be used to compare the performance of the standard methods, logistic regression and discrimination analysis, and the DEA models. For the Network DEA, we will need the two datasets, application and behavioural elements.

Application data: the following table contains the final list of variables to enter application model considered for the application model build.

Application Model Variables	
Time since start-up in days	Variance in Tangible Net Worth
Legal form code	Current Paydex Score
Number of subsidiaries	Percentage Trade Paid 60
Number of banks	Number of Accounts Placed for Collections L12M
Issued Capital	Number of Accounts Placed for Collections L36M
Paid up Capital	Number of Accounts Placed for Collections L60M
Calculated Raw Score	Number of Accounts Placed for Collections L84M
Calculated Risk Indicator	Sic Code
Delivered Risk Indicator	Quick ratio
Percentile Failure Score	Current Ratio
Financial Strength Indicator	Solvency Ratio
Total Liabilities	Debt to Worth Ratio
Financial Assets	Number of Current Directors Appointments L12M
Creditors Short Term	Number of Directors Resigned Last Year
Number of Directors Holding Shares	Number of Current Directors
Current Assets	Net Profit Loss
Tangible Assets	Tangible Net Worth
Fixed Assets	Working Capital
Total Receivables	Cash Flow
Cash in bank and hand	Number of Employees
Net Worth	Total Assets

TABLE 11: APPLICATION MODEL VARIABLES

Behavioural data: the following 10 variables represent the scores for answers on the selected questions. The questions are aimed to assess the behavioural of the obligor, to identify any adverse trends.

No	Behavioural Model Variables
1	Number of payments in last 12 months
2	Number of prompt payments in last 12 months
3	Number of late payments 1 to 30 days in last 12 months
4	Number of late payments 31 to 60 days in last 12 months
5	Number of late payments 61 to 90 days in last 12 months
6	Number of late payments 91 to 120 days in last 12 months
7	Number of late payments 121 to 180 days last 12 months
8	Number of late payments 180 days or worse in last 12 months
9	Total current Overdue balance
10	Total average Overdue balance

TABLE 12: BEHAVIOURAL VARIABLES

5.6 Portfolio profile

Descriptive statistics indicate the availability, central tendency and variability of the model variables.

Variable Name	Minimum	Mean	Maximum	Std Dev
Calculated Raw Score	0	1369	1913	449
Calculated Risk Indicator	0	999	1104	320
Cash Flow (000's)	-159000	725	624877	11258
Cash in bank and hand (000's)	-377	1562	3659406	45304
Creditors Short Term (000's)	-976	7465	11803187	166314
Current Assets (000's)	-140	9655	12558876	184917
Current Paydex Score	0	42	86	36
Current Ratio	-2	186	1319643	15559
Debt to Worth Ratio	-2114	4	3706	68
Delivered Risk Indicator	1	2	5	1
Financial Assets (000's)	-916	3378	2313946	53592
Financial Strength Indicator	0	1346	5453	1213
Fixed Assets (000's)	-902	6008	2313946	62717
Issued Capital (000's)	0	1604	1200252	20452
Legal form code	0	3117	20672	3145
Net Profit Loss (000's)	-3214417	118	624877	39447
Net Worth (000's)	-91922	3822	2095820	44745
Number of Accounts Placed for Collections L12M	0	0	12	0
Number of Accounts Placed for Collections L36M	0	0	32	1
Number of Accounts Placed for Collections L60M	0	0	33	1
Number of Accounts Placed for Collections L84M	0	0	33	1
Number of bank accounts	0	1	9	1
Number of Current Directors/Owners	0	3	17	3
Number of Current Directors/Owners	0	0	28	1
Number of Directors/Owners Holding Shares	0	0	30	1
Number of Directors/Owners Resigned Last Year	0	0	3	1
Number of Employees	0	36	19833	346
Number of subsidiaries	0	1	77	3

Paid up Capital (000's)	0	1468	1200252	19612
Percentage Trade Paid 60	0	4	100	11
Percentile Failure Score	0	64	100	35
Quick ratio	-2	3	1000	25
Sic Code	111	6379	9999	2327
Solvency Ratio	-45875	192	114819	2238
Tangible Assets (000's)	0	2048	693247	20536
Tangible Net Worth (000's)	-740524	4175	2095820	51965
Time since start-up in days	0	8231	85008	10127
Total Assets (000's)	-344	15699	12741486	210287
Total Liabilities (000's)	-21105	15602	12741486	210175
Total Receivables (000's)	-144	4325	4188821	62041
Variance in Tangible Net Worth (000's)	-1492489	600	1608498	33475
Working Capital (000's)	-619532	1715	2117391	34317

TABLE 13: DESCRIPTIVE STATISTICS (SOURCE: OWN CALCULATIONS)

Scoring models are built based on bank's experience of defaulting customers across different industrial sectors and business types.

Defining the dependent variable means defining the default. Default risk is the uncertainty regarding the borrower's ability to pay its obligations. Under Basel II, a default is considered to have occurred with regards to a particular obligor either or both of the following two events have taken place:

- Unlikely to pay: The bank considers the entity is unlikely to repay its debts in full,
- 90dpd: An entity is more than 90 days past due on any of its facilities.

The percentage of the nonperforming units has changed significantly over the 5 years. In 2008, the recession showed a higher number of defaults, with a bad rate of 3%. The situation didn't change the following year and the bad rate still touching 3%. Only in 2012, the percentage fell to 0.9%, indicating a healthier economy. In total, the analysed portfolio shows a 2.9% default rate.

Year	Bad Rate
2008	3.0%
2009	3.0%
2010	2.5%
2011	2.0%
2012	0.9%
Total	2.9%

TABLE 14: DEFAULT RATE BY YEAR (SOURCE: OWN CALCULATIONS)

The tendency can be better seen if represented graphically. The number of defaulted customers has been decreasing since 2008. It's not unexpected given the recession of 2008 and the consequences of the difficult economic conditions on companies as well as individuals.

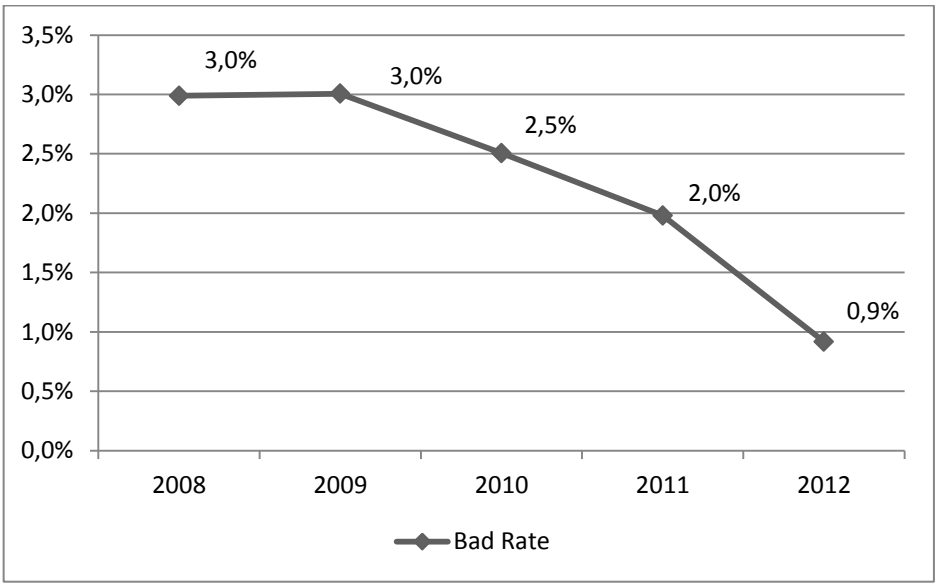

FIGURE 12: DEFAULT RATE TREND (SOURCE: OWN CALCULATIONS)

Credit scoring models are built to rank customers based on characteristics that make default more or less likely. Customers with similar characteristics are grouped into grades. Based on bank's own internal observation of how often customers with that grade default we can map each grade to a probability of default (PD), the probability that a customer will default within one-year. A PD of 0.5% means we would expect 1 in 200 similarly graded accounts to default within a year. So using internal default observations these percentages are mapped to PD bands/grades. The Probability of Default associated with each grade will stay stable over time. As credit quality for individual customers change, their grades will move up and down. This is called grade migration and is usually managed through a built behavioural model.

Since the PD is a continuous variable, taking values between 0 and 1, there are infinitely many possible ways to partition the 0-1 interval into a set of discrete intervals (the PD-buckets). The choice of the "optimal" buckets (sometimes referred to as "PD bucketing") is seldom tackled analytically by banks. Financial institutions often rely on a purely qualitative definition of the rating buckets (e.g., by defining labels like "excellent" or "AAA" and a set of rating criteria which help their analysts to sort obligors into different classes). (Krink, Paterlini, & Resti, 2008)

Organizing actual PDs into grading scales makes comparisons between counterparties easier. The grading scale in Table 15 proposes 16 bands, the smaller the grade, the less risky the customer is. At a PD grade of 16, there is 100% probability that the customer will default in the next 12 months.

PD Grade	PD Band (Low)	PD Band Mid-Point	PD Band (High)	PD Band Width
1	0.00%	0.02%	0.07%	0.07%
2	0.08%	0.10%	0.12%	0.04%
3	0.13%	0.16%	0.21%	0.08%
4	0.22%	0.27%	0.35%	0.13%
5	0.36%	0.45%	0.58%	0.22%
6	0.59%	0.74%	0.96%	0.37%
7	0.97%	1.24%	1.60%	0.63%
8	1.61%	2.06%	2.66%	1.05%
9	2.67%	3.46%	4.43%	1.76%
10	4.44%	5.72%	7.39%	2.95%
11	7.40%	9.54%	12.31%	4.91%
12	12.32%	15.89%	20.51%	8.19%
13	20.52%	26.47%	34.17%	13.65%
14	34.18%	44.10%	56.92%	22.74%
15	56.93%	73.48%	99.99%	40.06%
16	100.00%	100.00%	100.00%	0.00%

TABLE 15: GRADING SCALE (SOURCE: OWN CALCULATIONS)

The band width increases exponentially with each grade. It is to provide a higher granularity to the low PD grades.

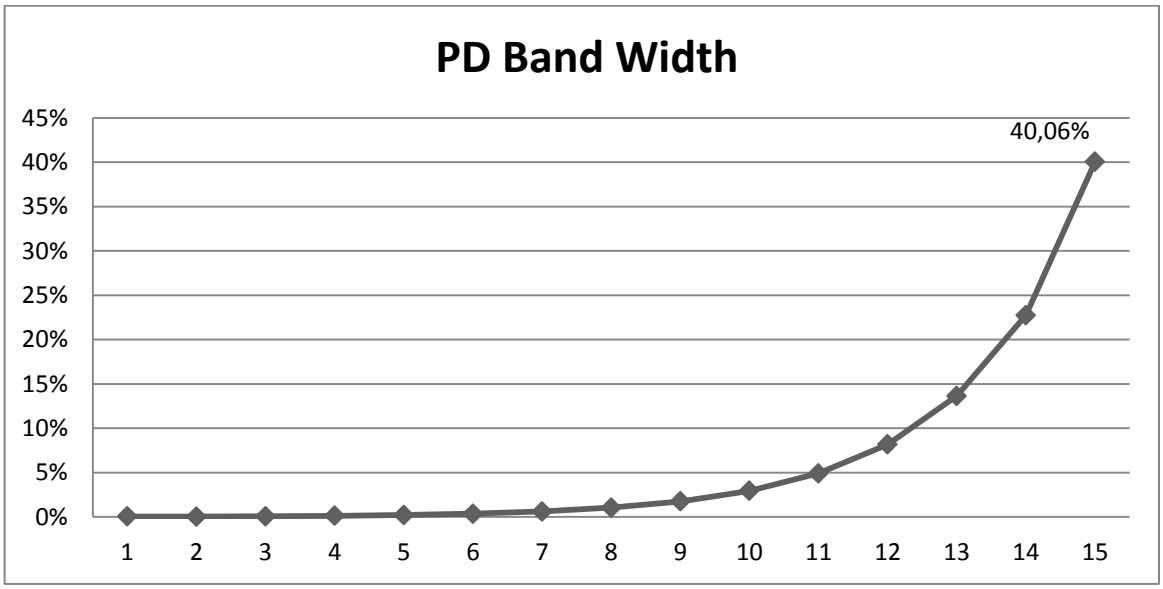

FIGURE 13: PD BAND WIDTH (SOURCE: OWN CALCULATIONS)

The portfolio is diversified across sectors. The biggest portion of the portfolio operates in services (35.61%) The second highest number of entities operates in manufacturing (15.71%).

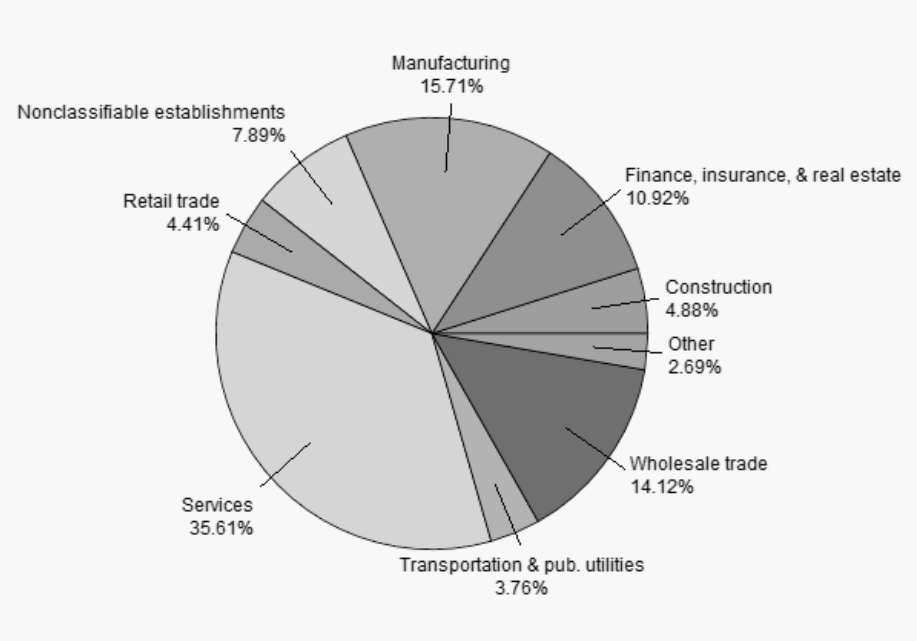

FIGURE 14: NUMBER OF ENTITIES BY SECTOR (SOURCE: OWN CHART)

As can be seen in Table 16, 16.96% of the portfolio is of a small or medium size. Only the remaining 4% do not fall into the usual SME definition.

Size	Frequency	Percent	Cumulative Frequency	Cumulative Percent
PLE	301	4.18	301	4.18
SME	6893	95.82	7194	100.00

TABLE 16: SIZE OF THE COMPANIES (SOURCE: OWN CALCULATIONS)

As defined in EU law, the category of micro, small and medium enterprises (SMEs) is made up of enterprises which employ fewer than 250 persons and which have an annual turnover not exceeding EUR 50 million, and/or an annual balance sheet total not exceeding EUR 43 million (Commission of the EU, 2003). Figure 16 shows that 95.82% of the portfolio is represented by SME customers.

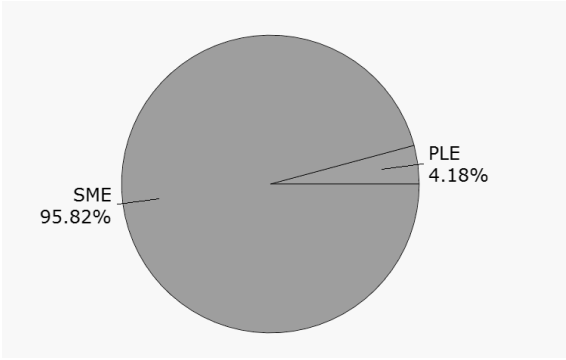

FIGURE 15: SME BOOK (SOURCE: OWN CHART)

6 LOGISTIC REGRESSION APPLICATION

6.1 Low default portfolio and selection of data

The difficulty here is the low ratio of default/non default. When the logistic regression doesn't have a big enough representation of the defaulted cases, the modelling becomes challenging. Cramer (Cramer, 2000) clearly points out that most techniques lose performance and result in low quality estimates modelling a rare event. The distinction between good and bad loans (or debtors) is used in a statistical analysis to link the probability of a loan going bad to the initial financial ratios of the debtor. The estimated risk is then used to score all loans and to establish a classification.

Logistic regression is not always the best choice for modelling credit grades when it comes to a low default portfolio. Low default portfolios are those for which banks have little default history, so that average observed default rates might not be reliable estimators of default probabilities (PDs). A key concern for regulators is that credit risk might be underestimated as a result of data scarcity. (Benjamin, Cathcart, & Ryan, 2006)

Low default portfolios are the ones where:

- Historically have experience a low number of defaults and are considered as low risk (such as banks, insurance companies, highly rated firms)
- Lack of historical data
- Low number of counterparties
- May have not incurred recent losses

The data set contains 7194 applications of which 6988 were classified as 'good' and 206 were classified as 'bad'. The overall sample bad rate is 2.9%.

Observations	7194
Defaults	206
Default Rate	2.9%

TABLE 17: DEFAULTS (SOURCE: OWN CALCULATIONS)

About 3 % default rate is very low, but with a good bad definition, the model will be robust enough. Given the low number of defaults, there is no room for dividing the data set into training and validation (called also out of sample) sets. The whole portfolio will be used to build the model on.

The regression model is constructed on a training set. Given there is not a sufficient representation of defaults within the dataset, we cannot divide the dataset into training and test set and we will have to use other methods to validate the model, such as cross validation (as described in section 3.4.2).

The optimal parameters are estimated by minimizing the squared difference between predicted and observed values from the training set.

6.2 Logistic model

The model was developed from internal data covering credit applications from January 2008 to December 2012 for 6759 unique entities. Some entities applied multiple times.

6.2.1 Assumptions

Unlike ordinary linear regression, logistic regression does not assume that the relationship between the independent variables and the dependent variable is a linear one. Nor does it assume that the dependent variable or the error terms are distributed normally. It does not require that the independents be interval or be unbounded. You can add explicit interaction and power terms. On the other hand, given this flexibility, it requires a bigger base of data to achieve stable and meaningful results.

The archive data are at application date or before. The performance indicators are at 12-18months after the application date to enable us to identify the bad customers, while modelling on the application data that is available at the time of application.

Bad definition consists of a few legal events that indicate the company is in bad shape. The list of events that define a company in default is shown in Table 18.

Legal events indicating default
Debtor
Liquidated Company
Moratorium
Debt Purge Terminated – Approval of the Offered Composition
Annulment of Bankruptcy After Objection
Annulment of Suspension by Expiration of the Agreed Term
Bankruptcy After Withdrawal of Suspension of Payment
Bankruptcy Terminated – Approval of the Offered Composition
Close Bankruptcy Proc. – Binding of the List of Creditors
Close the Bankruptcy Proceedings due to Lack of Assets
Liquidation Re-opened by Bankruptcy
Execution Sale
Suspension Terminated – Approval of the Offered Composition
Withdrawal of Suspension of Payment
Suspension of payment extension
Notification of meeting of creditors
Bankruptcy request
Bankruptcy
Annulment of dissolution by court

Seizure by tax authorities
Guardianship
Annulment of Guardianship
End Guardianship
Seizure by creditor
Dissolution by court
Debt purge
Withdrawal of Debt Purge
Debt Purge request
Company bankruptcy ended as principals received debt purge
Termination of bankruptcy by applying debt purge
Annulment of Debt Purge by Expiration of the Agreed Term
Request for Debt Purge not accepted
Debt Purge ended
Suspension of Payment after Withdrawal of Debt Purge
Bankruptcy after Withdrawal of Debt Purge
Suspension of Payments, Peremptory
Suspension of Payments, Provisional
Debt Purge, Provisional
Debt Purge, Peremptory

TABLE 18: BAD FLAG DEFINITION

The defaulting customers are flagged as 1, using bad flag as an indicator and dependant variable. The dependent variable is binary and that is the main reason why logistic regression has been used. The logistic regression can handle binary dependant variables and outputs the probability of the modelled value.

The Bad flag can only have two values. It is equal to 1 if the unit is in default and equal to 0 if the unit is performing. The bad rate is 2.9%. The other way of getting the default rate is to run the logistic regression without any predictor variables a we get the probability through the equation using the intercept value. The intercept value is the log of the odds of being in default.

6.2.2 Selection of variables

Data has been cleaned from any data quality issues and variables selected based on expert opinion. The SAS procedure offers another option for selecting variables. This can be done through the stepwise, forward or backward selection methodology. If this doesn't come up with satisfactory results, cluster analysis may be considered as additional selection method.

The stepwise option removed all but 3 variables that ended up in the final selection as yielding the most predictive combination. The threshold level for the variables to enter and remain in the model using the stepwise procedure was a p-value of 0.05.
The proposed set of 3 variables consists of:

1. **Delivered Risk Indicator** is a banded representation of the failure score. The Failure Score predicts the likelihood that a company will obtain legal relief from its creditors or cease operations over the next 12 month period. The Failure scorecard looks for the onset of failure such as meeting of creditors, administrator appointed, bankruptcy, receiver appointed, petition for winding-up among others legal events.
2. **Number of Current Directors** is the number of owners or directors that a company has. This indicates that the more the directors a company has, the more secure the company has i.e. the more directors indicates the size of the company, where small companies are riskier in general.
3. **Number of Resigned Directors L12M** is the number of directors of a company which have resigned in the last 12 months. High number of resignations indicates the company may have underlying issues and may have payment problems due to turnover in the staff. This can be an early indicator of a potential default.

We first want to make sure that the selected model variables are powerful predictors of the default. Tables and charts below are looking at the relationship of each of the variables and the bad flag that is our indicator of a default.

Delivered Risk Indicator	N	Number of bads	Bad Rate
1	2825	33	1.2%
2	605	12	2.0%
3	2355	82	3.5%
4	1041	45	4.3%
5	368	34	9.2%
All	7194	206	2.9%

TABLE 19: DELIVERED RISK INDICATOR VS DEFAULT (SOURCE: OWN CALCULATIONS)

It can be seen that the default rate is growing with the increasing value of the risk indicator. This means that the coefficient assigned to the variable by the model has to be positive.

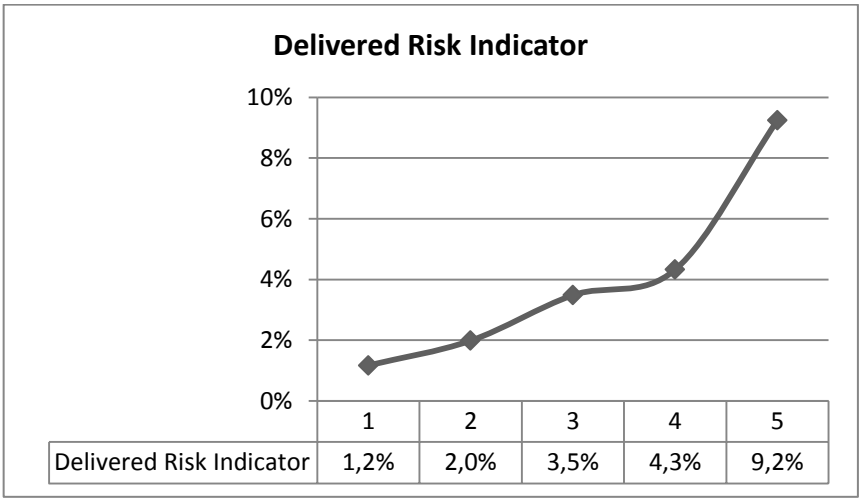

FIGURE 16: DELIVERED RISK INDICATOR VS DEFAULT TREND (SOURCE: OWN CALCULATIONS)

The same analysis is performed for the other 2 model variables. Table 20 shows that the higher the number of directors, the less chances the company will default.

Number of current directors	N	Number of bads	Bad Rate
0	1878	79	4.2%
1	1657	47	2.8%
2	1039	29	2.8%
>= 3	2620	51	1.9%
All	7194	206	2.9%

TABLE 20: NUMBER OF CURRENT DIRECTORS VS DEFAULT (SOURCE: OWN CALCULATIONS)

The decreasing trend of the default with higher number of directors is clearly shown on Figure 17.

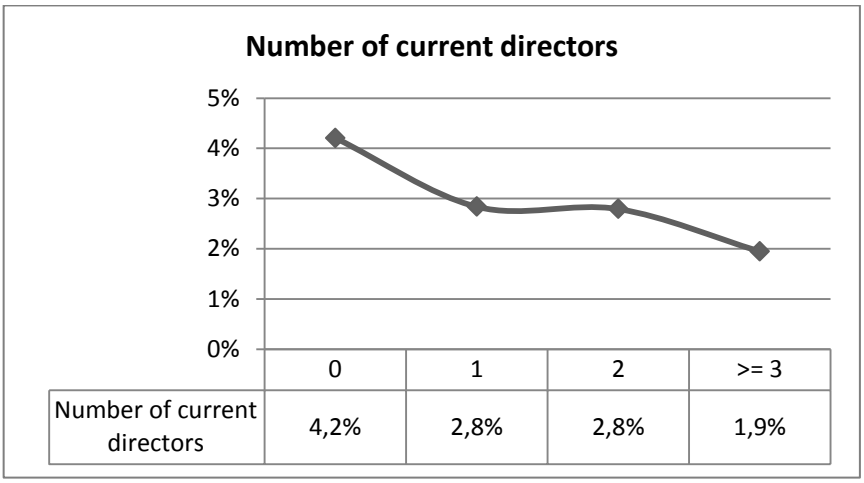

FIGURE 17: CURRENT DIRECTORS VS DEFAULT TREND (SOURCE: OWN CALCULATIONS)

High number of resigned directors is a clear sign of a difficulty within the company.

Number of directors resigned L12M	N	Number of bads	Bad Rate
0	6476	181	2.8%
1	430	15	3.5%
2	139	5	3.6%
>= 3	149	5	3.4%
All	7194	206	2.9%

TABLE 21: NUMBER OF RESIGNED DIRECTORS VS DEFAULT (SOURCE: OWN CALCULATIONS)

Increasing trend of the default rate for every additional resigned director during the last 12 months confirms positive direction of the logistic regression coefficient.

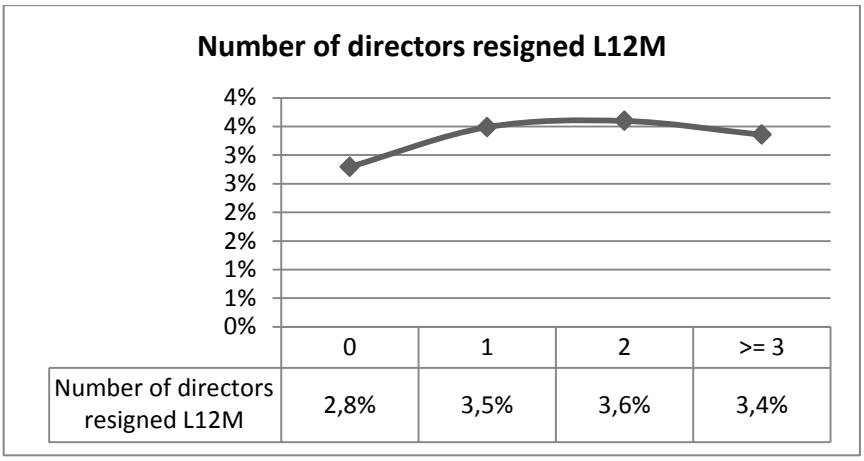

FIGURE 18: RESIGNED DIRECTORS VS DEFAULT TREND (SOURCE: OWN CALCULATIONS)

The above supports the selection of the variables of the model.

6.2.3 Logistic regression model

We are creating a model with a number of continuous predictor variables, describing the relationship between these variables and the log odds of being in default.

Here, the loans in default are the ones where bad flag = 1 (12 to 18 months after application date), but the dependent variables are taken at application date.

With logistic regression, we were looking for the predicted probabilities that a unit of the population under analysis will acquire the event of default as a linear function of the selected independent variables. Response level is descending, so the modelled level is bad flag = 1.

Response Profile		
Ordered value	**Bad flag**	**Total Frequency**
1	0	6988
2	1	206
Probability modelled is bad flag='1'.		

TABLE 22: RESPONSE PROFILE (SOURCE: OWN CALCULATIONS)

The model is a binary logit model and the maximum likelihood estimation is carried out with the Fisher-scoring algorithm[20], as described in section 3.3.

The full list of predictors is shown in the following table. Here the statistics test the null hypothesis that an individual coefficient is zero. The p values are smaller than 0.001 for all variables, therefore can be said as significant.

[20] More details on Fisher's scoring algorithm can be found in (Jennrich & Sampson, 1976)

Analysis of Maximum Likelihood Estimates						
	Parameter	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > Chi Sq**	**Standardized Estimate**
	Intercept	-4.6583	0.2219	440.7716	<.0001	
x_1	Delivered Risk Indicator	0.4689	0.0602	60.6591	<.0001	0.3287
x_2	Number of current directors	-0.1225	0.0366	11.2142	0.0008	-0.2244
x_3	Number of directors resigned L12M	0.5357	0.1312	16.6812	<.0001	0.1612

TABLE 23: PREDICTORS ESTIMATES (SOURCE: OWN CALCULATIONS)

The coefficients of the regression equation regress against the logit not the dependent variable itself. Taking the estimates of the parameters coefficients, the final equation of the logistic model is:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \qquad (6.2\text{-}1)$$

One unit change in the predictor variable will cause a change of the log odds by the respective coefficient (given the other variables stay constant). For example, for every unit change in x_2, the log odds of default decrease by 0.1225.

As with linear regression analysis, the parameter estimate can be conceptualized as how much mathematical impact a unit changes in the value of the independent variable has on increasing or decreasing the probability that the dependent variable will achieve the value of one in the population from which the data are assumed to have been randomly sampled. With all the variables equal to 0, the log-odds of default are equal to the intercept (-4.6583).

The logistic function of this model is shown below. It describes the relationship between the credit score, the logit and the predicted probability of default.

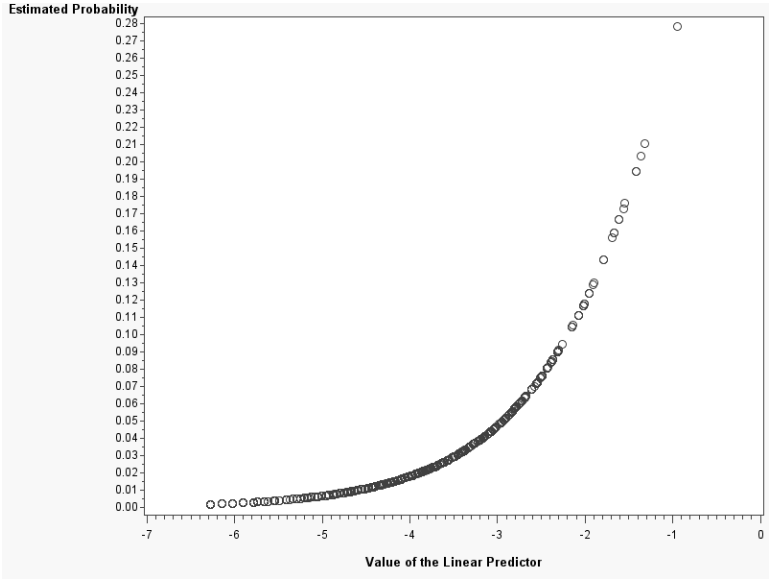

FIGURE 19: LOGISTIC FUNCTION (SOURCE: OWN CALCULATIONS)

Exponentiation of the parameter estimates for the independent variables in the model by the number e (about 2.17) yields the odds ratio, which is a more intuitive and easily understood way to capture the relationship between the independent and dependent variables.

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
Delivered Risk Indicator	1.598	1.420	1.798
Number of current directors	0.885	0.823	0.950
Number of directors resigned L12M	1.709	1.321	2.210

TABLE 24: ODDS RATIO ESTIMATES (SOURCE: OWN CALCULATIONS)

The odds ratio gives the increase or decrease in probability that a unit change in the independent variable has in the probability that the event of interest will occur. Taking the example of Number of current directors, one unit change in this variable will cause the log odds of default decrease by a factor of 0.885. The coefficient and the odds ratio provide the same information, only in two different ways.

The next step is to look at the performance and fit statistics of the model.

6.2.4 Model performance and fit statistics

There is no direct equivalent of R for logistic regression. It is the proportion of the variance in the dependent variable which is explained by the variance in the independent variables. Other statistics can be used when assessing the goodness of fit of the model.
The model convergence status describes whether the maximum-likelihood algorithm has converged or not. Table 25 shows that default criterion, the relative gradient convergence criterion (GCONV), is satisfied with the default precision of 10^{-8}.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

TABLE 25: MODEL CONVERGENCE STATUS

The output from SAS includes several tests of overall model adequacy which test the global null hypothesis that none of the independent variables in the model are related to changes in probability of event occurrence. Of these, the log-likelihood test is perhaps the most easy to interpret. The computation of and rationale for the log-likelihood test, among others, is found in Hosmer and Lemeshow (1989). The "Hosmer and Lemeshow Test" is a measure of fit which evaluates the goodness of fit between predicted and observed.

Partition for the Hosmer and Lemeshow Test					
Group	Total	Bad Flag = 1		Bad Flag = 0	
		Observed	Expected	Observed	Expected
1	641	5	3.46	636	637.54
2	795	8	7.49	787	787.51
3	974	12	12.24	962	961.76
4	651	16	11.79	635	639.21
5	805	18	19.61	787	785.39
6	476	11	14.02	465	461.98
7	717	18	23.75	699	693.25
8	716	34	26.67	682	689.33
9	673	30	32.63	643	640.37
10	746	54	54.35	692	691.65

TABLE 26: PARTITION FOR THE HOSMER AND LEMESHOW TEST (SOURCE: OWN CALCULATIONS)

We want this chi-squared value to be low and non-statistically significant if the predicted and observed probabilities match up nicely. The results support what we were looking for.

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
6.8319	8	0.5549

TABLE 27: HOSMER AND LEMESHOW GOODNESS-OF-FIT TEST (SOURCE: OWN CALCULATIONS)

The "Model Fit Statistics" table contains the Akaike Information Criterion (AIC), the Schwarz Criterion (SC), and the test statistic from the log-likelihood function (as defined in 4.3-9) for the intercept-only model and the fitted model. AIC and SC can be used to compare different models, and the ones with smaller values are preferred. The values though don't have much meaning themselves.

There are many options that one can choose from when running the model. The fit statistics that are calculated as part of the procedure help to choose the model with the best fit. Intercept only shows the value statistic with no predictor variables, just the binary response variable.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1871.932	1777.456
SC	1878.813	1804.980
-2 Log L	1869.932	1769.456

TABLE 28: MODEL FIT STATISTICS (SOURCE: OWN CALCULATIONS)

The values of the log-likelihood test statistic are usually negative, because the values that the density takes are usually smaller than 1 and its logarithm is then negative. But in our case, the distribution of the variable is different. It has a small standard deviation with density largely concentrated around 0. Obviously, this will cause that large values will be taken around this point and the logarithm positive.

The selection of variables was based on calculating the likelihood ratio. It responds to the question if including the variables explains better the model. Chi-square test was performed to assess the strength of explanation of these factors. Likelihood ratio is the most used indicator, but SAS calculates as well the Score Chi-Square and the Wald Chi-Square statistic. The difference between them is where on the log-likelihood function they are evaluated.

The DF defines the distribution of the Chi-Square test statistics and is defined by the number of predictors in the model. We have the final 3 variables selected to be kept in the model, therefore the degrees of freedom is equal to 3.

The small p-values reject the hypothesis that all slope parameters are equal to zero. Here the p values are smaller than 0.001, which means that there is at least one coefficient in the model not equal to zero.

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	100.4761	3	<.0001
Score	95.4608	3	<.0001
Wald	88.5190	3	<.0001

TABLE 29: NULL HYPOTHESIS TEST (SOURCE: OWN CALCULATIONS)

To assess the performance of the model and the level of discrimination, we can use the ROC (Receiver Operating Characteristics) curve. The larger is the area under the curve, the better the model can discriminate between the binary values (here default and non-default). The ideal situation would show a steep rise of the curve until it reaches 100%. The opposite would be a curve in the form of a straight line starting at 0 and ending at 100%. The vertical axis shows the accumulation of bads while the horizontal access shows the accumulation of total applications. The goal of the model is to select all the bads in the lowest scoring sections of the population. Therefore, if the model is powerful, it will accumulate more bads than goods as we move from left to right, and the model will have a steeper slope. A poor model will randomly select goods and bads and thus the graph will be linear.

The metric that calculates the area under the curve is also called the Gini coefficient or index. The ROC curve in Figure 20 represents the performance of our logistic model. It is also showing the different steps to reach the highest level of performance. The area under the curve is large enough to state that the model discriminates well.

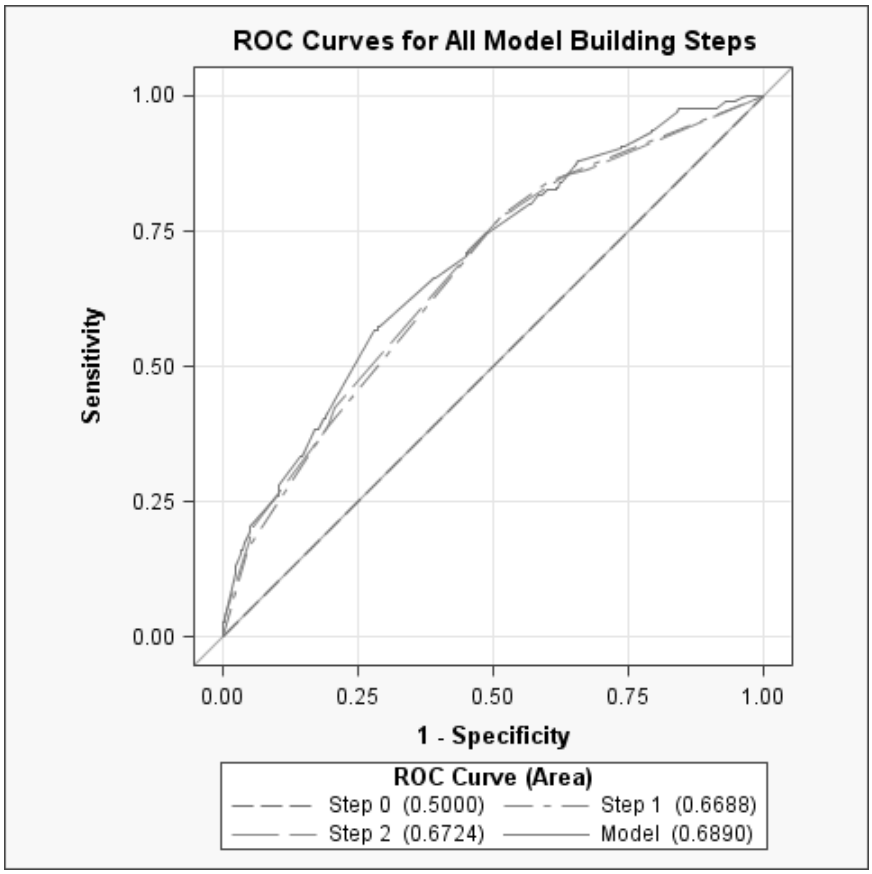

FIGURE 20: ROC CURVE (SOURCE: OWN CALCULATIONS)

Another measure of performance is used to select the right model. It is the Kolmogorov–Smirnov (KS) statistic. Same as the Gini index, it shows how good the model can discriminate between good and bad accounts. The KS metric is mainly used in the Unites States; European banks prefer to use the Gini index.

The KS test statistic shows the scorecards ability to discriminate between goods (those customers that have excellent performance history) and bads (Defaults/Charge-offs). The KS statistic is closely related to the ROC curve. It looks at the maximum vertical distance between the cumulative distribution function of the fitted distribution and the cumulative distribution of the data. On the graph, it would be the maximum difference between the linear 45 degree line and the ROC curve.

In Figure 21, the KS score of the proposed model is compared to that of two credit bureau scores shown to be powerful risk discriminators in their own right. It can be observed that the proposed model is significantly more powerful than both the Failure Score and the Paydex Score.

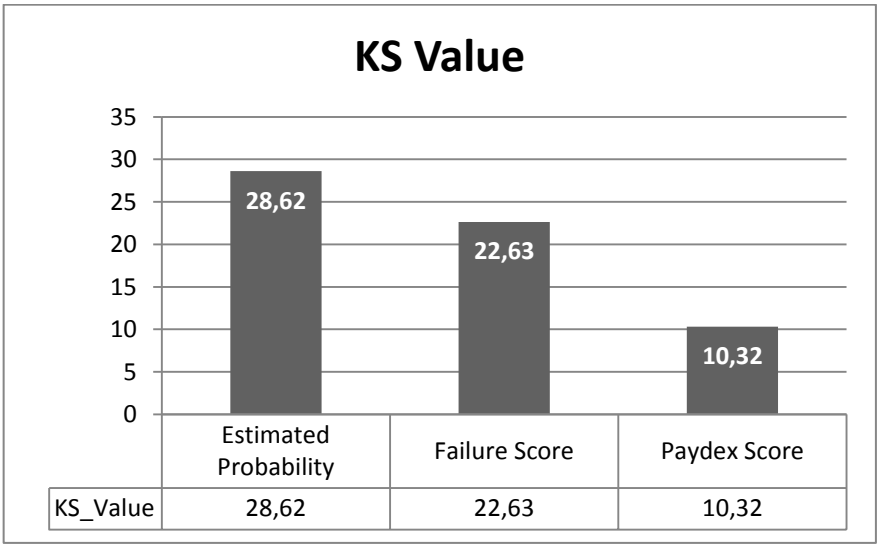

FIGURE 21: KS COMPARISON (SOURCE: OWN CALCULATIONS)

6.3 Cross validation

For the reasons explained in section 3.4.2., the most promising approach to validate the model on another set of real data is to use the sample data and *k*-fold cross validation technique, where $k = 5$ has been chosen.

We divide the dataset into 5 parts, randomly selecting and marking 20% of the dataset. We replicate the process 10 times to get a bigger sample. Each time, one of the 5 subsets is used as the test set and the other 4 subsets are put together to form a training set.

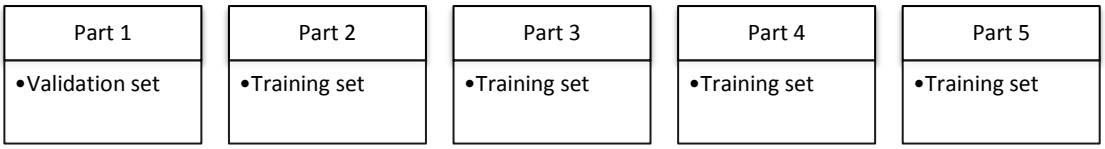

FIGURE 22: K-FOLD CROSS VALIDATION (SOURCE: OWN CHART)

We fit the model to the 4 parts and compute the error on predicting the 5_{th} part. Table 30 shows the final cross validation sample. Replicate is an indicator of the sampling round, while selected is showing whether the record has been selected into the training set or the test set. Percentagewise, we are seeing 20% of the original dataset to go into the test set and 80% of the original dataset to be selected to form the training set. This is true for all 10 sampling rounds or replicates.

Replicate	Selected	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	0	1438	2.00	1438	2.00
1	1	5756	8.00	7194	10.00
2	0	1438	2.00	8632	12.00
2	1	5756	8.00	14388	20.00
3	0	1438	2.00	15826	22.00
3	1	5756	8.00	21582	30.00
4	0	1438	2.00	23020	32.00
4	1	5756	8.00	28776	40.00
5	0	1438	2.00	30214	42.00
5	1	5756	8.00	35970	50.00
6	0	1438	2.00	37408	52.00
6	1	5756	8.00	43164	60.00
7	0	1438	2.00	44602	62.00
7	1	5756	8.00	50358	70.00
8	0	1438	2.00	51796	72.00
8	1	5756	8.00	57552	80.00
9	0	1438	2.00	58990	82.00
9	1	5756	8.00	64746	90.00
10	0	1438	2.00	66184	92.00
10	1	5756	8.00	71940	100.00

TABLE 30: CROSS VALIDATION SAMPLE (SOURCE: OWN CALCULATIONS)

After 10 rounds of rerunning the model, the average results of the cross validation need to be assessed. The best measure to assess model performance here is to see how well the probability of default reflects the real defaults.

Group	Predicted Probability: Bad flag=1				N	Bad flag	
	Min	Mean	Max	Sum		Sum	Mean
0	4.0%	6.1%	31.1%	176.483	2872	153	5.3%
1	3.1%	3.5%	4.0%	101.319	2880	105	3.6%
2	1.9%	2.5%	3.1%	72.56	2880	63	2.2%
3	1.1%	1.3%	1.9%	37.8	2860	52	1.8%
4	0.1%	0.7%	1.1%	20.738	2888	33	1.1%
All	0.1%	2.8%	31.1%	408.899	14380	406	2.8%

TABLE 31: CROSS VALIDATION PERFORMANCE MEASURES (SOURCE: OWN CALCULATIONS)

True picture of the predicting power of the model is seen on the trend of the default rate to the predicted Probability of default. The probability of default should follow the trend of the default rate. Figure 23 supports the statement.

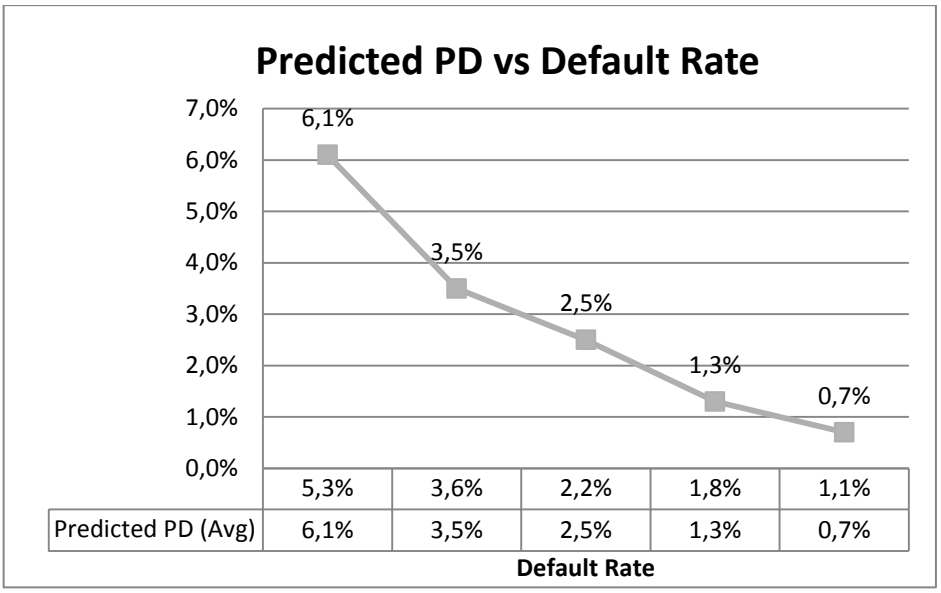

FIGURE 23: RESULTS OF CROSS VALIDATION (SOURCE: OWN CALCULATIONS)

7 DISCRIMINANT ANALYSIS APPLICATION

7.1 Assumptions

Compared to the logistic regression, the discriminant analysis can be used with small sample size datasets. This is true but it has a condition. It assumes a sufficient representation in each class/group.

More assumptions apply to this method. These are assumptions of normal distribution for the response variables.

We assume that the data are

- normally distributed within each group with equal covariances across groups (homoscedasticity),
- data are independent and a unit's value of one variable is independent to the values of the same variable for the other units,
- predicting variables are not strongly correlated (multicolinearity).

In case of this thesis, there are other assumptions related to the data definitions. These are the same as for the logistic regression.

The archive data are at application date or before and the performance indicators are at 12-18 months after the application date to enable us to identify the bad customers, while modelling on the application data that is available at the time of application.

While in cluster analysis we are looking to classify data into undefined groups, in discriminant analysis, the goal is to classify data into defined groups. This means that to perform a Discriminant analysis, initial group membership must be specified.

Here, the data are separated into two groups: performing and nonperforming. The variable to distinguish the two groups is called default. The non-performing loans with default = 1 represent the bad loans with the probability of default equal to 100%. As mentioned above, this means that it is more than likely that the customer will default within the next 12 months. The performing customers with default = 0 are the ones with the probability of default lower than 100%.

Bad definition consists of a few legal events that indicate the company is in bad shape. The list of events that define a company in default is shown in Table 18.

206 records are classified as bad and the remaining 6988 are classified as good. This pre classification is one of the basic input to the Discriminant analysis.

Class Level Information				
Bad flag	**Variable Name**	**Frequency**	**Weight**	**Proportion**
0	0	6988	6988	0.9714
1	1	206	206	0.0286

TABLE 32: CLASS LEVEL INFORMATION (SOURCE: OWN CALCULATIONS)

The defaulting customers are flagged as 1, using bad flag as an indicator and dependant variable. The Bad flag can only have two values. It is equal to 1 if the unit is in default and equal to 0 if the unit is performing. The bad rate is 2.86%.

7.2 Selection of variables

In order to conduct a discriminant analysis, a good training data set is required. We will use the same set of data as for the logistic regression. It contains 42 selected modelling variables and 7194 observations.

The Method for Selecting Variables is STEPWISE			
Total Sample Size	7194	Variable(s) in the Analysis	40
Class Levels	2	Variable(s) Will Be Included	0
		Significance Level to Enter	0.05
		Significance Level to Stay	0.05

TABLE 33: DATA SELECTION (SOURCE: OWN CALCULATIONS)

This data set is used to determine the combination of the responses which best describes each group.

As in the case of the logistic regression, we can use the stepwise selection of variables to enter the Discriminant analysis. The stepwise procedure selects variables that can be used as significant indicators of the differences between the two classes.

The Method for Selecting Variables is STEPWISE			
Total Sample Size	7194	Variable(s) in the Analysis	40
Class Levels	2	Variable(s) Will Be Included	0
		Significance Level to Enter	0.05
		Significance Level to Stay	0.05

TABLE 34: STEPWISE SELECTION (SOURCE: OWN CALCULATIONS)

With stepwise selection for discriminant analysis, variables are chosen to enter or leave the model according to one of two criteria:

- the significance level of an F test from an analysis of covariance, where the variables already chosen act as covariates and the variable under consideration is the dependent variable

- the squared partial correlation for predicting the variable under consideration from the class variable, controlling for the effects of the variables already selected for the model

Stepwise methods sequentially add variables, at each step adding the variable which contributes most to the discriminatory power of the model, as measured by Wilks' lambda[21]. When no other variables meet the criterion to enter the model, the process stops.

Out of the total of 42 factors, 5 variables were selected to enter the model build. 3 out of the 5 were previously identified as powerful by the stepwise procedure of the logistic regression.

Step	Entered	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda
1	Delivered Risk Indicator	0.0114	83.29	<.0001	0.9886	<.0001
2	No of Accts for Collections in L36M	0.0009	6.63	0.0101	0.9876	<.0001
3	No of Accts for Collections in L12M	0.0018	13.30	0.0003	0.9858	<.0001
4	No of Dirs Resigned L12M	0.0009	6.66	0.0099	0.9849	<.0001
5	No of Current Directors	0.0010	7.33	0.0068	0.9839	<.0001

TABLE 35: SELECTED VARIABLES (SOURCE: OWN CALCULATIONS)

The meaning of the selected variables and their relationship with a default event will help to validate them in terms of their predictive power:

1. **Delivered Risk Indicator** is a banded representation of the failure score. The Failure Score predicts the likelihood that a company will obtain legal relief from its creditors or cease operations over the next 12 month period. The Failure scorecard looks for the onset of failure such as meeting of creditors, administrator appointed, bankruptcy, receiver appointed, petition for winding-up among others legal events.
   It has been shown in Section 7.2.2 that the Delivered Risk Indicator has a positive relationship with the Default event. The higher is the indicator, the higher is the chance of a default.

2. **Number of Current Directors** is the number of owners or directors that a company has. This indicates that the more the directors a company has the more secure the company has i.e. the more directors indicates the size of the company, where small companies are riskier in general.
   It has been shown in Section 7.2.2 that the Number of Current Directors has a negative relationship with the Default event. Higher number of directors is a sign of a bigger company and a small chance of this company to go bankrupt or insolvent.

[21] Wilks' lambda distribution is a probability distribution introduced by Samuel S. Wilks and used in multivariate analysis of variance to test whether there are differences between the means of identified groups of subjects on a combination of dependent variables.

3. **Number of Resigned Directors** L12M is the number of directors of a company which have resigned in the last 12 months. The greater the number of resignations indicates the company may have underlying issues and may have payment problems due to turnover in the staff. This can be an early indicator of a potential default.
   It has been shown in Section 7.2.2 that the Number of Resigned Directors L12M has a positive relationship with the Default event. It is obvious that stable companies do not witness frequent resignations of their directors.

The stepwise selection sees two other variables as potentially predictive.

- Number of Accounts Placed for Collections in last 12 month
- Number of Accounts Placed for Collections in last 36 month

By the name, they seem to be looking at the same information, but in a different period. This would mean a high correlation between predictors, something that we would like to avoid. We can confirm this by calculating the Pearsons Correlation Coefficient (Table 36) and by looking at the cluster analysis in Section 5.4.1.

Pearson Correlation Coefficients, N = 7194		
	No of Accts for Collections in L36M	
No of Accts for Collections in L12M	0.84474	
	<.0001	

TABLE 36: CORRELATION BETWEEN TWO SELECTED PREDICTORS (SOURCE: OWN CALCULATIONS)

To identify which of the above two variables has better predictive power, we will undergo the same analysis as in Section 7.2.2.

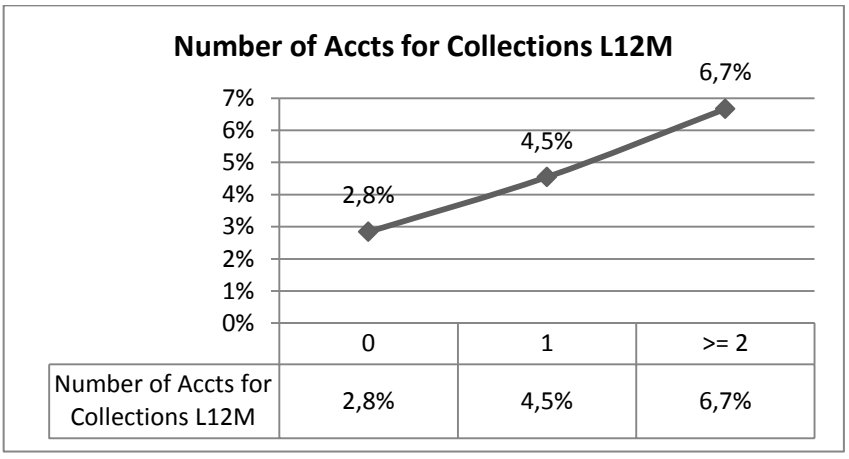

FIGURE 24: NO OF ACCTS FOR COLLECTIONS L12M VS DEFAULT (SOURCE: OWN CALCULATIONS)

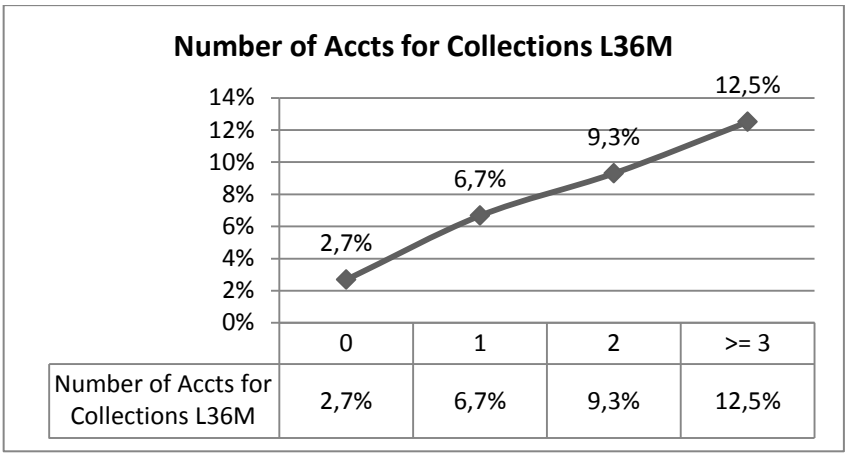

FIGURE 25: NO OF ACCTS FOR COLLECTIONS L36M VS DEFAULT (SOURCE: OWN CALCULATIONS)

The 3 years overview clearly provides more information than the 12 months view. This variable will enter the model.

7.3 Linear Discriminant model

The goal now is to find a discriminant function based on the 4 variables selected that best classifies the units into the two classes.

In SAS, the discriminant procedure is called PROC DISCRIM. It computes various discriminant functions for classifying observations. Linear or quadratic discriminant functions can be used for data with approximately multivariate normal within-class distributions. Nonparametric methods can be used without making any assumptions about these distributions. (SAS Institute Inc., 1999)

The model has been calculated with the selected final number of 4 explanatory variables, highlighted below. SAS procedures have been used for the analysis.

Table 37 shows the summary information about the input to the model.

Total Sample Size	**7194**	**DF Total**	**7193**
Variables	7	DF Within Classes	7192
Classes	2	DF Between Classes	1

TABLE 37: SUMMARY INFORMATION (SOURCE: OWN CALCULATIONS)

Unless prior probabilities are specified, the default is usually set as equal priori probabilities.

We do not know the exact proportional for the two groups. We will assume that there is 50% for the company to end up in the default group, so is 50% chance to remain in the good book.

The analysis gives the best results analysis when applied onto a randomly selected dataset and testing it on a different test sample. The dataset needs to be divided into 5 parts, where 1 part forms the test dataset and the rest is used to train the model on.

Class Level Information			
Bad flag	**Frequency**	**Proportion**	**Prior Probability**
0	5574	0.9684	0.5000
1	182	0.0316	0.5000

TABLE 38: PRIOR PROBABILITY (SOURCE: OWN CALCULATIONS)

The squared distances between the classes are shown in the table below, calculated as

Generalized Squared Distance to bad flag		
From bad flag	0	1
0	0	0.6557
1	0.6557	0

TABLE 39: SQUARED DISTANCE TO DEFAULT (SOURCE: OWN CALCULATIONS)

Since I am using a linear model, I will expect to get good results only if there are differences between the groups:

Bad flag	N Obs	Variable	Minimum	Mean	Maximum
0	5574	No of Current Directors	0	2.7488	17
		Delivered Risk Indicator	1	2.368	5
		No of Dirs Resigned L12M	0	0.1636	3
		No of Accts for Collections in L36M	0	0.0486	8
1	182	No of Current Directors	0	1.6484	10
		Delivered Risk Indicator	1	3.2473	5
		No of Dirs Resigned L12M	0	0.1868	3
		No of Accts for Collections in L36M	0	0.1813	6

TABLE 40: GROUP DESCRIPTIVE STATISTICS (SOURCE: OWN CALCULATIONS)

The coefficients are chosen to maximize separation between groups. LDA is based on a linear model, and assumes normality and homoscedasticity.

Linear Discriminant Function for Bad flag		
Variable	**0**	**1**
Constant	-3.0499	-4.5154
No of Current Directors	0.5088	0.4451
Delivered Risk Indicator	1.9840	2.4880
No of Dirs Resigned L12M	0.1305	0.5636
No of Accts for Collections in L36M	-0.3765	0.6212

TABLE 41: COEFFICIENTS OF LINEAR DISCRIMINANT FUNCTION (SOURCE: OWN CALCULATIONS)

7.4 Model performance and error rates

The performance of a discriminant function can be evaluated by estimating error rates (probabilities of misclassification) or cross validation. The error-count estimates give the proportion of misclassified observations in each group. These error rates can be biased as they are calculated on the same training dataset.

Error Count Estimates for bad flag			
	0	1	Total
Rate	0.3105	0.4066	0.3586
Priors	0.5	0.5	

TABLE 42: ERROR COUNT ESTIMATES (SOURCE: OWN CALCULATIONS)

We compare the original group memberships to the LDA group assignments and measure the percentage "correctly" classified. The model has identified 108 nonperforming cases, out of 182, which is 59% performance. In total, we can count 1805 incorrectly classified units out of 5756. This represents an average misclassification rate of 35.86%.

Number of Observations and Percent			
Classified into Bad flag			
From	**0**	**1**	**Total**
0	3843	1731	5574
	68.95	31.05	100
1	74	108	182
	40.66	59.34	100
Total	3917	1839	5756
	68.05	31.95	100
Priors	0.5	0.5	

TABLE 43: ERROR RATES ON TRAINING SET (SOURCE: OWN CALCULATIONS)

We wanted to separate 20% of the original dataset to run a validation and confirm the error rates stated above. The results of the validation are shown in Table 44. The total misclassification rate is 43.28%.

Number of Observations and Percent			
Classified into Bad flag			
From	**0**	**1**	**Total**
0	956	458	1414
	67.61	32.39	100
1	13	11	24
	54.17	45.83	100
Total	969	469	1438
	67.39	32.61	100
Priors	0.5	0.5	

TABLE 44: ERROR RATES ON TESTING SET (SOURCE: OWN CALCULATIONS)

Same as in the case of Logistic regression, we can look at the discriminative power measure by Kolmogorov–Smirnov (KS) statistic. In Figure 26, the KS score of the proposed model is compared to that of two credit bureau scores shown to be powerful risk discriminators in their own right. It can be observed that the proposed model is more powerful than both the Failure Score and the Paydex Score.

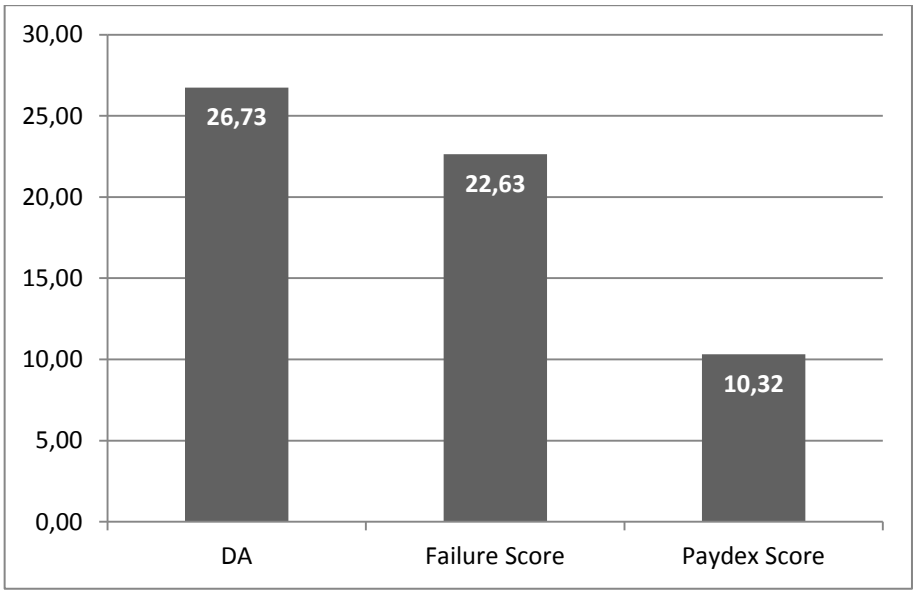

FIGURE 26: KS COMPARISON (SOURCE: OWN CALCULATIONS)

8 DATA ENVELOPMENT ANALYSIS APPLICATION

We have already described the basic models and formulations in Chapter 6. In this chapter, we will look at the application of such models to our dataset of a credit portfolio. We will use the CCR output-oriented model assuming constant returns to scale, as defined in (5.1-7).

Once the DEA model is built, the linear programming technique will attempt to find the highest value of the efficiency for each of the units. The program goes through a number of iterations until the efficiency of one or more units hits the number 1. If the efficiency is equal to 1 and the optimal values of the slacks (s_k^+,s_i^-) are zero, the DMU is said to be Pareto efficient. Any result lower than one is considered inefficient. This means that a linear combination of other units from the sample could produce the vector of outputs using a smaller vector of inputs.

8.1 Data selection

The DEA needs to be applied with care and judgement. There is a need for large enough number of similar units and this number must be much greater than the number of inputs and outputs chosen. Otherwise it could cause weak discrimination between units.

The data consists of a group of units (DMUs) and the values of their inputs and outputs to be included in the analysis. Here the group of units is made of bank's credit applicants.

Selection of the inputs and outputs is the core issue of the DEA analysis. Inappropriate selection or inappropriate number of inputs and outputs can distort the analysis and lead to incorrect results. Zero values as well as missing values are not allowed in DEA. Units with missing values have to be omitted from the data set or substitute values agreed upon. SAS has been used to enable all data manipulation, data mining and analysis. Most of the data manipulation has been done before applying the traditional methods above, the logistic regression and the linear discriminant. The main issues of the data cleaning and manipulation were to delete any duplicate records and deal with missing and incorrect values.

The DEA will be applied on the same data set as the other two methods, so that we can compare the results and assess the performance of the two traditional methods versus the new practical approach of the data envelopment analysis. Observations with missing values for selected inputs and outputs will be removed.

8.1.1 Information Value ranking

As mentioned in section 5.4, one of the most used methods for selecting variables is the Information Value (IV) metric that looks at the differences between the distributions of good and bad accounts. The output is then showing how powerful each of the variables is in terms of predicting the default. The IV is sometimes called Divergence.

Top 12 strongest variables are shown in Table 45. To no surprise, the main model variables used to predict default with a logistic and discrimination function are present in the list.

Rank	Variable	No of Bins	KS value	Info Value
1	Delivered Risk Indicator	2	26.5955	0.3224
2	Calculated Risk Indicator	2	24.3995	0.2945
3	Percentile Failure Score	2	20.5039	0.1777
4	Calculated Raw Score	2	20.461	0.1770
5	Cash in bank and hand	2	18.5336	0.1436
6	Number of Current	3	12.6054	0.0978
7	Net Profit Loss	2	13.6403	0.0967
8	Financial Strength Indicator	2	13.8608	0.0795
9	No of Accts for Collections in L36M	2	6.2457	0.0686
10	Time since start-up in days	3	11.161	0.0552
11	Issued Capital	2	11.208	0.0511
12	Tangible Net Worth	2	9.4952	0.0365

TABLE 45: TOP 10 VARIABLES WITH HIGHEST INFORMATION VALUE (SOURCE: OWN CALCULATIONS)

To calculate the information value, the Weight of Evidence (WOE) methodology was used. The WOE method is commonly used to help categorize variables into logical buckets and measures the strength of an attribute of a characteristic in differentiating good and bad accounts. Weight of evidence is based on the proportion of good to bad applicants at each group level. Negative values indicate that a particular grouping is isolating a higher proportion of bads than goods. The Weight of Evidence of an attribute is defined as the logarithm of the ratio of the proportion of "goods" in the attribute over the proportion of "bads" in the attribute. It is in fact the second part of the IV formula. High negative values correspond to high risk; high positive values correspond to low risk.

The Information Value is the weighted sum of the Weights of Evidence of the characteristic's attributes. The sum is weighted by the difference between the proportion of "goods" and the proportion of "bads" in the respective attribute, as defined in (6.4-1).

A few studies have explored the scoring models using the weight of evidence (WOE) measure, or in terms of poor as good or good and bad credit, also results were comparable with those from other techniques (Siddiqi, 2006), (Banasik & Crook, 2003), (Bailey, 2004).

Siddiqi states the rule of thumb regarding the Information Value (Siddiqi, 2006):

- Less than 0.02 is unpredictive
- 0.02 to 0.1 is weak
- 0.1 to 0.3 is medium
- More than 0.3 is strong

All of the top 10 selected variables fall into weak, medium and strong categories.

8.1.2 Input and output selection

The DEA requires classifying variables into inputs and outputs. We take the list of top 12 selected variables and indicate whether this can be seen as an input to the process or output from the process.

Rank	Variable	Info	Min/Max	Input/Output
1	Delivered Risk Indicator		Min	I
2	Calculated Risk Indicator		Min	I
3	Percentile Failure Score		Max	O
4	Calculated Raw Score		Max	O
5	Cash in bank and hand		Max	O
6	Number of Current Directors		Max	O
7	Net Profit Loss		Max	O
8	Financial Strength Indicator		Min	I
9	No of Accts for Collections in L36M		Min	I
10	Time since start-up in days		Max	O
11	Issued Capital		Max	O
12	Tangible Net Worth		Max	O

TABLE 46: POSSIBLE INPUTS AND OUTPUTS (SOURCE: OWN CALCULATIONS)

From the collinearity and cluster analysis, we know that Calculated Risk Indicator, Calculated Raw Score and Percentile Failure Score should not enter any model together as they could bias the results. Only one of these can be selected.

Having tested various combinations of inputs and outputs, the best results are given by the following set of model variables, presented in Figure 27.

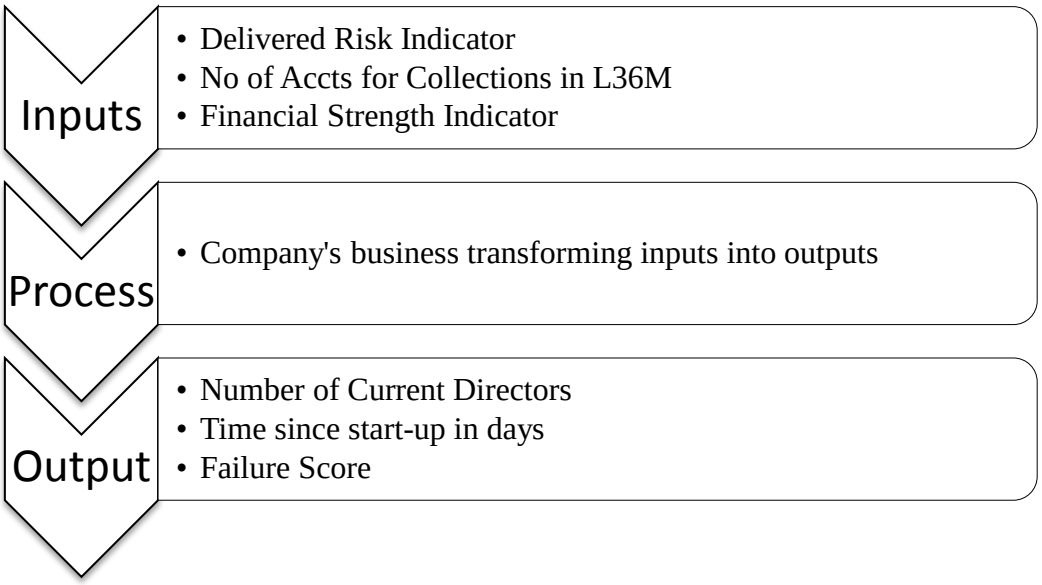

FIGURE 27: INPUT AND OUTPUT SELECTION (SOURCE: OWN CHART)

The inputs are to be minimized, while the outputs are to be maximized.

8.2 DEA software tools

Today's DEA users and researchers have a wide range of solution technology choices. Barr provided a survey listing of some of the best of the commercial and non-commercial DEA tools available in 2004 (Barr, 2004). Some of them are independent software programs, such as Frontier Analyst, DEA Solver Pro or EMS; other solutions represent add-ins to existing tools such as Excel. These software products provide a wide range of available models, features and capabilities, user interfaces, reporting options, model solution speeds, and acquisition costs.

The disadvantages of some of those tools are the incompatibility with other software products where from the data are fed or go to. Another important criterion the researchers and potential users look at is the limit to the size of data. The add-ins are usually related to widely used product, such as Excel. They are easy to install and their scope of use is limited by the Excel specifications. The size of the records that Excel can handle is limited to 60000 rows. This would be sufficient for small models, but not for credit scoring models run daily on thousands of clients.

The least different software products exist in the bank's IT system, the smaller the possibility of data errors, losses or unwanted transformations. The better the compatibility and linkage of the various software products, the more automated process one can get. If using a separate independent software product for calculation of DEA efficiencies, we first need to export the data from the source database in a format that would be acceptable by the independent DEA software, then import again, run the model and export results and store.

The above limitations give priority to software systems that can handle enormous databases, provide all sort of data mining, statistical, mathematical and reporting functions and are compatible with widely used products. SAS offers all of that and provides even specialized packages suitable for various industries and business activities. Nowadays, many organizations such as banks, universities, hospitals, telecommunications and airlines use SAS software as a database system and/or statistical analysis tool.

SAS allows transforming data about customers, performance, financials and more into information and predictive insight that lays the groundwork for solid and coherent decisions. As presented by the SAS Institute (Basel Commitee on Banking Supervision, 2005)., SAS is used at more than 50,000 sites in over 100 countries, including 93 of the top 100 companies on the 2010 FORTUNE Global 500® list.

This fact forced the developers in SAS to look at different industries and offer, beside its basic set of functions and operations (called Base SAS), specialized packages that can be used to facilitate treatment of specific problems. A specialized package for optimisation[22] programming with all types of constraints was developed. SAS/OR is a set of procedures for

[22] The process of determining the optimal values for the decision variables, so the objective is either maximized or minimized, is called optimisation.

exploring models of distribution networks, production systems, resource allocation problems, and scheduling problems using the tools of Operations Research.

SAS/OR software can be used to solve a wide variety of optimisation problems. Kearney defines the basic optimisation problem as that of minimizing or maximizing an objective function subject to constraints imposed on the variables of that function (Kearney, 1999). The absence of a special procedure for calculating DEA scores requires building a DEA model using the available functions of SAS/OR.

There are two procedures that can be used to implement a DEA model. The first procedure, that was introduced for DEA modelling by (Paradi, Asmild, & Simak, 2004) is called the linear programming (LP) procedure. Sabah introduced the second approach using the OPTMODEL procedure at the SAS Global Forum (Sabah, 2011). It is based on several programming techniques, the linear programming including. The difference is in the way of programming, the data preparation, the length of the code and the time it takes to resolve the problem.

The LP procedure within SAS solves linear and mixed integer programs with a primal simplex solver. The LP procedure provides various control options and solution strategies. (Charnes, Cooper, & Rhodes, 1978) state that there are no restrictions on the problem size in the LP procedure, but the bigger the problem, the more resources of memory and time you need.

8.3 DEA model using the OPTMODEL procedure

All datasets are taken from the same testing dataset used in the previous two methods, the regression analysis and the discriminant analysis. For the purpose of DEA, we need to separate the variables to be used in the model, into 2 groups. The inputs, being the variables to be minimized entering the production process of a unit and the outputs, being the variables to be maximized as a result of the production process.

An example of a data file containing inputs and called **Inputs** is as follows. The same format is used for the dataset containing output variables and called **Outputs**.

DMU	Input1	Input2
1		
2		

TABLE 47: INPUTS SOURCE

The OPTMODEL procedure provides a modelling environment that is tailored to building, solving, and maintaining optimisation models. (SAS Institute Inc., 2010) In this thesis, the OPTMODEL procedure is used to rank credit applicants using Data Envelopment Analysis Techniques. The OPTMODEL uses a number of solvers including linear and non-linear

programming. In addition to invoking optimization solvers directly with PROC OPTMODEL, the OPTMODEL language can be used purely as a modelling facility. The coding for modelling DEA with the OPTMODEL procedure is shorter than with the LP procedure only and the code runtime is smaller.

Basic concept of the linear programming or linear optimisation lies in the search for the optimal outcome of a process. It is a technique used to optimize a certain problem that can be represented with a linear model. An objective function is optimized subject to constraints that limit the feasible choices of variables being evaluated. In terms of DEA, the objective function is to maximize the efficiency subject to the efficiency of all units being less than 1. To be able to use the LP procedure, data need to enter the analysis in a certain form, so that SAS can recognize the variables needed to run the linear programming steps. Two different formats are possible, the sparse and the dense model.

The model data used in this thesis enter the analysis in the sparse model. The LP procedure within SAS solves linear and mixed integer programs with a primal simplex solver. The LP procedure provides various control options and solution strategies. There are no restrictions on the problem size in the LP procedure. The sparse format to PROC LP is designed to enable you to specify only the nonzero coefficients in the description of linear programs, integer programs, and mixed-integer programs. (Charnes, Cooper, & Rhodes, 1978)

The SAS data set that describes the sparse model must contain at least four SAS variables:

- type of variable (max, equal, etc.)
- column variable (here names of inputs and outputs or right hand side (RHS) column)
- row variable (conditions or objective function) and
- coefficient variable (values of inputs and outputs)

Each observation in the data set associates a type with a row or column, and defines a coefficient or numerical value in the model.

The data is generated in the sparse format. The variable names are the structural variables, the rows are the constraints, and the coefficients are given as the values for the structural variables.

row	_col_	_type_	_coef_
Constraints	Variable names and _rhs_	Max/Min/EQ/LE/GE	Values

TABLE 48: SPARSE FORMAT

The macro DEA_OPTMODEL (see Appendix for code) does all the necessary transformation of the original data into a format that is readable by the SAS procedure, as well as the running of the model and providing results. Two source datasets used are the Inputs and Outputs datasets, in the format described above.

```
%DEA_OPTMODEL (Inputdata=Inputs, Outputdata=Outputs);
```

The macro here consists of 2 parts. The first part aims at preparing the data, looking for:

- number of inputs and outputs,
- number of DMUs.

The model part describes the system of equations, the parameters of the model and the variables taken directly from the source files. Once the model to be run is formulated, the solver can be called.

In order to accumulate benchmarks, efficiency values and produce a graphical comparison of efficiencies, the PROC APPEND is used. It is a SAS procedure that adds the observations from one SAS data set to the end of another SAS data set.

Solution Summary gives information about the solution that was found, including whether the optimizer terminated successfully after finding the optimum. When PROC LP solves a problem, it uses an iterative process. First, the procedure finds a feasible solution that satisfies the constraints. Second, it finds the optimal solution from the set of feasible solutions. The Solution Summary lists information about the optimization process such as the number of iterations, the infeasibilities of the solution, and the time required to solve the problem. (SAS Institute Inc., 2010).

NOTE: The problem has 6 variables (0 free, 0 fixed).
NOTE: The problem has 4779 linear constraints (4778 LE, 1 EQ, 0 GE, 0 range).
NOTE: The problem has 28671 linear constraint coefficients.
NOTE: The problem has 0 nonlinear constraints (0 LE, 0 EQ, 0 GE, 0 range).
NOTE: The OPTMODEL presolver is disabled for linear problems.
NOTE: The OPTLP presolver value AUTOMATIC is applied.
NOTE: The OPTLP presolver removed 0 variables and 83 constraints.
NOTE: The OPTLP presolver removed 498 constraint coefficients.
NOTE: The presolved problem has 6 variables, 4696 constraints, and 28173 constraint coefficients.
NOTE: The DUAL SIMPLEX solver is called.
Objective
Phase Iteration Value
2 1 25.615817
2 11 0.933846
NOTE: Optimal.
NOTE: Objective = 0.933845503.

TABLE 49: SOLUTION SUMMARY FOR SAMPLE UNIT (OWN CALCULATIONS)

For every unit, the solution summary provides the optimal solution and the number of iterations needed to find the final solution.

8.4 Distribution of the efficiency scores

The calculated DEA scores vary from 0 to 1. Scores equal to 1 are considered as the best and most efficient ones. In average, we expect the bad units to have a predicted efficiency at a lower level than the good ones. We can compare the basic statistics and confirm that it is the case.

Analysis Variable : Efficiency								
Bad flag	N Obs	N	Minimum	25th Pctl	Mean	Median	75th Pctl	Maximum
0	6251	6251	0.0023	0.4628	0.6989	0.814	0.98	1
1	187	187	0.0119	0.1742	0.5126	0.5218	0.8125	1

TABLE 50: COMPARISON OF AVERAGE EFFICIENCIES (SOURCE: OWN CALCULATIONS)

All basic statistics for the bad accounts are lower than for the good. Only 5 of the bad units received an efficiency of 1 which indicates good level of discrimination of the portfolio between good and bad.

Figure 28 shows the distribution of the DEA scores. The units are spread across the scale with main concentration of units from 0.8 to 1. There are 855 fully efficient units according to the results of the DEA run. These units form the efficient envelope. The majority are not located on the frontier, but their target values are not far from the efficient ones.

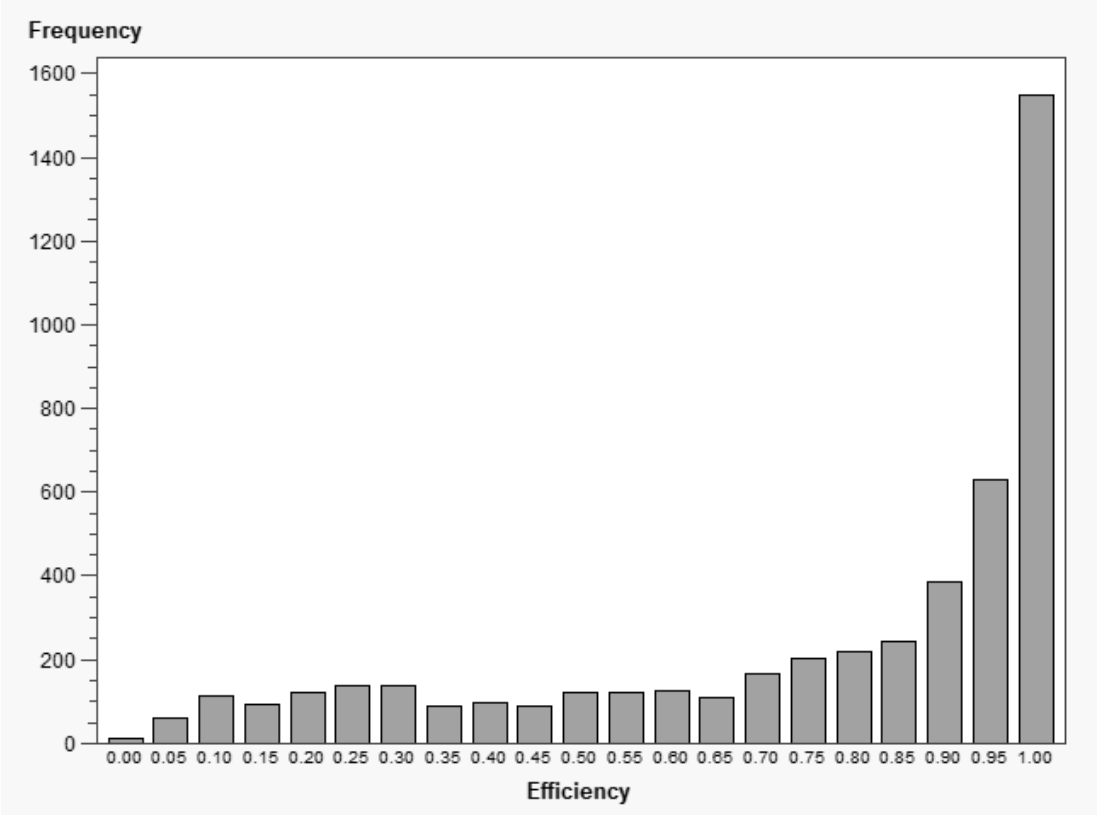

FIGURE 28: DEA SCORES DISTRIBUTION IN A CHART (SOURCE: OWN CALCULATIONS)

As said before, the efficiency calculated by the DEA algorithm is not absolute but relative. It means that the units having a smaller DEA score are more likely to go to default than the ones with a higher score.

Table of Efficiency by Bad flag				
		Bad flag		Total
		0	1	
Efficiency		606	49	655
Below 0.19	Frequency			
	Percent	9.41	0.76	10.17
	Row Pct	92.52	7.48	
0.20 - 0.29	Frequency	458	12	470
	Percent	7.11	0.19	7.3
	Row Pct	97.45	2.55	
0.30 - 0.39	Frequency	316	19	335
	Percent	4.91	0.3	5.2
	Row Pct	94.33	5.67	
0.40 - 0.49	Frequency	361	8	369
	Percent	5.61	0.12	5.73
	Row Pct	97.83	2.17	
0.50 - 0.59	Frequency	351	15	366
	Percent	5.45	0.23	5.68
	Row Pct	95.9	4.1	
0.60 - 0.69	Frequency	430	15	445
	Percent	6.68	0.23	6.91
	Row Pct	96.63	3.37	
0.70 - 0.79	Frequency	535	17	552
	Percent	8.31	0.26	8.57
	Row Pct	96.92	3.08	
0.80 - 0.89	Frequency	680	22	702
	Percent	10.56	0.34	10.9
	Row Pct	96.87	3.13	
0.90 - 0.99	Frequency	1673	25	1698
	Percent	25.99	0.39	26.37
	Row Pct	98.53	1.47	
1	Frequency	841	5	846
	Percent	13.06	0.08	13.14
	Row Pct	99.41	0.59	
		6251	187	6438
Total	Frequency	6251	187	6438
	Percent	97.1	2.9	100

TABLE 51: DEA SCORES DISTRIBUTION (OWN CALCULATIONS)

The decreasing bad rate trend with increasing efficiency is clearly shown in Figure 29. The percentages are based on the calculations in Table 51 (row pct).

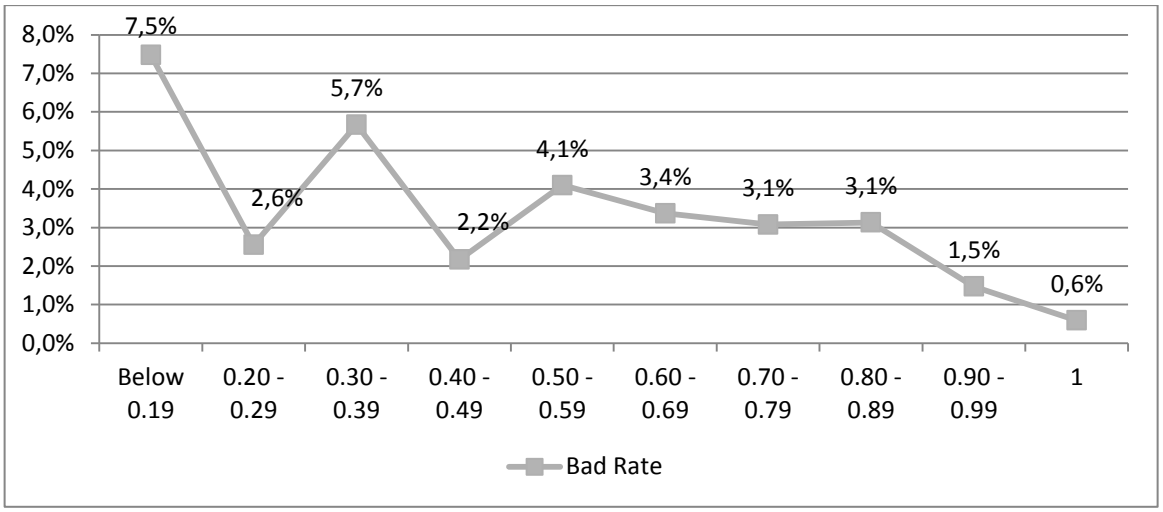

FIGURE 29: BAD RATE TREND BASED ON EFFICIENCY SCORES (SOURCE: OWN CALCULATIONS)

The credit scoring managers have to decide on the way they interpret the results. The bank can set a threshold DMU to be compared with. This can be the DMU with the lowest inputs and highest outputs or a DMU with a certain score. Any DMU assessed having a lower score than the threshold will be seen as bad (Min & Lee, 2004).

KS statistic can be computed for any score. The discriminant power of the DEA Efficiency score is displayed in Figure 30.

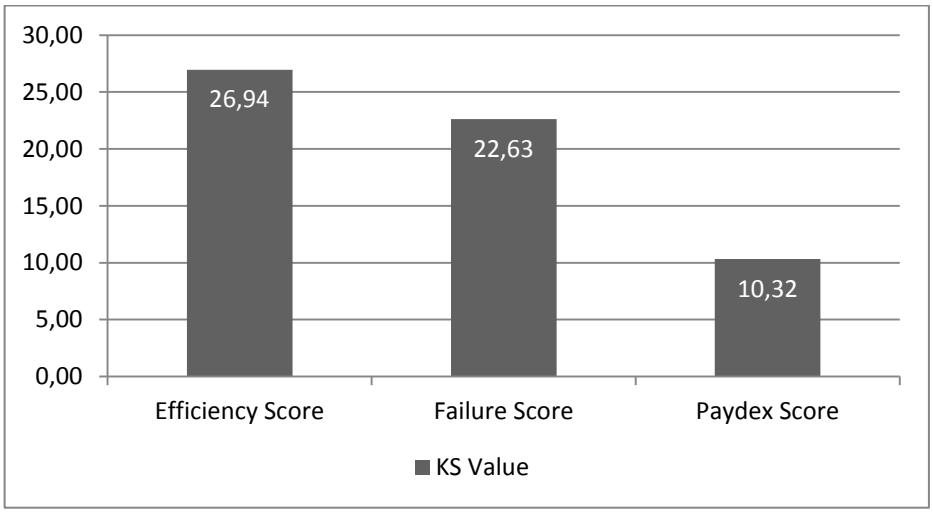

FIGURE 30: KS COMPARISON (SOURCE: OWN CALCULATIONS)

8.5 Regression model with censored data

To facilitate future calculation of the scores for new applicants, we need to apply a regression model looking for relationship between the DEA scores and the financial predictors (inputs and outputs). This approach provides a fitted regression relationship that can be used for every new entering unit without the need to run the DEA analysis on the whole lot.
We could use the ordinary linear regression, but there is a risk of having predictions outside of the range of the efficiency score. We know that the DEA efficiency is constrained to 0-1.

Tobit models, also called censored regression models, are designed to estimate linear relationships between a limited dependent variable and other independent variables. The dependent variables can be censored from left or right or both.
In the case of this thesis, censoring from both sides is necessary, where lower bound is equal to 0 and upper bound is equal to 1.

Summary Statistics of Continuous Responses							
Variable	Mean	Standard Error	Type	Lower Bound	Upper Bound	N Obs Lower Bound	N Obs Upper Bound
Efficiency	0.6935	0.3072	Censored	0	1	0	846

TABLE 52: CENSORING OF RESPONSE VARIABLE (SOURCE: OWN CALCULATIONS)

As shown below, all the variables have expected directions; the inputs coefficients are negative while the outputs coefficients are positive, meaning the higher the output, the better the score. All except one are statistically significant under the level of 0.1.

Parameter Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**	**Approx Pr > |t|**
Intercept	1	0.2631	0.0080	32.83	<.0001
No of Accts for Collections in L36M	1	-0.0035	0.0018	-1.96	0.0504
Financial Strength Indicator	1	-0.0001	0	.	.
Delivered Risk Indicator	1	-0.0339	0.0015	-22.77	<.0001
Failure Score	1	0.0082	0.0001	125.14	<.0001
No of Current Directors	1	0.0054	0.0003	15.72	<.0001
Time in business in days	1	0.000002	0	.	.
_Sigma	1	0.0744	0.0008	96.11	<.0001

FIGURE 31: REGRESSION COEFFICICENTS (SOURCE: OWN CALCULATIONS)

Tobit regression coefficients are interpreted in the similar manner to OLS regression coefficients. However, the linear effect is on the uncensored latent variable, not the observed outcome (McDonald & Moffitt, 1980).

Having this linear relationship between the DEA scores and the variables, we can calculate an approximate DEA score for any new observation. The fitted DEA scores may differ slightly from the actual DEA scores, but the difference is not significant.

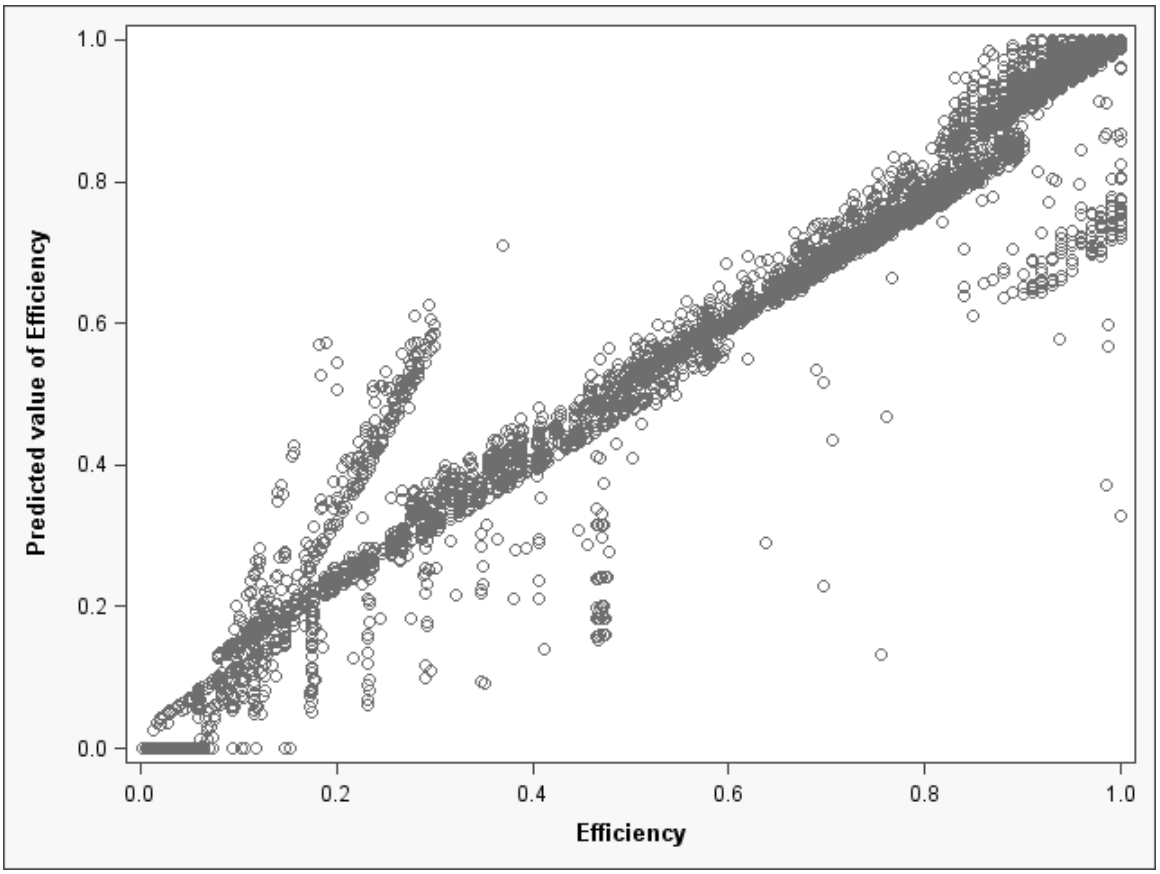

FIGURE 32: OBSERVED TO PREDICTED SCORES (SOURCE: OWN CALCULATIONS)

Further enhancement to the approach can bring the most recent modifications in the form of the Network DEA. Suggestion of how this methodology could be applied on our case study is shown in the section 10.

9 COMPARISON OF THE STANDARD METHODS AND DEA

One can think of a few different perspectives for comparison of the three methods that were applied in this thesis. This chapter presents all of these perspectives and compares results from the calculations.

9.1 Use of each method

The first perspective is the use of the method. This would include the transparency and complexity of the method. Not only it is important to understand how the method works, but it is important that also this can be easily explained to other departments than credit scoring or modelling teams and to the high management. Non transparent approaches will not get far in the private world. Innovation is certainly a good thing and employees are usually motivated to bring new ideas, but these have to be explained and confirmed with results.
From this point of view, the Logistic Regression is probably the easiest to present. It is due to the fact that the logistic regression is close to the ordinary linear regression and that is a term taught in most schools.

Discriminant analysis is not as popular as Logistic regression and would have to be presented with special care to ensure the model and methodology will get approved by the management. Obviously, alternative methods such as the proposed one, is not known to the larger public or even professionals. Proposal and implementation of such methodology requires supportive management and well prepared presentations. The idea of DEA is though easy to explain and can be shown on real life examples.

Also the use of the method involves IT resources. Methods requiring a lot of iterations can be heavy on IT resources, when it comes to a larger portfolio. This is a clear disadvantage of DEA. Comparing all of the units to the whole set is a significant task and can cause problems. On the other hand, the proposed methodology requires the DEA algorithm to be run only once to define the model (same as for the other two methods) and not at each time of application. Discriminant analysis and logistic regression are very similar in terms of resources.

9.2 Assumptions and limitations

All of the three methods have their assumptions and shortcomings, as well as advantages.

If we are to compare the standard methods - logistic regression and discriminant analysis, we can state that the latter is limited by more assumptions and restrictions than the logistic regression. On the other hand, discriminant analysis can be used with small sample size datasets. Logistic regression is not build on many assumptions, but requires a bigger base of data to achieve stable and meaningful results. Unlike ordinary linear regression, logistic regression does not assume that the relationship between the independent variables and the

dependent variable is a linear one. Nor does it assume that the dependent variable or the error terms are distributed normally. It does not require that the independents be interval or be unbounded. You can add explicit interaction and power terms. Both methods are sensitive to the levels of representation in each class/group.

The main difference and advantage of DEA compared to the other two methods, is that the method does not stand on prior classification of the units and is non-parametric. Beside these, we can account for no need of prior specification of weights or functional forms of relations between inputs and outputs, transparency and user friendliness compared to black box methods such as neural networks.

9.3 Results

The outcomes from the methods are not consistent and that is also one of the reason to consider when choosing the right method.

The usefulness of the logistic regression is in the direct output from the model. It is the probability of an event. Here we talk about default and the probability of default. Equation (3.3-3) can be used to transform the predicted results of the model to the probability of default.

Discriminant analysis also provides a probability of an observation belonging to one group or another. In this case, the output shows the probability of being classified as defaulted. The output though is not as transparent and granular as in the case of the logistic regression.

Output from DEA is a relative measure of efficiency, compared to the rest of the group. Same as in case of LDA, it does not provide the direct output in form of the probability of default that can be mapped to the bank's grading scale and grade the customers based on their PDs.

The details of the results from each of the method are outlined below.

9.3.1 Logistic Regression Results

The dependant variable in the model is the logit, or in other words, logarithm of the odds of being in default. As explained in section 3.3.2, logarithm of the odds is just another way of representing probabilities. The log-odds are calculated using the regression equation.

Based on the relationship between the log-odds and probability (3.3-3), we can derive the raw PD. The raw PD can then be mapped directly to the PD Grade, using the PD Grading Scale in Table 15.

Variable	Variable Name	Coefficients (Weights)	Value	Weighted Value
	Intercept	-4.6583		-4.6583
1	Delivered Risk Indicator	0.4689	4	1.8756
2	No of Current Directors	-0.1225	2	-0.2450
3	No of Dirs Resigned L12M	0.5357	1	0.5357
			Log odds	-2.4920
			Raw PD	7.64%
			PD Grade	11
			Midpoint PD	9.54%

TABLE 53: SAMPLE PD CALCULATION (SOURCE: OWN CALCULATIONS)

The logistic regression provided a granular output in terms of PD grading scale. Using the approach explained above, the distribution of the population across the PD grades is presented in Table 54. The Bad Rate across the sample portfolio is 2.86% (206 bads out of 7194). Based on this model, the majority of the applicants will fall into grades 6-10, covering 91% of all the population.

PD_GRADE	BADS	GOODS	ALL	POPULATION_PERC	BAD_RATE
3	0	22	22	0.31%	0.00%
4	0	101	101	1.40%	0.00%
5	2	242	244	3.39%	0.82%
6	3	724	727	10.11%	0.41%
7	28	1553	1581	21.98%	1.77%
8	27	1177	1204	16.74%	2.24%
9	67	1968	2035	28.29%	3.29%
10	46	953	999	13.89%	4.60%
11	27	235	262	3.64%	10.31%
12	5	12	17	0.24%	29.41%
13	1	1	2	0.03%	50.00%

TABLE 54: PD GRADE DISTRIBUTION (SOURCE: OWN CALCULATIONS)

Figure 33 clearly shows the trend of the bad rate versus the population distribution across the PD grades. The relationship is as expected; the uplift in Bad rate is outside the tail of the normal distribution. Based on this relationship, the cut-off is recommended at PD=10.
By selecting this cut-off point, we accept the maximum default rate of 4.60% within the approved applications. The declined high risk applications will remain below 4%.

Subsequently, when the sample distribution is compared with the linear PD Grades and in particular the granularity of each of the PD Grades (i.e. the width of each band), it is apparent that the peak of the distribution is centred about the most sensitive part of the PD scale, implying that this mapping will give us the greatest level of control and transparency over the majority of new applications passed through the scorecard.

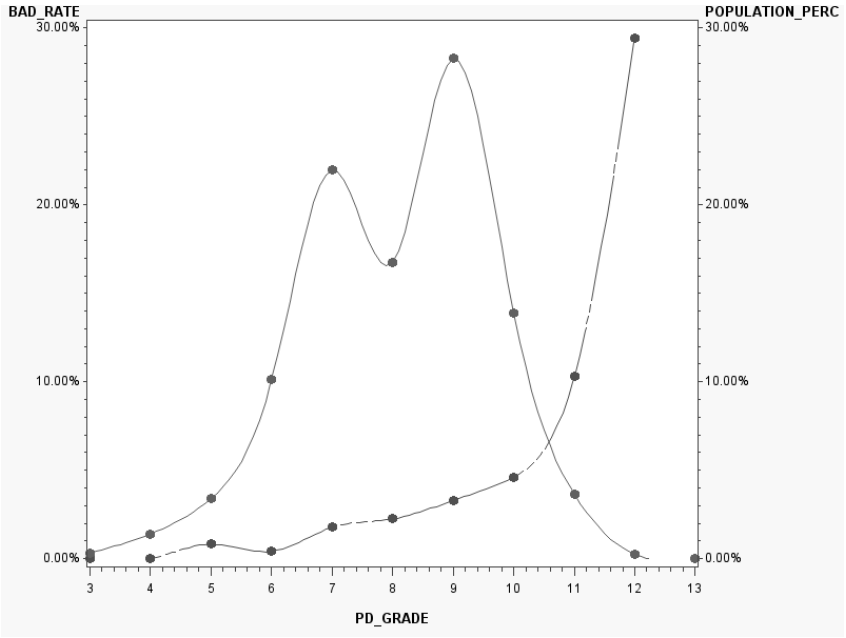

FIGURE 33: BAD RATE VS POPULATION DISTRIBUTION (SOURCE: OWN CALCULATIONS)

The Probability of Default is the main ingredient to the expected loss calculation and therefore to the amount of capital that the bank would need to hold. The other ingredients are exposure and the recovery rate[23].

9.3.2 Discriminant Analysis Results

The output of the discrimination analysis is a binary variable equal to 0 if the unit is in the performing class, equal to 1 if the unit is in the non-performing class. The classes are separate and each unit belongs to one of the groups. Regardless of whether the observation will be placed in that group, each observation will be assigned a probability of belonging to that group based on the distance of its discriminant function from that of each group mean.

Similar to the logistic regression, we can use the calculated probability of an observation belonging to one or the other group to assign a PD grade. On the same 16 band scale, the applications flagged as bad are all spread across PD grades 11 to 15.

Obs	PD_GRADE	BADS	GOODS	ALL	POPULATION_PERC	BAD_RATE
1	11	0	29	29	0.40%	0.00%
2	12	2	413	415	5.77%	0.48%
3	13	29	2164	2193	30.48%	1.32%
4	14	87	2911	2998	41.67%	2.90%
5	15	88	1471	1559	21.67%	5.64%

TABLE 55: PD GRADE DISTRIBUTION (SOURCE: OWN CALCULATIONS)

[23] 1- Recovery rate = Loss Given Default (LGD)

The trend of the bad rate versus the population across the PD grading scale can be seen in Figure 34. The trend of the default rate is correct. The higher is the PD grade, the bigger is the number of bads.

The difficulty here would be to recommend the correct cut-off point. Based on the relationship, we can see that the bad rate is quite small in all PD grades, but the majority of the population is in the high risk 4 grades. If we select as cut-off point PD grade 14, we will approve almost 79 % of all applications, accepting a default rate of 4.71 %.
A lower cut off would exclude big portion of the applications and would have a negative on the business.

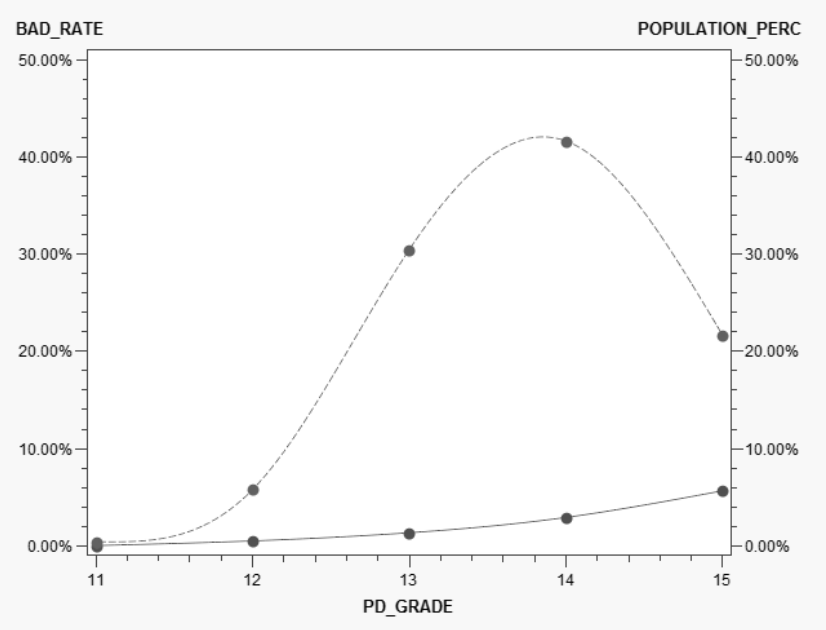

FIGURE 34: BAD RATE VS POPULATION DISTRIBUTION (SOURCE: OWN CALCULATIONS)

No method comes only with positive sides. The two methods represented above have their disadvantages and a researcher or users need to understand both the advantages and the disadvantages to assess if the advantages and the performance of such method can overcome the disadvantages or if there is a different method that can be used or modifications that can be done in order to get better results and ensure that the risk of losing money is well managed.

This method gives a solution for problems with use of categorical variables, although requiring them to be continuous. There are also the other inevitable shortcomings: the model assumes that the distributions of independent variables are normally distributed, but in practice the data are often not completely normal distribution, resulting in the unreliability of statistical results. A disadvantage of this approach is that it does not yield estimated PDs.

A new alternative to the standard methods, requiring only ex post information could represent the DEA.

9.3.3 DEA results

The outcome of the DEA is the Efficiency score. In comparison to the other methods, it is not an absolute measure of the default. It is a relative efficiency of the company within the analysed group of units. This is a certain disadvantage.

On the other hand, we have seen that the low default rate is causing difficulties for both standard methods. With no sufficient numbers in each class, the model does not have enough to build on. This limitation is not valid in the case of the DEA. This method doesn't need any previous separation and the calculations take into account the whole distributions of values of the predictors.

Apart from the efficiency score, we get extra information about the target values of the predictors. This can also be useful to understand how far the unit in question is from the ideal situation within the sample.

Rank	Efficiency			N	Bad flag	
	Min	Mean	Max		Sum	Mean
0	1	1	1	846	5	1.0%
1	0.99	0.99	1	457	6	1.0%
2	0.95	0.97	0.99	627	7	1.0%
3	0.9	0.93	0.95	645	12	2.0%
4	0.81	0.85	0.89	643	21	3.0%
5	0.69	0.75	0.81	641	20	3.0%
6	0.53	0.61	0.69	648	20	3.0%
7	0.35	0.44	0.53	644	26	4.0%
8	0.2	0.27	0.35	643	21	3.0%
9	0	0.11	0.2	644	49	8.0%
All	0	0.69	1	6438	187	3.0%

TABLE 56: EFFICIENCY DISTRIBUTION VS DEFAULT RATE (SOURCE: OWN CALCULATIONS)

9.4 Final comparison

The discriminative power of the scores (PD in case of logistic regression and Discriminant Analysis and Efficiency score in case of DEA) is the main indicator of performance of a credit scoring model.
The main purpose is to identify the bad units to avoid losses, therefore the KS statistic that we calculated for all three methods, is a good comparison measure between the three methods.

The KS statistic represents the maximum difference between cumulative percentages of bads and goods for each value of the scores.

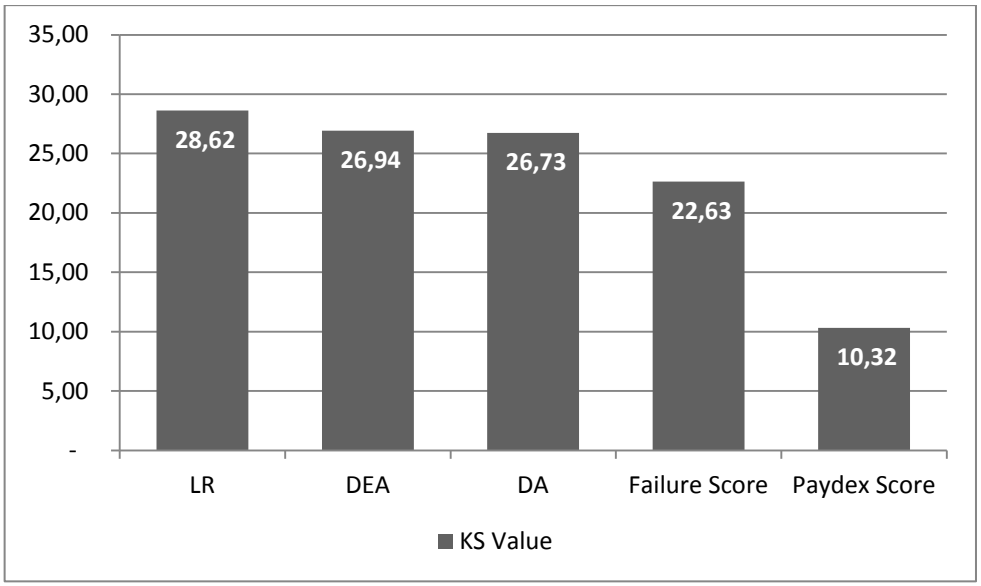

FIGURE 35: DISCRIMINATIVE POWER OF ALL METHODS (SOURCE: OWN CALCULATIONS)

As we can see, the best performance reached the PD assigned via logistic regression. This means that the probability of default is the best measure to identify the defaulted cases among all methods. Second best is the DEA Efficiency score and not far away is the Discriminant analysis probability.

How the final PDs and score follow the default rate is another way to compare the performance of the models. The scores get sorted from the worst to the best and ranked into 10 groups. Each group is assigned the number of bads. The bad rate is then calculated as the percentage of the bas within the group population. The bad rate should be decreasing, if the best scores are assigned to good applications. All three methods stand well in this comparison and provide reliable results.

	DEA			**Discriminant Analysis**			**Logistic Regression**		
Rank	% Population DEA	Bad flag		% Population DA	Bad flag		% Population LR	Bad flag	
		Sum	Bad Rate DEA		Sum	Bad Rate DA		Sum	Bad Rate LR
0	10%	49	8%	9%	52	8%	10%	54	7%
1	10%	26	4%	11%	32	4%	10%	31	4%
2	10%	21	3%	12%	35	4%	10%	33	5%
3	10%	20	3%	10%	18	3%	10%	19	3%
4	10%	20	3%	8%	16	3%	10%	17	2%
5	10%	21	3%	12%	19	2%	10%	14	2%
6	10%	12	2%	7%	11	2%	7%	13	3%
7	13%	5	1%	13%	11	1%	13%	12	1%
8	7%	6	1%	8%	7	1%	11%	8	1%
9	10%	7	1%	11%	5	1%	9%	5	1%
All	100%	187	3%	100%	206	3%	100%	206	3%

TABLE 57: SCORE PREDICTION ALIGNEMENT COMPARISON (SOURCE: OWN CALCULATIONS)

The trend is better seen in Figure 36. The worst scored units display higher default rates, while the best scores would only allow for 1% of bads.

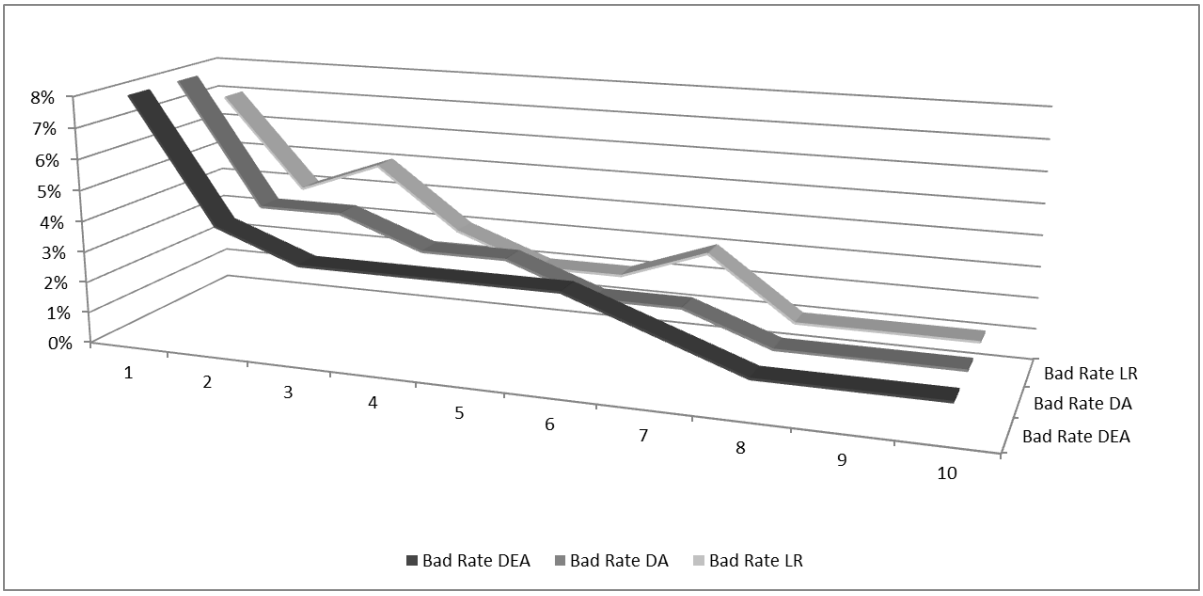

FIGURE 36: BAD RATE TREND COMPARISON (SOURCE: OWN CALCULATIONS)

10 BREAKING THE BLACK-BOX: NETWORK DEA

Network DEA is a fairly new term and methodology introduced in 1995. While traditional DEA focuses on organisational level measures, it does not provide sufficient detail for management to identify the specific sources of inefficiency embedded in interactions among business divisions that comprise the organisation.

First attempts to apply the Network DEA in risk management appeared few years after its introduction. (Matthews, 2011) evaluated bank performance in risk management practices using a Network DEA approach where an index of risk management practice and an index of risk management organisation are used as intermediate inputs in the production process.

Network DEA facilitates managerial insight regarding specific areas of improvement at various levels or stages. Identifying the structure of Sub-DMUs is a challenging exercise. Sexton & Lewis suggest a simple configuration that can describe many common organizational structures (Sexton & Lewis, 2003).

- Acquisition/Production: The Stage 1 Sub-DMU acquires resources required for production in Stage 2.
- Marketing/Sales: The Stage 1 Sub-DMU attracts potential customers and the Stage 2 Sub-DMU closes the sales.
- Processing/Finishing: The outputs of one machine are the inputs to another.

In the case of a bank, labour and fixed capital can be used to generate deposits, which in turn is used to generate interest earning assets. The deposits can be viewed as an intermediate output which is an intermediate input to produce interest bearing assets in the second stage of production.

In general, we have two types of credit models. The first type is called Application Models. These kinds of models are built to provide credit check on applications. In other words, this is the first credit decision and grade that is assigned to the customer. These models are based on mostly financial variables that come from a credit bureau or the application forms. Credit Bureaus usually represent the main source of financial and other data on customers. The automated models ensure immediate but thorough credit checks. The second type is called Behavioural Models that are used to track the payment behaviour of the customer. If the customer's behaviour is bad, the bank can lower its credit limit and increase the PD and therefore reduce its exposure to the client or put the customer of a watch list. These models are based on the borrower's own payment history with the bank.
This article assesses the possibility of application of the two stage network DEA to the credit scoring models. As mentioned above, the traditional DEA cannot address analysis of sub-processes, but only considers the overall efficiency of the whole system. Network DEA can be decomposed into a product of efficiencies of the sub-processes.

In credit scoring, we can look at the system as the combination of the two model types where the final efficiency is an overall measure of each unit. The combination of the above types allows for better evaluation of the company's financial situation as well as its payment behaviour towards the bank. The financial situation doesn't always follow the payment history and vice versa. The outcome of the standard approach would be discrimination between bads and goods based on their efficiency ratios. We can think of the two types of the credit scoring models as the two stages of the network or system. And we can analyse the efficiency based on the application data and the obligor's behaviour separately.

The data consists of a group of units (DMUs) and the values of their inputs and outputs to be included in the analysis. Here the group of units is represented with credit applicants. SAS has been used to enable all data manipulation and the DEA model application, using the procedures described in the previous section. Taking as an example a subset of 100 decision making units (DMUs) from the original dataset, the following scenarios will be considered and their results compared:

1. Traditional CRS will be applied on the aggregated system – the so called black box
2. Network DEA model will be applied on the two stages, taking into account the links between the two models

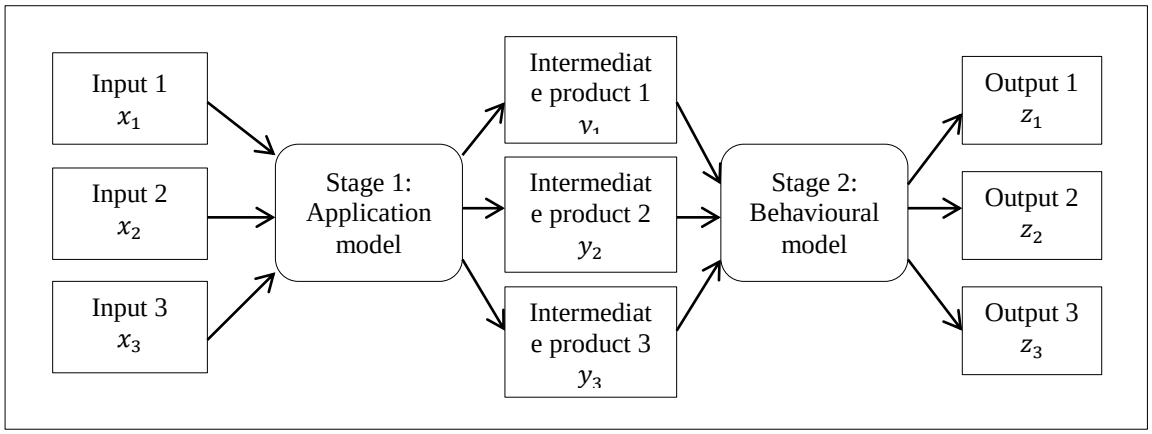

FIGURE 37: TWO-STAGE NETWORK (SOURCE: OWN CALCULATIONS)

Scenario 1: Using the traditional approach means looking at the entire system as one piece and calculating the overall efficiency. We cannot calculate the impact of each stage on the efficiency.

ID	Description	Type
x_1	Delivered Risk Indicator	Application
x_2	No of Accts for Collections in L36M	Application
x_3	Financial Strength Indicator	Application
z_1	Number of prompt payments in last 12 months	Behavioural
z_2	Number of late payments 1 to 30 days in last 12 months	Behavioural
z_3	Average current overdue balance	Behavioural

TABLE 58: AGGREGATED MODEL VARIABLES (SOURCE: OWN CALCULATIONS)

In traditional DEA, every activity belongs to either the input or output, but never to both of them. Thus they cannot deal with intermediate products (Tone & Tsutsui, 2008).

Scenario 2: The first stage is at the application time, when the customer first applies for a certain type of credit. At this stage, the bank or financial institution looks at the financial data they can get to assess the efficiency of the customer to turn inputs, such as assets and other investments into a growing business, showing the increasing ratios of liquidity, productivity and size of the business itself. The company utilizes its equity, assets and borrowings to produce revenues and profits. All the outputs from the first stage are the only inputs to the second stage. The outputs from the first stage to the second stage are called intermediate measures or products.

ID	Description	Type
x_1	Delivered Risk Indicator	Application
x_2	No of Accts for Collections in L36M	Application
x_3	Financial Strength Indicator	Application
y_1	Time since start-up in days	Application
y_2	Percentile Failure Score	Application
y_3	Number of Current Directors	Application
z_1	Number of prompt payments in last 12 months	Behavioural
z_2	Number of late payments 1 to 30 days in last 12 months	Behavioural
z_3	Average current overdue balance	Behavioural

TABLE 59: NDEA MODEL VARIABLES (SOURCE: OWN CALCULATIONS)

The overall efficiency score is the weighted mean of the Stage 1 and Stage 2 scores. The weights are determined in accordance with the importance given to the particular Stage. A DMU is overall efficient if and only if it is efficient at both stages.

The results of the 2 scenarios above are displayed in the table below (a sample of 10 DMUs are displayed).

Units	Aggregated Efficiency	NDEA Overall Efficicency	Stage 1 Efficiency (0.4 weight)	Stage 2 Efficiency (0.4 weight)
1	79.46%	88.77%	100.00%	71.93%
2	100.00%	100.00%	100.00%	100.00%
3	43.30%	87.12%	100.00%	67.79%
4	22.55%	78.08%	100.00%	45.19%
5	22.10%	50.44%	17.41%	100.00%
6	37.49%	75.60%	91.00%	52.49%
7	12.99%	27.14%	18.36%	40.31%
8	38.88%	71.37%	100.00%	28.42%
9	19.44%	67.92%	100.00%	19.80%
10	33.32%	73.97%	100.00%	34.93%

TABLE 60: AGGREGATED AND NDEA SCORES (SOURCE: OWN CALCULATIONS)

As can be seen in Figure 38, the efficient units do get recognized by both approaches. However, looking at both phases separately can unveil hidden inefficiencies that aren't visible when looking at the entire system. Where some units might look efficient in one phase, they are not doing well in the second phase of the network (e.g. units 1,3,4 on the graph).

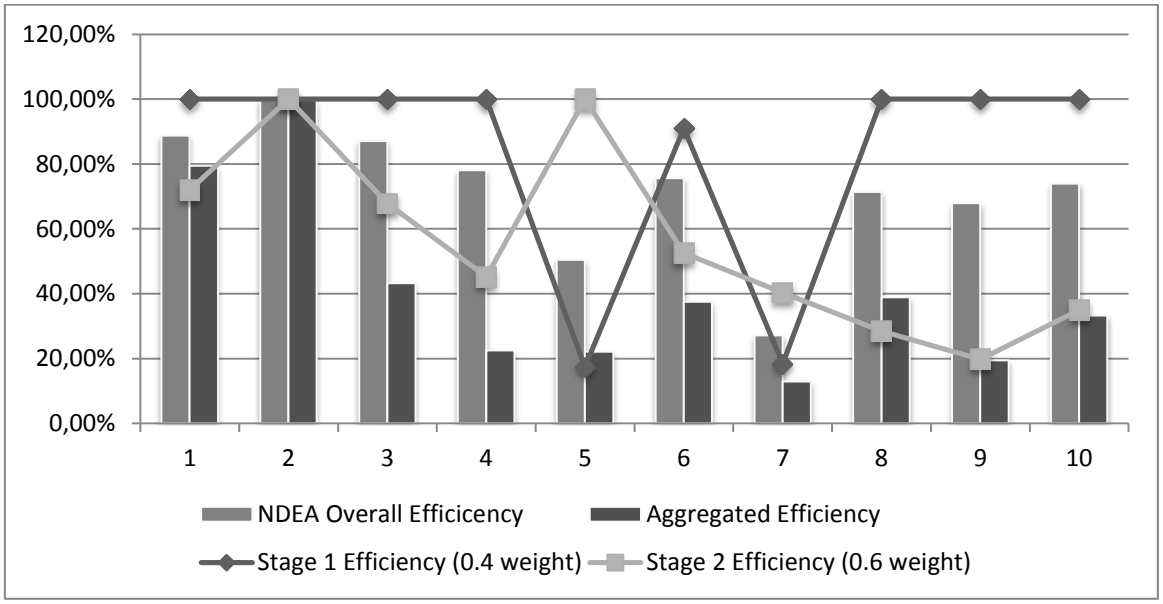

FIGURE 38: OVERALL DIFFERENCE BETWEEN TWO SCENARIOS (SOURCE: OWN CALCULATIONS)

CONCLUSION

The financial crisis of 2008-to date has revealed significant weaknesses in Financial Institutions' risk management in general and credit scoring in particular. The grading of customers and the models used appeared to be insufficient and lead to the situation where banks hold less capital than needed to overcome growing bad book. The purpose of this paper was to suggest a new approach to the credit scoring and to apply two standard methods and the new approach based on the DEA methodology on the same data.

With the right amount of knowledge and openness to try new ideas, financial institutions could potentially reap the benefits of applying novel analytical. This use of innovation for modelling their credit risk portfolios would also encourage institutions to not fall behind in other sectors in the use of novel analytical techniques, as well as challenge the regulators to show that advanced analytical techniques can in fact lead to better models and better estimations of risk (Brown, 2012).

There is merit to using regression techniques and the discriminant analysis due to their clarity and ease of use. Furthermore, the logistic regression provides results in the form of the percentage of probability of default. But it can perform poorly when it comes to low portfolios or portfolios with many outliers. Advanced analytical techniques need to be fully understood before data is thrown into them.

DEA is non-parametric, can handle multiple inputs and outputs and doesn't need prior information about classification of the units. These are the main advantages that the DEA methodology can offer to the credit scoring managers. Beside these, we can account for no need of prior specification of weights or functional forms of relations between inputs and outputs, transparency and user friendliness compared to black box methods such as neural networks.

The scoring model based on the DEA approach can be run at a granular level. The banks might prefer to compare performance of their clients within each sector, country or other grouping based on various characteristics. Furthermore, this approach could help the banks to identify their most efficient clients and give them priority in granting a loan or offering them other services. On the other hand, the DEA could serve in identifying the worst cases of the bunch and put them a watch list and hold provisions for the event of default.

The solution of the DEA model provides not only the relative efficiency level but the peer group and the targets outputs and inputs. The target values provide information about potential improvement of the less performing clients.

What is certain is that none of the standard methods neither the DEA can deal with bad quality data. If the supplied financial indicators of the borrowers are far from being correct, the results will surely be significantly biased. This is the main task, before applying any of the above methods, to ensure that their databases are clean and accurate.

Scoring borrowers as accurately as possible and identifying the bad ones, can save the bank a lot of money and helps therefore to overcome any sort of recession.
The diagram in Figure 39 shows how inaccurate scoring influences various levels of banks decisions.

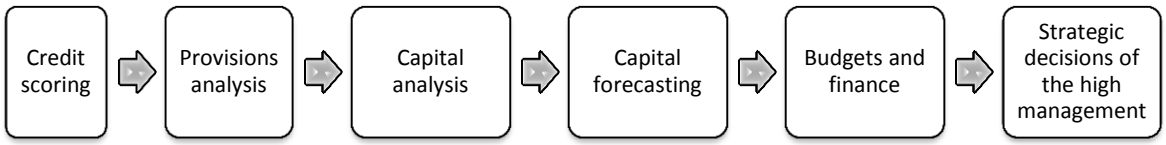

FIGURE 39: DECISION LEVELS (SOURCE: OWN CHART)

The credit grading and modelling especially play a significant role for the bank. An accurate model helps the bank to better manage its finances and to meet the strict regulatory requirements. The model performance is key to capital adequacy. And the cost of granting a loan to a defaulter is much higher than the cost of rejecting a good applicant.

The purpose of this thesis was to show how recently introduced modifications of the DEA models can contribute to the development of alternative approaches to the credit scoring problems. The difficulties the risk managers have to face remain the same. Every application for credit has to be checked and a decision taken if approval is recommended or not. These decisions need to be made as fast as possible, but thorough at the same time.

The reason for developing statistically based models that can be implemented into the credit systems is to provide an easy way for the underwriters to ensure of the repayment capacity of the applicant. The concern is not only about making a wrong credit recommendation, but to avoid putting restrictions on the business.

The Data envelopment analysis and its modifications can offer new and innovative ways of managing risks. There is no need for prior classification of bads and goods and the NDEA can put some light onto the different stages of credit scoring, the application stage and the behavioural stage. This is another step towards considering the DEA models as an alternative methodology to be used in credit scoring.

REFERENCES

Altman. (1968). Financial ratios, discriminant analysis and the Prediction of Corporate Bankruptcy. *Journal of Finance, 23*(4), 589-609.

Altman, E. (2000). Predicting financial distress of companies: Revisiting the Z-Score and Zeta models.

Altman, E. I., & Saunders, A. (2001). An analysis and critique of the BIS proposal on capital adequacy and ratings. *Journal of Banking & Finance, 25*(1), 25-46.

Andersen, P., & Petersen, N. (1993). A procedure for ranking efficient units in data envelopment analysis. *Management Science, 39*(10), 1261-1265.

Angelini, E., Tollo, G. d., & Roli, A. (2008). A neural network approach for credit risk evaluation. *The Quarterly Review of Economics and Finance, 48*(4), 733-755.

Arnold, V., Bardhan, I., Copper, W. W., & Gallegos, A. (1998). *Primal and dual optimality in computer codes using twostage solution procedures in DEA.*

Avkiran, N. K. (2010). Sensitivity analysis of network DEA illustrated in branch banking. Working Paper Series No. WP12/2010.

Bailey, M. (2004). *Credit Scoring: The Principles and Practicalities.* White Box Publishing.

Banasik, J., & Crook, J. (2003). Sample selection bias in credit scoring models. *Journal of the Operational Research Society*, 822-832.

Banker, R. D., & Morey, R. C. (1986). Efficiency Analysis for Exogenously Fixed Inputs and Outputs. *Operations Research, 34*(4), 513-521.

Banker, R. D., Cooper, W. W., Seiford, L. M., Thrall, R. M., & Zhu, J. (2004). Returns to scale in different DEA models. *European Journal of Operational Research, 154*(2), 345-362.

Banker, R. D., Charnes, A., & Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management science*.

Banker, R., & H.Chang. (2006). The super-efficiency procedure for outlier identification, not for ranking efficient units. *European Journal of Operational Research, 175*(2), 1311-1320.

Barr, R. (2004). DEA Software Tools and Technology: A State-of-the-Art Survey. V W. W. Cooper, L. M. Seiford, & J. Zhu, *Handbook on Data Envelopment Analysis* (1. vyd., stránky 539-566). Boston: Springer (Kluwer Academic Publishers).

Basel Commitee on Banking Supervision. (2004). An Explanatory Note on the Basel II IRB Risk Weight Functions.

Basel Commitee on Banking Supervision. (2005). An Explanatory Note on the Basel II IRB Risk Weight Functions.

Basel Committee of Banking Supervision. (1988). International convergence of capital measurement and capital standards.

Beaver, W. H. (1966). Financial Ratios as Predictors of Failure. *Journal of Accounting Research, 4*, 71-111.

Benjamin, N., Cathcart, A., & Ryan, K. (2006). Low Default Portfolios: A Proposal for Conservative Estimation of Default Probabilities.

Berg, S. A., Førsund, F. R., & Jansen, E. S. (1992). Malmquist Indices of Productivity Growth during the Deregulation of Norwegian Banking, 1980-89. *The Scandinavian Journal of Economics, 94*, 211-228.

Bernstein, P. L. (1996). *Against the Gods: The remarkable story of risk.* New York: John Wiley and Sons.

Boussofiane, A., Dyson, R., & Thanassoulis, E. (1991). Applied data envelopment analysis. *European Journal of Operational Research*, 1-15.

Brown, I. (2012). *Five key challenges in credit risk modelling.* Načteno z SAS.
Cannata, F., & Quagliariello, M. (14. January 2009). The Role of Basel II in the Subprime Financial Crisis: Guilty or Not Guilty? *SSRN: http://ssrn.com/abstract=1330417*.
Caves, D. W., Christensen, L. R., & Diewert, W. E. (1982). The Economic Theory of Index Numbers and the Measurement of Input, Output and Productivity. *Econometrica, 50*(6), 1393-1414.
Commission of the EU. (2003). Commission Recommendation concerning the definition of micro, small and medium-sized enterprises. *Official Journal of the European Union*, 36-41.
Cook, R. D. (1977). Detection of Influential Observations in Linear Regression. *Technometrics, 19*(1), 15-18.
Cramer, J. (2000). Scoring Bank Loans That May Go Wrong: a Case Study. Tinbergen Institute Discussion Paper.
Crook, J., & Banasik, J. (2004). Does rejectinference really improve the performance of application scoring models? *Journal of Banking and Finance, 28*(4), 857-874.
Crosbie, P., & Bohn, J. (2003). Modeling Default Risk. Moody's KMV Company.
Czepiel, S. A. (2011). Maximum likelihood estimation of logistic regression models: Theory and implementation.
Desai, V. S., Crook, J. N., & Overstreet, G. A. (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research, 95*(1), stránky 24-37.
Durand, D. (1941). Risk Elements in Consumer Instalment Financing.
Eisenbeis, R. A. (1978). Problems in applying discriminant analysis in credit scoring models. *Journal of Banking & Finance, 2*(3), 205-219.
Emrouznejad, A. (2000). An Extension to SAS/OR for Decision System Support. *SUGI 25 Proceedings.* Indiana.
Emrouznejad, A. (2006). A SAS® Application for Measuring Efficiency and Productivity of Decision Making Units. *SUGI 27.*
Emrouznejad, A. (nedatováno). *DEA Zone*. Získáno 2012, z http://www.deazone.com
Färe, R., & Grosskopf, S. (1996). Productivity and Intermediate Products: A Frontier Approach. *Economics, 50*, 65-70.
Färe, R., & Grosskopf, S. (2000). Network DEA. *Socio-Economic Planning Sciences, 34*, 35-49.
Färe, R., & Whittaker, G. (1995). An Intermediate Input Model of Dairy Production using Complex Survey. *Journal of Agricultural Economics, 46*(2), 201-213.
Färe, R., Grosskopf, S., & Whittaker, G. (2007). *Modeling Data Irregularities and Structural Complexities in Data Envelopment Analysis.* Springer US.
Färe, R., Grosskopf, S., Roos, P., & Lindgren, B. (1992). Productivity changes in Swedish Pharmacies 1980-1989: A Non-Parametric Malmquist Approach. *Journal of Productivity Analysis, 3*, stránky 85-101.
Farrell, M. J. (1957). The Measurement of Productive Efficiency. *Journal of the Royal Statistical Society* , stránky 253-290.
Feruś, A. (2008). The DEA method in managing the credit risk of companies. *Ekonomika*, 109-118.
Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annale of Eugenics*.
Golany, B., & Roll, Y. (1993). Alternate methods of treating factor weights in DEA. *OMEGA Int. J. of Mgmt Sci., 21*, stránky 99-109.
Greene, W. H. (2003). *Econometric Analysis.* New Jersey.

Grifell-Tatjé, E., & Lovell, C. (1996). Profits and Productivity. *Management Science, 45*(9), 1177-1193.
Hand, D. J., & Henley, W. E. (1997). *Journal of the Royal Statistical Society Series A, 160*(3), stránky 523-541.
Hosmer, D. W., & Stanley, L. (1989). *Applied Logistic Regression.* John Wiley & Sons, Inc.
Huang, C.-L., Chen, M.-C., & Wang, C.-J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications, 33*(4), 847-856.
Hughes, J. P., Lang, W., Mester, L. J., & Moon, C.-G. (1996). Efficient Banking under Interstate Branching. *Journal of Money, Credit and Banking, 28*(4), 1045-1071.
Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research, 2*(6), 429-444.
Chen, Y. (2005). Measuring super-efficiency in DEA in the presence of infeasibility. *European Journal of Operational Research, 161*(2), 545-551.
Chen, Y. (2005). On preference structure in data envelopment analysis. *International Journal of Information Technology & Decision Making, 4*(3).
Chen, Y., & Zhu, J. (2004). Measuring Information Technology's Indirect Impact on Firm Performance. *Information Technology and Management, 5*(1-2), 9-22.
Chen, Y., Liang, L., & Zhu, J. (2009). Equivalence in two-stage DEA approaches. *European Journal of Operational Research, 193*(2), 600-604.
Cheng, E. C. (2007). Alternative approach to credit scoring by DEA: Evaluating borrowers with respect to PFI projects. *Building and Environment*.
International Organization for Standardization. (2009). ISO/DIS 31000: Risk management — Principles and guidelines on implementation.
Jennrich, R. I., & Sampson, P. F. (1976). Newton-Raphson and Related Algorithms for Maximum Likelihood Variance Component Estimation. *Technometrics, 18*(1), 11-17.
Kao, C., & Hwang, S.-N. (2008). Efficiency decomposition in two-stage data envelopment analysis: An application to non-life insurance companies in Taiwan. *European Journal of Operational Research, 185*(1), 418-429.
Kearney, T. D. (1999). Advances in Mathematical Programming and Optimization in the SAS System. *SUGI 24 Proceedings.* Cary: SAS Institute Inc.
Koopmans , T. C. (1951). Analysis of production as an efficient combination of activities. V T. C. Koopmans, *Activity analysis of production and allocation* (Sv. 58). Cowles Commission Monograph.
Krink, T., Paterlini, S., & Resti, A. (2008). The Optimal Structure of PD Buckets. *Journal of Banking & Finance, 32*(10), 2275–2286.
Lee, T.-S., & Chen, I.-F. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications, 28*(4), 743-752.
Lopez, J. A. (2000). Evaluating Credit Risk Models. *Journal of Banking & Finance*.
Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 49-55.
Malmquist, S. (1953). Index numbers and indifference surfaces. *Trabajos de Estadistica y de Investigacion Operativa, 4*(2), 209-242.
Markowitz, H. (1952). Portfolio selection. *The Journal of Finance, 7*, 77-91.
Matthews, K. (28. March 2011). Risk Management and Managerial Efficiency in Chinese Banks: A Network DEA Framework . HKIMR Working Paper No.10/2011. Načteno z SSRN: http://ssrn.com/abstract=1797468
McDonald, J. F., & Moffitt, R. A. (1980). The Uses of Tobit Analysis. *The Review of Economics and Statistics, 62*(2), 318-321.

Medema, L., Koning, R. H., & Lensink, R. (2009). A Practical Approach to Validating a PD Model. *Journal of Banking & Finance*.
Menard, S. (2002). *Applied Logistic Regression Analysis.* Sage Publications.
Merton, R. C. (1973). On the pricing of corporate debt The risk structure of interest rates. *Journal of Finance*.
Min, J. H., & Lee, Y. (2004). A Practical Approach to Credit Scoring. *ICEB.*
Mok, J.-M. (2009). Reject Inference in Credit Scoring.
Myers, J. H., & Forgy, E. W. (1963). The Development of Numerical Credit Evaluation Systems. *Journal of the American Statistical Association, 58*(303), 799-806.
Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Predictionof Bankruptcy. *Journal of Accounting Research, 18*(1), 109-131.
Olson, D., & Wu, D. D. (2010). Enterprise risk management: a DEA VaR approach in vendor selection. *International Journal of Production Research, 48*(16), 4919-4932.
Ong, C.-S., Huang, J.-J., & Tzeng, G.-H. (2005). Building credit scoring models using genetic programming. *Expert Systems with Applications*, 41-47.
Paradi, C. J., Asmild, M., & Simak, P. C. (2004). Using DEA and Worst Practice DEA in Credit Risk Evaluation. *Journal of Productivity Analysis, 21*(2), 153-165.
Plummer, W. C., & Young, R. A. (1940). *Sales Finance Companies and Their Credit Practices.*
Rao, C. R. (1973). *Linear Statistical Inference and its Applications.* John Wiley & Sons, INC.
Rencher, A. C. (2002). *Methods of Multivariate Analysis, Second Edition.* A JOHN WILEY & SONS, INC. PUBLICATION.
Sabah, S. (2011). The Final Frontier: A SAS Approach to Data Envelopment Analysis. *SAS Global Forum Online Proceedings.*
SAS Institute Inc. (nedatováno). Načteno z Business Analytics and Business Intelligence Software: www.sas.com
SAS Institute Inc. (1999). SAS/STAT(R) 9.2 User's Guide, Version 8.
SAS Institute Inc. (2010). SAS/OR® 9.22 User's Guide: Mathematical Programming.
SAS Institute Inc. (2010). *SAS/OR® 9.22 User's Guide: Mathematical Programming.* Cary: SAS Institute Inc.
Sexton, T. R., & Lewis, H. F. (2003). Two-Stage DEA: An Application to Major League Baseball. *Journal of Productivity Analysis, 19*, 227–249.
Shuai, J. J., & Li, H. L. (2005). Using Rough Set and Worst Practice DEA in Business Failure Prediction. *Lecture notes in Computer Science, 3642*, 503-510.
Siddiqi, N. (2006). *Credit Risk Scorecards Developing and Implementing Intelligent Credit Scoring.* John Wiley & Sons.
Soriano, J. (1993). Global minimum point of a convex function. *Applied Mathematics and Computation*, 213-218.
Statistical Consulting Group. (nedatováno). *Introduction to SAS.* Načteno z UCLA: Academic Technology Services: http://www.ats.ucla.edu/default.htm
Tamari, M. (1966). Financial Ratios as a Means of Forecasting Bankruptcy. *Management International Review, 6*(4), 15-21.
Tone, K., & Tsutsui, M. (nedatováno). Network DEA: A slacks-based measure approach.
Troutt, M., Rai, A., & Zhang, A. (1996). The potential use of DEA for credit applicant acceptance systems. *Computers & Operations Research, 23*(4), 405-408.
Vasicek, O. A. (2002). The distribution of loan portfolio value. *Journal of Risk*.
West, D. (2000). Neural network credit scoring models. *Computers and Operations Research*.
Yobas, M. B., Crook, J. N., & Ross, P. (2000). Credit scoring using neural and evolutionary techniques. *IMA Journal of Mathematics Applied in Business and Industry, 11*, stránky 111-125.

Zhu, J. (1996). Data Envelopment Analysis with Preference Structure. *The Journal of the Operational Research Society, 47*(1), 136-150.
Zhu, J. (2001). Super-efficiency and DEA sensitivity analysis. *European Journal of Operational Research, 129*(2), 443-455.
Zhu, J. (2009). *Quantitative Models for Performance Evaluation and Benchmarking.* Springer.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Full form
BCC	Banker Charnes Cooper
BIS	Bank of International Settlements
CCR	Charnes, Cooper and Rhodes
CRS	Constant returns to scale
DEA	Data envelopment analysis
DMU	Decision making unit
EL	Expected Loss
FICO	Fair Isaac Company
IMF	International Monetary Fund
OLS	Ordinary Least Squares
PLE	Public Large Enterprise
IRB	Internal Ratings Based
LGD	Loss Given Default
M	Maturity (effective maturity)
MLE	Maximum Likelihood Estimation
PD	Probability of default
NDEA	Network Data Envelopment Analysis
RWA	Risk weighted assets
SME	Small and medium enterprises
VRS	Variable returns to scale

APPENDIX

Default rates

Year	Month	Goods	Bads	All	Bad Rate		Year	Month	Goods	Bads	All	Bad Rate
2008	1	1343	49	1392	3,5%		2010	7	104	1	105	1,0%
2008	2	589	16	605	2,6%		2010	8	100	4	104	3,8%
2008	3	380	12	392	3,1%		2010	9	87	1	88	1,1%
2008	4	320	7	327	2,1%		2010	10	80	1	81	1,2%
2008	5	215	5	220	2,3%		2010	11	80	1	81	1,2%
2008	6	173	3	176	1,7%		2010	12	74		74	0,0%
2008	7	205	9	214	4,2%		2011	1	14		14	0,0%
2008	8	178	3	181	1,7%		2011	2	6		6	0,0%
2008	9	171	6	177	3,4%		2011	3	16		16	0,0%
2008	10	246	8	254	3,1%		2011	4	10		10	0,0%
2008	11	153	5	158	3,2%		2011	5	7		7	0,0%
2008	12	150	4	154	2,6%		2011	6	7	1	8	12,5%
2009	1	158	4	162	2,5%		2011	7	4	1	5	20,0%
2009	2	151	7	158	4,4%		2011	8	6		6	0,0%
2009	3	156	7	163	4,3%		2011	9	4		4	0,0%
2009	4	118	2	120	1,7%		2011	10	8		8	0,0%
2009	5	93	6	99	6,1%		2011	11	6		6	0,0%
2009	6	113	4	117	3,4%		2011	12	11		11	0,0%
2009	7	112	2	114	1,8%		2012	1	9		9	0,0%
2009	8	85	3	88	3,4%		2012	2	10		10	0,0%
2009	9	112	2	114	1,8%		2012	3	9		9	0,0%
2009	10	122	4	126	3,2%		2012	4	14		14	0,0%
2009	11	108	1	109	0,9%		2012	5	5	1	6	16,7%
2009	12	124	3	127	2,4%		2012	6	5		5	0,0%
2010	1	116	7	123	5,7%		2012	7	4		4	0,0%
2010	2	110	4	114	3,5%		2012	8	6		6	0,0%
2010	3	151	2	153	1,3%		2012	9	5		5	0,0%
2010	4	103	1	104	1,0%		2012	10	9		9	0,0%
2010	5	95	2	97	2,1%		2012	11	11		11	0,0%
2010	6	106	7	113	6,2%		2012	12	21		21	0,0%

Data mining SAS code

```
/*Basic statistics*/

proc means data=model_data N NMiss min p1 p25 mean p50 p75 p95
p99 max Mode StdDev;
run;

/*Correlation*/

PROC CORR DATA=model_data nomiss PLOTS(MAXPOINTS=NONE)
PLOTS=matrix(histogram);
VAR &sel_vars;
RUN;

/*Cluster Analysis*/

proc varclus data=model_data centroid outtree=tree;
var &sel_vars;
run;

/*Default Rate*/

proc freq data=model_data;
tables Bad flag*app_year/list missing out=BAD_RATE;
tables Bad flag*app_year*app_mth/list missing;
run;
```

KS Macro SAS code

```
/*Macro for KS calculation:*/

 %MACRO KS (DSNAME,PERFVAR,WGHTVAR,SCOREVAR);

 proc freq data=&DSNAME(keep= &perfvar &scorevar &wghtvar);
  where &perfvar=0;              *Bad;
 %if %length(&wghtvar) gt 0 %then %do;
 weight &wghtvar;
 %end;
 tables &scorevar / outcum noprint missing
out=freqbad(drop=cum_freq rename=(count=_nbad cum_pct=_cumbd
percent=_pbad));
  run;

  proc freq data=&DSNAME(keep= &perfvar &scorevar &wghtvar);
  where &perfvar=1;              *Good;
 %if %length(&wghtvar) gt 0 %then %do;
 weight &wghtvar;
 %end;
 tables &scorevar / outcum noprint missing
           out=freqgood(drop=cum_freq rename=(count=_ngood
cum_pct=_cumgd percent=_pgood));
  run;
```

```
  data freqmerg_&scorevar (keep=&scorevar _ngood _nbad _cumgd
_cumbd _pgood _pbad _cumdiff);
      merge freqgood (in=a) freqbad (in=b);
           by &scorevar;
  retain cumgd 0 cumbd 0;
  if a then cumgd=_cumgd;
  else do;
      _ngood=0;
      _pgood=0;
      _cumgd=cumgd;
 end;
  if b then cumbd=_cumbd;
  else do;
      _nbad=0;
      _pbad=0;
      _cumbd=cumbd;
 end;
  _pbad=_pbad/100.0;
  _pgood=_pgood/100.0;
  _cumbd=_cumbd/100.0;
  _cumgd=_cumgd/100.0;
  _cumdiff=abs(cumgd-cumbd);
  run;

  proc means data=freqmerg_&scorevar noprint;
     var _cumdiff;
      output out=outks_&scorevar max=KS_Value;
  run;
  title2 "&Scorevar";
  proc print data=outks_&scorevar noobs;
     var KS_value;
  run;
  title2 " ";

 %MEND KS;
```

Logistic regression SAS code

```
/*Logistic regression*/

proc logistic data=model_data simple namelen=50
    outmodel=parameters plots=EFFECT plots=ROC;
    model Bad flag(event='1')=&sel_vars. /rsquare
    selection=stepwise slentry=0.05  slstay=0.05 stb corrb
    outroc=roc lackfit;
 output out=outdata xbeta=logit p=estprob;
run;

/*Predicted and observed alignement*/

 title "Predicted and Actual alignment";
 proc rank data = outdata out = kgb_rank descending groups = 10;
       ranks p_rank ;
       var estprob;
```

```
run;

proc tabulate data =kgb_rank missing;
      class p_rank;
      var  Bad flag   estprob  ;
      table p_rank=' ' ALL,(estprob )*(min mean max sum) n (Bad
      flag )*(sum mean )/box= 'Predictive';
run;

/*Logistic Function*/

title 'Logistic function';
symbol1 v=circle l=32  c = green i=none;
proc gplot data = outdata;
plot estprob * logit ;
run;

/*Bad rate vs Population Distribution*/

title1 "Bad Rate vs Population Distribution";
symbol1 interpol=spline value=dot color=vibg height=1;
symbol2 interpol=spline value=dot color=depk height=1;

proc gplot data= all;
      plot bad_rate*PD_GRADE / vaxis=0 to 0.3 by 0.1;
      plot2 population_perc*pd_grade / frame
                          vaxis=0 to 0.3 by 0.1;
run;
quit;

/*Cross validation sample*/

%let K=5;
%let rate=%sysevalf((&K-1)/&K);

proc surveyselect data=model_data out=cv seed=231258
       samprate=&rate outall reps=10;
run;
```

Linear Discriminant SAS code

```
/*Variable selection for Discriminant Analysis*/

proc stepdisc
      data=model_data
      method=stepwise SLENTRY=0.05 SLSTAY=0.05;
class Bad flag;
var &sel_vars.;
run;

/*Cross validation sample*/

%let K=5;
%let rate=%sysevalf((&K-1)/&K);
```

```
 proc surveyselect data=model_data out=cvda seed=231258
  samprate=&rate outall;
 run;

data datest datrain;
 set cvda;
 if selected = 0 then output datest;
 if selected = 1 then output datrain;
 run;

proc means data=datrain print min mean max;
 var &inputs;
 class Bad flag;
run;

/*Discrimination analysis*/

PROC DISCRIM DATA=datrain
      method=normal
      POOL=yes
      outd=outdat
      outstat=DAResults
      testdata=datest
      testout=tout;
      PRIORS equal;
      CLASS Bad flag;
      VAR &inputs;
RUN;
```

DEA SAS code

The SAS code for the DEA efficiency calculation is using parts of the macros developed by Ali Emrouznejad (Emrouznejad, 2006).

```
/*Create Input and Output data sets */

      data Inputs;
      set cohort ;
      length DMU $ 4;
      DMU=_n_;
      rename
            &input1 = Input1
            &input2 = Input2
            &input3 = Input3
            ;
      keep DMU &inputs;

      data Outputs;
            set cohort;
            length DMU $ 4;
      DMU =_n_;
      rename
            &output1 = Output1
            &output2 = Output2
```

```
            &output3 = Output3;
      keep DMU &outputs;
      run;

      proc datasets library=work;
      delete efficiencies countins countouts dualoutput optoutput
      benchmarks;

      data inputs;
            set inputs;
            drop DMU;

      data inputs;
            set inputs;
            DMU =_n_;
            run;

      data outputs;
            set outputs;
            drop DMU;

      data outputs;
            set outputs;
            DMU =_n_;
      run;

/*DEA Macro*/

      %MACRO DEA_OPTMODEL(Inputdata=,Outputdata=);

      /*Sort Data by DMU*/
      proc sort data=&Inputdata;
            by DMU;
      run;
      proc sort data=&Outputdata;
            by DMU;
      run;

      /*Count DMUS*/
      data _null_;
            set &Inputdata;
            call symput('DMU_COUNTER',_N_);
      run;

      /*Count Inputs*/
      proc transpose data=&Inputdata out=countins;
      run;
      data _null_;
            set countins;
            call symput('_nInput',_N_-1);
      run;

      /*Count Outputs*/
      proc transpose data=&Outputdata out=countouts;
      run;
      data _null_;
```

```
      set countouts;
      call symput('_nOutput',_N_-1);
run;

/*Loop Through DMUs*/
%do LOOP_COUNT=1 %to &DMU_COUNTER;

proc optmodel printlevel=0;

      /*Declare Sets and Parameters*/
      set Inputs = 1.. &_nInput;
      set Outputs = 1.. &_nOutput;
      set <num> DMU;
      set n = /&LOOP_COUNT/;

      number scale{DMU,Inputs};
      number objective{DMU,Outputs};

      /*Read in Data*/
      read data &Inputdata
           into DMU = [DMU]
{d in Inputs} < scale[DMU,d]=col("INPUT"||d) >;


      read data &Outputdata
           into DMU = [DMU] {e in Outputs} <
objective[DMU,e]=col("OUTPUT"||e) >;

      /*Declare Variables and System of Equations*/
      var x{n,Inputs}>=0, y{n,Outputs}>=0;
      Max Efficiency = sum{i in n}(sum{j in
Outputs}(Objective[i,j]*y[i,j]));
      con Scaling: sum{i in n}(sum{j in
Inputs}(Scale[i,j]*x[i,j])) = 1;
      con CapToOne {i in DMU}: sum{a in n}(sum{b in
Outputs}(Objective[i,b]*y[a,b])) - sum{c in n}(sum{d in
Inputs}(Scale[i,d]*x[c,d])) <= 0;

      /*Call Solver*/
      solve;

      /*Create Benchmarking Dataset*/
      create data DualOutput from DMU=&LOOP_COUNT {h in DMU} <
col("BENCHMARK"||h)=CapToOne[h].dual >;

      /*Create Efficiency DataSet*/
      create data OptOutput from DMU = &LOOP_COUNT
Efficiency=Efficiency;

quit;

/*Append Report Set*/
proc append base=Benchmarks data=DualOutput;
run;

/*Append Plotting Set*/
```

```
proc append base=Efficiencies data=OptOutput;
run;

%end;

/*Merge Sets for Report*/
data FinalReport;
     merge Efficiencies Benchmarks;
     by DMU;
run;

/*Output Efficiency Table*/
proc print data=FinalReport noobs;
run;

/*Manipulate Data Set for Plotting*/
data Efficiencies;
     set Efficiencies;
     if int(efficiency) = 1 then EfficientTrue=Efficiency;
     else EfficientFalse=Efficiency;
run;

%MEND;

%DEA_OPTMODEL(Inputdata=Inputs, Outputdata=Outputs);
```

Censored Regression SAS code

```
proc qlim data=sasuser.DEAReport;
model Efficiency = &inputs;
endogenous efficiency ~ censored (lb=0 ub=1);
OUTPUT OUT=SASUSER.QLIMPred predicted;
run;


/* Define symbol characteristics */
symbol1 interpol=spline value=dot color=vibg height=1;
TITLE;
TITLE1 "Scatter Plot";
proc sgplot data=SASUSER.QLIMPred;
scatter x=Efficiency y=p_efficiency;
run;
```