A New Model for Predicting the Loss Given Default

Master Thesis Business Analytics

Author: P.W.F. Alons
Supervisor 1: drs. E. Haasdijk
Supervisor 2 : dr. M. Jonker
Supervisor ABN AMRO: MSc. A. Wyka
VU University Amsterdam
Faculty of Sciences
Master Business Analytics
De Boelelaan 1081a
1081 HV Amsterdam

March 2013

Preface

In the autumn of 2012 and the winter of 2013 I worked on this master thesis about a new Loss Given Default (LGD) model for the ABN AMRO. The master thesis is the final part of the master Business Analytics (before Business Mathematics & Informatics) at the VU University. The aim of this thesis is to combine Mathematics and Informatics on a study of a real problem in the Operation Research. The master thesis is part of an internship at a company of choice. My internship took place at the Credit Risk Modelling department of the ABN AMRO bank.

I would like to thank Evert Haasdijk for his critical insights during this research. I would also like to thank my supervisor Arek Wyka of the ABN AMRO for all his help. Finally I would like to thank all my colleagues for the nice months and for all their help during my research.

I hope you enjoy reading this thesis.

Summary

To quantify the risk that an obligor is not able to pay back his loan, it is important to know the creditworthiness of the counterparty. To quantify the creditworthiness of the counterparty banks would like to know what the expected losses on their credit loans outstanding to the counterparty are. There are, in general, four parameters, which influence this expected loss [3]:

- Probability of Default (PD)
- Exposure at Default (EAD)
- Loss Given Default (LGD)
- Maturity

Here the PD is the probability that the counterparty is not able to pay back the loan. The EAD is the extent to which the bank is exposed if the counterparty goes in default. The LGD is defined as the fraction of the EAD that cannot be recovered by a default. For a given maturity the expected loss can then be expressed in the following formula: $EL = PD * EAD * LGD$.

Since Basel II, it is possible for banks to compute own estimates for those parameters [3].

Within a bank, two different types of LGD are considered, namely the LGD for performing loans and the LGD for non-performing loans, also called the in-default LGD. Performing loans are loans that are not in default. Non-performing loans are loans that have been already in default, but the losses to the bank are not yet known as these are still in the recovery. Only after the recovery process it is known what is recovered and what is eventually lost.

The current model of the ABN focuses on the coverage ratio to predict the LGD. In the current model the LGD for performing loans is divided in classes, based on coverage ratio. The classes are divided in: unsecured, 0%-30%, 30%-60%, 60%-90%, 90%-120%, 120%-150%, and >150% covered. For each class the expected LGD is estimated. For the in-default LGD the ABN adjusts this LGD model for performing loans with another model (the so called UCR factor model), which depends on their Unified Credit Rating (UCR): the higher the UCR rating, the higher the UCR factor.

With the help of statistical and data mining techniques I tried to build a new model for both LGDs. The model techniques that are used in this thesis are linear regression, regression tree, model tree ,and decision tree.

First a new model for the in-default LGD was build. This was first tried with a dataset that contains both LGD as PD information. However, by combining the PD and LGD dataset 90% of the data was lost, because there was no rating information available for 90% of the instances. This resulted in a dataset that contained only 518 instances, from which 504 instances had no loss.

In the models, based on this small dataset, no PD variable was involved. That's why it was also tried to build a better model with the large dataset, consisting of 4791 instances. This dataset did contain PD variables, but these variables had only values for 518 of the 4791 instances. For all the other instances, these values were missing in the dataset. However, with this large dataset it was not possible either to find a better model for the in-default LGD than the current model.

Next, I tried to predict whether a loss occurred or not, and then tried to build separate models for the LGNL and LGL. Unfortunately, this did not result in a better model than the current model. Moreover, I tried to get a better model by predicting the total LGD, which is the combined LGD of the counterparty and the LGD of the guarantor. This did also not result in a better model than the current model.

Moreover, it was tried to build an in-default LGD model based on the cashflow dataset. This cashflow dataset contains the quarter snapshots of the provision. Predicting the LGD directly did not result in a better model than we already had. If we first predict whether a loss occurred or not, and then built models on both the LGL and LGNL separately, still no better model was received.

Finally, I also tried to find a model for the LGD for performing loans. Again first a model was built for the LGD directly. This gave already a slightly better RMSE and MAE than the current model. Second, it was tried to build a model for the total LGD again. Both the regression tree technique as the model tree technique gave lower RMSE and MAE than the current model.

So for the in-default LGD no better model was found in this thesis.

For the LGD of performing loans it is optimal to predict the total LGD. Which of the models to use, depends on what the bank really wants. With the regression tree it is possible to hold the same model structure as the current model, but now with different risk drivers involved. Figure 16 shows what the model looks like if we use this technique. .

The model tree is the most complex model and outputs a continuous function. The model tree cuts the distribution of the LGD into different pieces and predicts each piece of the distribution by a linear function. This results in a so called piecewise linear function. It might be difficult for the bank to implement a (piecewise) linear function. However, it can be possible to divide the output of this function into buckets. Then we have buckets based on the LGD, rather than on coverage ratio as in the current model.

For further research it would be interesting to add variables like seniority of the loan and the reason of default. Especially the last one can be interesting, as it might explain the high number of totally recovered instances.

For further research I would also recommend to try to get a much larger dataset with both LGD as PD variables. I would recommend investigating why there is no rating information available for 90% of the instances. This, because a dataset of 518 instances is too small to build real confident model on (at least with the techniques that are used in this thesis). Besides, that dataset also contains only 14 non-cured case, which is too less and not representative to build a confident model on.

So if it is possible to get a larger and more representative dataset with both PD and LGD variables, with also the variables seniority and reason of default added to this, then it might be possible to get a much better model with one of the techniques that are used in this thesis.

Table of Contents

1 Introduction

Risk Management is becoming an increasingly important part of banking. Risk management is about identifying, and quantifying risks and determining control measures. With control measures we mean the activities that affect the probability of occurrence or the impact of the risk events [1].

In 1988 it was decided to introduce a multinational accord to strengthen the stability of the international banking system, called the Basel accord. The Basel accord is a settlement between the banks containing recommendations on the banking laws and regulations [2]. In 2006 a review of this accord was made, which was called Basel II. According to these Basel accords a bank should hold enough capital behind to cover their risk.

To quantify the credit risk it is important to know the creditworthiness of the counterparty. To quantify the creditworthiness of the counterparty banks would like to know what the expected losses on their credit loans outstanding to the counterparty are. There are, in general, four parameters, which influence this expected loss [3]:

- Probability of Default (PD)
- Exposure at Default (EAD)
- Loss Given Default (LGD)
- Maturity

Here the PD is the probability that the counterparty is not able to pay back the loan. The EAD is the extent to which the bank is exposed if the counterparty goes in default. The LGD is defined as the fraction of the EAD that cannot be recovered by a default. For a given maturity the expected loss can then be expressed in the following formula: $EL = PD * EAD * LGD$.

So for a bank it is very important to estimate those parameters as well as possible.

Since the introduction of Basel II, it is possible for banks to compute their own estimates for those parameters [3]. It is now possible for the bank to develop their own internal ratings which they can use to make their own internal estimates of the parameters.

In this thesis I only consider the LGD. Within a bank two different types of LGD are considered, namely the LGD for performing loans and the LGD for non-performing loans. Performing loans are loans that are not in default. Non-performing loans are loans that have been already in default, but the losses to the bank are not yet known as these are still in recovery. Only after the recovery it is known what is recovered and what is eventually lost.

In this thesis I investigate if it is possible to build a new model, which predicts the LGD better than the current model of the ABN. I only focus on the LGD for the non-retail business of the ABN. The non-retail business of the ABN only focuses on loans, which are outstanding to corporate clients, rather than loans that are outstanding to private clients. Moreover, I look to both the LGD of performing loans as the LGD of non-performing loans, which is better known as the in-default LGD.

Probably one of the most important components of the LGD is the coverage ratio. This is the percentage of the credit loan that is covered with collateral. The current model of the ABN focuses on this coverage ratio to predict the LGD. In the current model the LGD for performing loans is divided in classes, based on coverage ratio. The classes are divided in: unsecured, 0%-30%, 30%-60%, 60%-90%, 90%-120%, 120%-150%, and >150% covered. For each class the expected LGD is estimated. For the in-default LGD the ABN adjusts this LGD model for performing loans with another model (the so called UCR factor model), which depends on their UCR rating: the higher the UCR rating, the higher the UCR factor.

The advantage of the current method is that it is very easy to understand for everyone who has to work with it. A disadvantage of this model is that the LGD is now only based on one risk driver.

Although the coverage ratio is probably one of the most important components of the LGD, it is probably not the only important component. Another important risk driver could be the provision that the banks take during the recovery period. At the end of each reporting period, which is each quarter, the ABN assesses whether impairments on loans are needed. Impairment (provision) occurs if it is probable that the counterparty will not be able to pay all amounts, and a loss will probably occur.

The literature [3] tells us that the seniority of the loan and the industry in which the firm is active are also very important components of the LGD. Another important risk driver of the LGD is the state of the economy. In times of recession the LGD might be higher than in times of economic prosperity. More interesting risk drivers of the LGD might be the leverage of the firm, which is $\frac{Total\ assets}{Total\ liabilities}$, the country in which the firm is incorporated, and the reason of default.

With those extra risk drivers I try to build a model, which predicts the LGD “better” than the current model. I try to build different kind of models, to see which model approximates the LGD the best. By building the models I try to find a balance between accuracy and transparency. On the one hand it would be desired to have a model that is very accurate. A disadvantage of such a model is that it might be very complex and not transparent, such that it is hard to implement in the business lines of the ABN. On the other hand a too transparent model might give no better results than the current model. So a good balance has to be found.

To build a new model, I make use of different techniques from the Statistics and Machine Learning. One technique that I use is a regression analysis. An advantage of this model is that it is pretty transparent, and may give accurate results, because this technique produces a function that outputs continuous values. A disadvantage of this model is that regression analysis assumes that there is a linear structure in the data, which might not be the case.

Another technique that is used is regression tree. The advantage of this technique is that it is does not assume any structure in the data. A disadvantage of regression tree is that it outputs classes, rather than continuous values. So the output might be less accurate. A last technique is a model tree. This is the same as the regression tree, but now the output classes contain a regression function. This technique is less transparent, but might be more accurate.

The thesis is organized as follows. In Chapter 2 the Basel accords are explained and the capital requirements arising from those accords are explained. In this chapter the beginning of the Basel accords is also discussed. Next in this chapter the important pillars of the Basel accords are discussed. Furthermore, the differences between Basel I, II, and III are given. Finally, the capital requirements for a bank, recording from the Basel accords, are described.

In the third chapter the default and the LGD are described in more detail. In this chapter the definition of default, the default process, the capital structure of a firm, and the anatomy of a Bankruptcy are described. Furthermore, in this chapter some common characteristics of the LGD, the distribution of the LGD and how to measure the LGD are described. In Chapter 4 the current and the new model are discussed. This chapter includes the techniques I used to build my model and includes the assumptions I made to build the model. In Chapter 5 the data is discussed. In this chapter the pre-processing steps I made are discussed, and other important information about the data is given. In the sixth chapter the results are given, and finally in the seventh chapter the conclusions are summarized.

2 Basel Accords

This chapter describes the history of the Basel Accords. The first section gives the history of the Basel Accords[4]. In the second section, Basel 1 is described[4][5][6][7][8][22]. In the third section, Basel II is described[7][9][14][22], and in the last section, Basel III is described [8][10] [11][12][22].

2.1 History of the Basel Accord

At the end of 1974 the central bank governors of the group of ten countries established the Basel Committee, due to the serious problems on the international currency and banking markets [4]. The main objective of the Committee was to improve the quality of supervision banking in the world.

The recommendations of the Basel Committee were not binding; neither did the Basel Committee have any supranational authority. The Committee only provided standards, recommendations, and guidelines. They hoped, but also expected that the individual countries would implement those in their own national system.

In 1988 the Basel Committee decided there was a need for a multinational accord to strength the stability of the international banking system, and to remove competitive inequality arising due to differences between national capital requirements. So the first Basel Capital Accord became a fact.

2.2 Basel I

The Basel Capital Accord of 1988, also known as Basel I, primarily focused on credit risk [5], which is the risk that an obligator is not able to pay back the loan. The objective was to ensure that the international banks had minimal capital behind to catch up losses occurring from credit risk. In the first paragraph the four pillars of Basel I are described. In the second paragraph the criticism on Basel I is described.

2.2.1 Pillars

The first pillar, known as the Constituents of Capital, divides capital into two types, namely Tier 1 and Tier 2 capital [7]. Tier 1 capital includes own equity and retained earnings after taxes, which are capital elements that all countries have in common. Besides, it is one of the best indicators showing how well a bank can catch up his losses occurring from credit risk [6][8]. Tier 2 capital consists of capital from elements, which are allowed as capital by the national supervisors, such as hidden reserves. So tier 2 capital is not the same in all countries, because the own national supervisors may decide what is allowed as Tier 2 capital.

According to the second pillar the credit risk was divided in 5 categories: 0%, 10%, 20%, 50%, and 100%. Each asset was assigned to one of these categories. The third pillar, also known as the 8% rule, says that the Cooke ratio or Capital Adequacy ratio, which is Capital/ Risk-weighted assets, should be bigger than or equal to 8%. The required capital that a bank should hold for each asset was 8% of the risk-weighted asset. Moreover, 8% of the risk-weighted asset should be covered with Tier 1 and Tier 2 capital [7]. Finally, 4% of the risk-weighted asset should be covered with Tier 1 capital.

So for an asset of 100 euros falling into the 100% category, 8 euros should be covered by both Tier 1 and Tier 2 capital, and at least 4 euro should be covered by Tier 1 capital.

The fourth pillar was about the implementation of Basel 1. In this pillar the governments of the countries were requested to ensure that the central banks of the countries create strong surveillance on following the Basel Accord.

2.2.3 Criticism on Basel I

Although Basel 1 was introduced in 1988, Basel 1 was really implemented in all the member countries at the end of 1992.

Quick after introducing Basel I a few disadvantages were found already. Disadvantages of this model were that Basel I only focused on credit risk and was targeted only on the G-10 countries. Credit risk was not the only important risk that should be considered. Other risks like market risk and operating risk were also important to take into account. Another criticism on Basel 1 was the pretty simplistic approach for setting the credit risk weights. Those risk weights were for the most part arbitrary, because they were not based on particular insolvency probability standards [5].

Finally this accord did not consider the obligator himself [6][8]. An unsecured loan which is borrowed to a triple A (best rating) client is weighted at the same way as a borrower with a D (lowest rating) status.

As a response to the banking crisis in the 1990 the Basel Committee suggested to replace the Capital Accord with a new more risk-sensitive Accord. In 2004 a new set of guidelines was introduced, named Basel II.

2.3 Basel II

In Basel II the pillar structure remains intact. Basel II is divided in three pillars. The first pillar is about the minimal capital requirements to cover credit risk, market risk, and operational risk [5]. The second pillar focus on supervisory review. In this pillar banks are advised to develop their own capital requirements to cover all their risks, and not only the risks of pillar 1. This capital is better known as Economical Capital. The objective of this pillar is to ensure that the bank always has enough capital. The third pillar focus on better reporting to the financial markets. The structure is summarized in Figure 1.

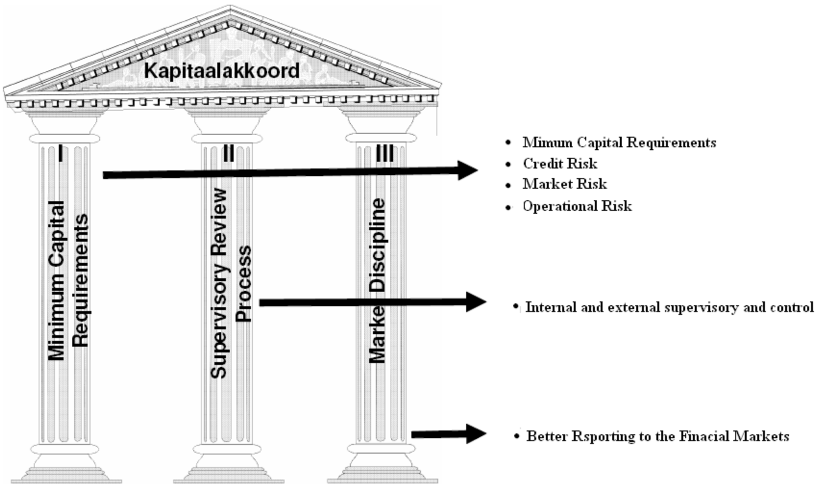

Kapitaalakkoord
I
Minimum Capital Requirements
II
Supervisory Review Process
III
Market Discipline
• Mimum Capital Requirements
• Credit Risk
• Market Risk
• Operational Risk
• Internal and external supervisory and control
• Better Rsporting to the Finacial Markets

Figure 1: Graphical representation of the Capital Accord [13]

2.3.1 Pillar 1

According to the first pillar there are three ways to rate the credit risk of bank assets. The first method is the so called Standardized Approach. This method is an extension of the method used in Basel 1, and includes external ratings. In this method the credit risk categories are the same, but now the category in which the asset falls depends on external ratings. The range of the external rating goes from AAA to D. Depending on which rating the asset has and depending on which kind of asset it is, it is assigned to a risk category [7]. This can be seen in Table 1.

Table 1: Risk weighted Categories under Standardized Approach

Rating	Sovereign debt	Banks debt	Corporates debt
AAA to AA-	0%	20%	20%
A+ to A-	20%	50%	50%
BBB+ to BBB-	50%	50%	100%
BB+ to BB-	100%	100%	100%
B+ to B-	100%	100%	150%
Below B-	100%	150%	150%
Unrated	100%	50%	100%

The other two methods are both known as the Internal Rating Based Approach (IRB). In this approach banks are allowed to use own internal ratings. By increasing the risk-weighted reserves by 6% if banks use the Standardized approach, banks are encouraged to use those own internal ratings. So if banks use internal ratings the bank is allowed to have fewer reserves, which is more profitable for the bank. Both IRB approaches are based on four parameters used to quantify credit risk for performing loans, which are loans that did not went in default (yet) [3].

1. Probability of Default (PD)
2. Exposure at Default (EAD)
3. Loss Given Default (LGD),
4. Maturity

Here the PD is the probability that the counterparty is not able to pay back the loan, because of a default. The EAD is the extent to which the bank is exposed if the counterparty will go in default. The LGD is the fraction of the EAD that cannot be recovered by a default. For a given maturity the expected loss can be expressed in the following formula: $EL = PD * EAD * LGD$.

In the first internal method, known as the Foundation IRB, only the PD may be assigned internally. The LGD and the EAD are both fixed. The LGD is based on supervisory values even as the EAD in cases the measurement is not clear. Finally, in this method the average maturity is assumed to be 2.5 years.

In the second internal method, which is known as the Advanced IRB, the LGD may also be assigned internally.

As already mentioned, according to Basel II banks have to hold enough capital to cover their operational risk and their market risk. Basel II contains three methods to handle operational risk. Operational risk is the risk, which is accompanied by failures in internal processes, in decision making of individuals, equipment, and other external events [7].

The first method is the Basic Indicator Approach. In this method the bank should hold 15% of their average gross-income earned the past three years behind to cover their operational risk.

The second method is the Standardized Approach. In this method the bank is divided in its business line to determine the capital that should be hold behind to cover their operational risk. Each of the business lines are weighted by its relative size to receive the percentage of the assets that should be hold behind. This can be seen in Table 2.

Table 2: Standardized Approach Reserve targets [7]

Business line	% of Profits needed in Reserves
Corporate Finance	18%
Sales & Trading	18%
Retail Banking	12%
Commercial Banking	15%
Settlement	18%
Agency Services	15%
Asset Management	12%
Retail Brokerage	12%

As can be seen; the more risky the business line, the higher the target.

The third method is the Advanced Measurement Approach. This method is a little more complex, and can be compared to the IRB approach of credit risk. In this method own calculations are made for the operational risk. Of course those calculations should be approved by the supervisors.

The last risk that has to be covered in Basel II is the market risk. Market risk is the risk of loss due to movement in the asset prices [7]. Within this risk, distinctions are made between fixed income and other products, like equity and commodities. Basel II also separates market risk into interest rate risk and volatility risk [7]. Interest rate risk is the risk that the value of the asset decreases due to fluctuations of the interest rate. Volatility risk is the risk that the value of the asset decreases as a result of changes in the volatility of a risk factor.

For the fixed income assets there is a measure, called the “Value at Risk” (VaR), to determine the capital that banks need to cover both risks. To determine this VaR, a bank can create own calculations, provided that they are approved by their supervisors.

If a bank is not able or does not want to develop their own VaR models, Basel II recommends two methods to cover the two kinds of risks. For the interest rate risk the amount of capital that one should hold depends on the maturity of the asset. This can be seen in Table 3.

Table 3: Interest Rate Risk weightings [7]

Time to Maturity	% of Profits needed in Reserves
1 month or less	0.00%
6 month or less	0.70%
1 year or less	1.25%
4 years or less	2.25%
8 years or less	3.75%
16 years or less	5.25%
20 years or less	7.50%
Over 20 years	12.50%

It is seen that the bank holds between the 0 and 12.50% of the value of the asset as capital to cover interest rate risk.

To cover the volatility risk of the fixed income assets, Basel II suggests risk weighting corresponding to the credit risk ratings, which can be seen in Table 4.

Table 4: Interest Rate Risk weightings for volatility risk [7]

Rating	Risk weigting
AAA to AA-	0.00%
A+ to BBB	0.25%
BB+ to B-	8.00%
Below B-	12.00%
Unrated	8.00%

The amount of capital needed to cover the fixed income assets against market risk is the sum of all the assets multiplied by both risk weightings.

For other market based products, like stocks and commodities, the risk weightings are based on other methodologies. Because there are many complex methods for these, only the three main methods are described.

The first method is called the Simplified Approach. This method works a little bit like the Non VaR models of the fixed income assets. The assets are divided by maturity, origin, and volatility. The weights are assigned along a range of values from 2.25% to 100%, depending on how risky the asset is.

The second method is called Scenario Analysis. This method is much more complex than the Simplified Approach. This method assigns the risk weights according to possible scenarios that the asset may encounter in each countries markets[7].

The final method is called the Internal Model Approach (IMA). In this model the bank may develop their own models to calculate the market risk for those other market based products.

The total amount of capital the bank should hold is now the sum of the capitals to cover credit risk, operational risk, and market risk. Besides, in Basel II the amount of Tier 1 capital should also be the same as the amount of Tier 2 capital.

2.3.2 Pillar 2

Pillar 2 is less complex than Pillar 1, and focuses mainly on the interaction between supervisor and the bank, and gives certain rights to the supervisor. According to this pillar the supervisor has the right to see the internal risk models of Pillar 1, and may change them to simpler and more conservative models if the bank is unable to manage the credit, operational, and market risks independently. The supervisor should also hold the senior management responsible if the bank misrepresents its risk position. Moreover, the supervisor has also the right to penalize the banks if they do not report their risk profiles[7].

Finally, Pillar 2 gives two mandates to the supervisor. The first mandate is to create an extra buffer of capital next to the minimum capital requirements, mentioned in Pillar 1, if the bank is "skirting" around the capital accords. The second mandate is to take immediately action when the capital requirements of a bank fall below the minimum levels. This is to avoid a financial crisis in countries.

2.3.3 Pillar 3

The last pillar focuses on better reporting. With Basel II the disclosures of the bank's capital and risk position should be published in public, instead of only visible for the supervisors which was the case in Basel I. So each quarter the banks have to publish their amount of Tier 1 and Tier 2 capital, their risk weighted ratios, their capital requirements to cover credit, operational, and market risk, and a description of which risk mitigation they undertake.

This way they hope that the shareholders force the bank to not take too much risk.

2.3.4 Capital requirements

Up till now the different rules for the capital requirements are discussed. A question that had not yet been answered completely is why a bank needs capital. As already said the bank needs to hold capital to cover his risks. The difficult thing about this is, is that a bank does not know its risk in advance. So a bank has to predict its risk, to have an idea what its losses could be. The losses of a bank can be divided in expected losses and unexpected losses. The expected loss is the average loss on a defaulted loan. In Figure 2 the density function of the loss distribution is given. The expected loss is thus the mean of the loss distribution.

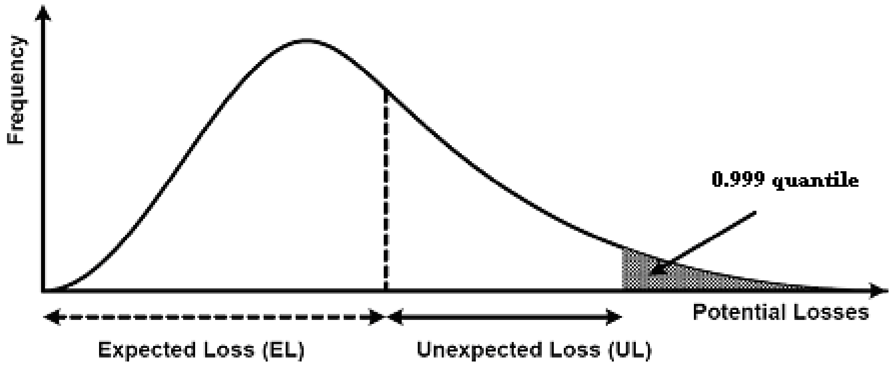


Figure 2: Probability density function of the Loss [14]

The expected loss is in majority covered with premiums, which are paid by the counterparty. For a bank it is not enough to cover only its expected losses. The bank should also cover its unexpected losses. The unexpected loss is defined as the 1-α quantile of the loss distribution minus the expected losses. In the Basel II accord this α is set to 0.001, which means that the total loss should be in 99.9% of the cases not higher than both EL and UL. The minimum amount of capital that should be held to cover both the EL and UL is defined as Regularity Capital, which is described in pillar 1 of the Basel II accord. Next to Regularity Capital there is also Economic Capital. Economic Capital is the amount of capital that the bank really holds to cover its risk. This is in most situations higher than the Regular capital. For their Economic Capital the ABN set their α to 0.0005, such that in 99.95% of the cases the total loss is covered with Economic Capital.

2.3.5 Critic

The main criticism on Basel II is that the accord only gives recommendations for the G10 members ,and not for developing countries. Although there are separate standards for the emerging markets, but due to the fact that these standards are obscure, they have less impact on the international banking. The large banks, rating agencies, and institutions see Basel II as the standard for the regulations of the world economies, and because of the fact that the standards for the emerging markets are unclear and less precise, they ignore the emerging market sectors.

Another criticism on Basel II is that because of too much trust in the rating agencies for valuing the risk, unfavourable implications are caused. This because small borrowers cannot afford the services of those agencies, which will cause banks to have less diversification on their loan books, which will make them more sensible for economic shocks. A second reason is that banks are allowed to choose their own rating agency, which may cause that banks search for the rating agency that gives the highest ratings, such that they need the least capital to cover their risk.

The last criticism on Basel II is that due to the new possibilities to develop own internal ratings, banks probably amplify recessions, and cause inflation in times of economic boom [7]. The reason for this is that due to own internal ratings, the risk weights are based on future expectation. This will ensure that banks will withdraw credit in recession times, and will extend more credit in times of economic prosperity. This method does ensure that the banks are well protected against additional risk, but it is generally known that the economic forecasters tend to exaggerate their predictions. So because the forecasters exaggerate the recession, banks tend to be too careful, which might only amplify the recession.

The current financial crisis also shows that the current capital requirements of Basel II are not sufficient. The last years showed that there is more risk, to which banks are exposed, which is not identified by Basel II.

For example, banks lower their capital by combining loans and sell them to so called Special Purpose Vehicles (SPVs). These SPVs, on their turn, issued shares and sold these shares. This way more credit risk came in the trading book, which are subjected to lower capital requirements [6]. Besides, the bank could use this money to issue new loans.

This and all the other criticisms made the Basel Committee decide to develop a new accord, called Basel III.

2.4 Basel III

Since 2009 the Basel committee works on this new accord, which will be released in June 2013. The idea is that Basel III will improve the capital framework.

Basel III consists of five main topics[22]. The first topic is to improve the quality of the capital. The second topic is to improve the risk coverage of the capital requirements. The third is the introduction of the leverage ratio. The fourth is reducing the procyclicality. The last topic is about how to tackle the system risk. The last paragraph discusses some remarks on Basel III.

2.4.1 Improving the quality of capital

The first topic of Basel III is about improving the quality of the capital. The financial crisis showed that only capital of the highest quality (Tier 1 capital) is useful to capture losses [10]. Therefore the bank should hold more Tier 1 capital. In Basel II 4% of the risk weighted assets should be covered with Tier 1 capital. In the new accord this increased to 6%. Furthermore, 4.5% of this 6% should come from common equity, which consist of common shares and retained income [11]. In the new accord the 8% rule stays intact, so still 8% of the risk-weighted assets should be covered by both Tier 1 and Tier 2 capital, but now at least 6% should be of Tier 1 capital.

However, the bank should also hold extra capital buffers to cover procyclical effects and system risk, which makes that the total capital requirements can increase to a maximum of 15.5% of the total risk weighted assets. The extra buffer requirements are explained in the paragraphs below.

2.4.2 Improving the risk coverage

The second topic of the accord is about improving the risk coverage of the capital requirements. When trading with options, futures, and swaps, there is a risk that the underlying value changes. In the new accord there is a new capital requirement for the risk to which a bank is exposed, when the creditworthiness of counterparty deteriorates.

Besides, in the new accord the banks are stimulated, with the help of lower capital requirements, to settle their derivatives as much as possible by a central counterparty. These parties ensure that the contracts are netted in a common way, which will lower the system risk.

2.4.3 Leverage ratio

The third topic in the accord is about the introduction of the leverage ratio. The leverage ratio is the ratio between Tier 1 capital and Total exposure. Total exposure is the gross sum of all active and off-balance sheet items. This ratio is introduced to prevent the bank building up a too large liability position. The new Accord proposes to let this leverage ratio be 3% at most.

2.4.4 Decreasing the procyclicallity

The fourth topic is about decreasing the procyclicality. As already said in the former section, one of the criticisms on Basel II was that due to the internal ratings, the banks amplify recessions and causes inflation in times of economic boom. To decrease the procyclicality, the new Basel accord suggests to build up acyclic capital requirements. The first measure is the so called capital conservation buffer. This buffer should only hold in times of economic boom, and consists of an extra 2.5% common equity Tier 1 capital buffer that should be held. During times of recession the bank does not have to hold this extra capital.

A separate extra capital requirement is the so called counter-cyclic buffer. This is a buffer between 0%-2.5% of the risk weighted assets that should be held to decrease the procyclicallity. The exact amount will be determined by the relevant supervisor in each jurisdiction, depending on the systemic risk a bank has built up as a result of too much credit growth [12].

2.4.5 Tackling the system risk

The last topic is how to tackle the system risk. Some banks are, because of their size, complexity, and their connections with other banks, so important for the financial system, that they are “too big to fail”. Often there is too much confidence within those banks (if something goes wrong the government will prevent that we fail), that they take more risk than other banks. To tackle the system risk, the banks should also hold an extra buffer between 0%-2.5%, called the SIFI buffer, to cover the system risk.

2.4.6 Remarks on Basel III

Before introducing Basel III some remarks should already be noticed. Because of the new rules in Basel III the capital requirements shall increase. If all the banks do this on the same time, this might lead to an increase of the prices, which will be paid by the customers and which will not be good for the economic growth of the world economy. A phased implementation can avert this[22].

Another remark on Basel III is about how the bank will get their extra capital. The banks should make more profit to be attractive for potential investors. On the other hand the politics want to reduce the risk of the banks, which will lead to low efficiency. The main question here is if those investors are still interested if the efficiency is low.

Finally the main important remark that should be mentioned is that we should not think that the new capital requirements results in a safe and stable financial system. All these extra measures only affect the consequences of the risk and not the reasons. There is still a lot of risk, but hopefully the banks are now (more) resistant to it.

3 Default and LGD

This chapter, based on [3] and [14], describes the default process and the LGD more in detail. The first section describes the definition of default and the default process. The second section describes the properties of the LGD and how it is measured.

3.1 Default an default process

In this section the definition of default and the process of the default are given. The first paragraph gives the definition of default. In the second paragraph the anatomy of the bankruptcy is given. In the third paragraph the capital structure of a firm is given.

3.1.1 Definition of default

Basel II gives the following definition of default [15]: *"A default is considered to have occurred with regard to a particular obligator when one or more of the following events has taken place.*

(a) It is determined that the obligator is unlikely to pay its debt obligations (principal, interest or fees) in full;

(b) A credit loss event associated with any obligation of the obligor, such as charge-off, specific provision, or distressed restructuring involving the forgiveness or postponement of principal, interest or fees;

(c) The obligor is past due more than 90 days on any credit obligation; or

(d) The obligor has filed for bankruptcy or similar protection from creditors."

The LGD depends on the definition of default. The broader the definition of default, the lower the average LGD will be and the other way around the narrower the definition, the higher the average LGD will be. For example it can be possible that a company that goes in default according to "c", pay all its obligations and has a recovery of 100%, so a LGD of 0%. If the bank does not count this as a default, the LGD will be overestimated.

3.1.2 Anatomy of the bankruptcy

In this section the anatomy of the bankruptcy is described. Although not all the defaults are the result of bankruptcy, as we saw in the former section, some of them are. So it is good to see how such a bankruptcy takes place. In Figure 3 a timeline is given of a firm that goes bankrupt.

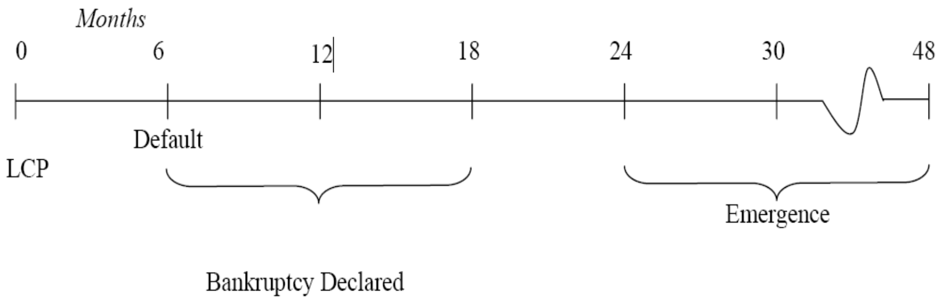


Figure 3: Timeline of firm that goes bankrupt [3]

A bankruptcy process always starts with the last cash paid (LCP). This of course is not known in advance, so this is determined later. At some moment later there is a default. For bonds this is the moment that a coupon or an interest payment is missing. Since coupons for bonds are paid twice a year this is six months later than the LCP. Then within 12 months the firm is often declared bankrupt. This can take a while, because a firm that goes in default does not necessarily have to be declared bankrupt. It is possible for the firm to negotiate with its creditors, to avoid a bankruptcy. The last part is the Emergence, which occurs between 2 to 4 years after the LCP.

It is possible that there is some cash flow during the period, but most of the cash is divided after Emergence.

3.1.3 The capital structure

In this paragraph the capital structure of a firm is described. The capital structure can be divided in eight dimensions, which can be seen in Figure 4.

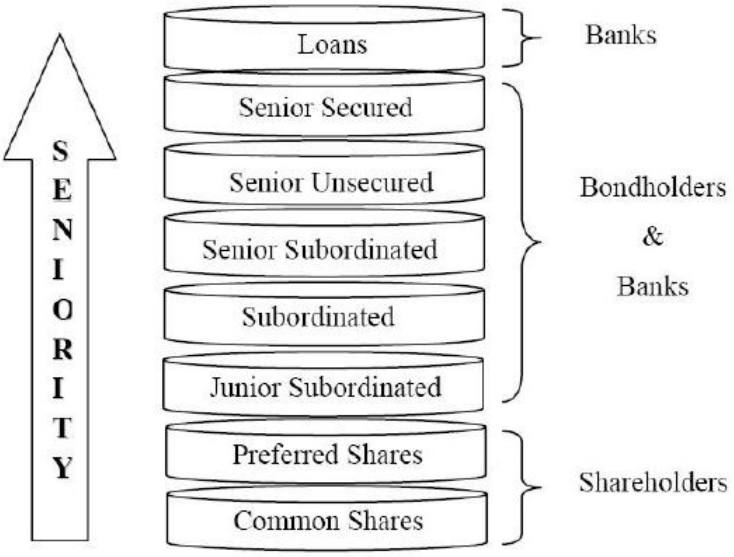


Figure 4: Capital structure of a firm [3]

The capital structure of a firm gives the order of the distribution of the value among the different debt types. The most senior creditors are paid first and most junior creditors are paid the last. This is done according to the absolute priority rule, which is a feature of the Bankruptcy Law. The APR says [25]:

"The absolute priority rule states that a bankruptcy firm's value is to be distributed to suppliers of capital such that senior creditors are fully satisfied before any distributions are made to more junior creditors, and junior creditors are paid in full before common shareholders."

However, this is the theoretical distribution. In practice the APR is in most cases violated. It has been shown that in 65%-80% of the cases the APR is violated. In these cases junior claimants already received money, while the senior claimants are not yet fully paid. The main reason for this is that most claimants agree to violate the APR if this speeds up the resolution.

3.2 LGD

In this section the LGD is described in more detail. The first paragraph describes some common facts about the losses and recovery. The second paragraph describes the distribution of the LGD. The third paragraph is about measuring the LGD.

3.2.1 Some common facts on losses and recovery

By analysing different academic studies on losses and recovery, some common characteristics has been found [3]. Note that the LGD is $1 - recovery$.

1) It has been found that recoveries are often either high (around 80%) or low (around 20%).

2) The most important risk driver for recoveries is whether or not it is secured and the seniority of the loan.

3) Recoveries are much lower in times of recession than in times of economic boom.

4) The industry of the firm is also an important risk driver. It has been shown that tangible asset industries have in most cases higher recoveries than service factor firms.

5) The size of the exposure does not have much influence on the losses.

3.2.2 Recovery distribution

If the recoveries are studied without looking to the different risk drivers, characteristic "1" of the former paragraph can be seen. In Figure 5 the probability distribution of the recoveries is displayed.

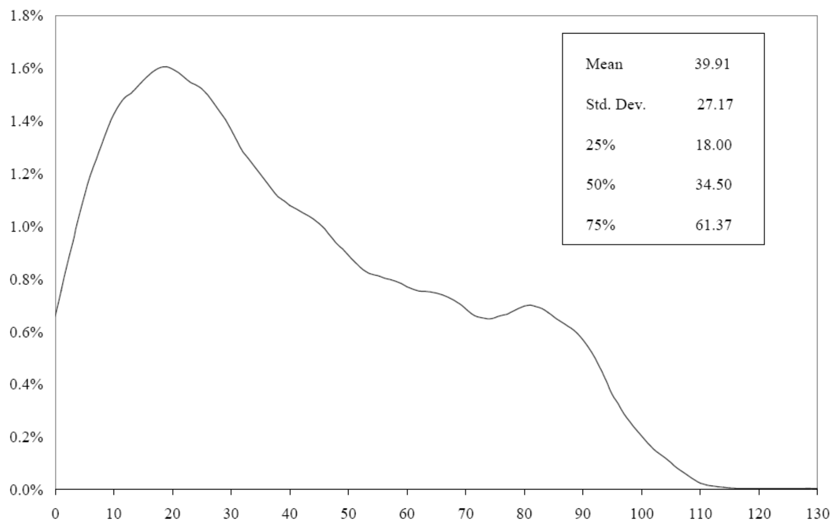


Figure 5: Probability distribution of the recoveries [3]

It is seen that there are indeed two peaks in the probability distribution of the recoveries. The first peak is around 20%, the other around 80%. That is why the literature says that distribution of the recoveries is called to be bimodal.

However, I do not really agree with this. There is indeed some kind of peak around 80%, but that peak is in my opinion too low to call the distribution bimodal.

The peaks can be explained by noticing that the recovery is probably high when the loan is secured and probably low when the loan is unsecured.

3.2.3 Measuring the LGD

The LGD is the part of the exposure at default that is lost. After a default, the LGD consist of the following losses[3]:

- Loss of principal
- Loss due to the carrying costs of non-performing loans
- Loss due to the cost of collecting the defaulted loan

There are three ways of measuring the LGD, depending on the instrument considered [16]:

- **Market LGD**. For traded assets, like bonds or securitized loans, the LGD is obtained from the market prices. As long as those defaulted bonds and loans are traded, the price can be observed directly. The prices of those loans and bonds at default namely reflect the expected recovery of the investor.

- **Workout LGD**. For non-traded bank loans, it is not possible to use the market LGD. Here the LGD is obtained from the discounted cash flows of the distressed loan. So all the discounted cash flows are taken into account from the period between the default and the resolution date. This period between default and the resolution date is also called the workout period. The value of the LGD is discounted to the value at default.

- **Implied Market LGD.** Here the LGD is obtained by looking at the credit spreads of the non-defaulted risky-free bonds, which are currently traded on the market. This can be done, because the spread above the risk-free bonds is a good indicator of the risk premium requested by the investors. However, what should be noticed is that this premium is, as already mentioned in the former chapter, a measure for the EL and not for the LGD. Fortunately, there are models which make it possible to separate this premium into the different parameters. Those models also found that the LGD obtained according to this method gives on average higher LGDs than the actual LGDs.

4 Modeling the LGD

This chapter describes both the current model that the ABN uses to predict the LGD, as well as the new models I used to predict the LGD. In the first section the current model, based on [17] is described more in detail. The second section the describes the new model techniques more in detail. The third section gives the quality measures for evaluating the models.

4.1 Current model

This section describes the current model that the ABN uses to predict the LGD. The current model is a rating system. The model produces rating classes, which correspond to a LGD percentage. The LGD classes are based on some explicit risk drivers:

- Segmentation
    - Corporates NL,
    - Corporates non-NL,
    - Private Individuals
        - Large (>= 1 mln) exposures
        - Small (< 1mln) exposures
- Collateral values
- Collateral types
- Guarantees
- Subordination
- Counterparty size
- Legal Risk (Score)
- UCR Factor
- Recovery rates
    - Collateral type
    - Recoveries related to the corresponding collateral

Those risk drivers are not based on any statistical or mathematical technique. They are chosen by experts on rational reasons.
In Table 5 the essence of the current model for performing loans is displayed.

Table 5: Essence of the current LGD model [17]

	Collateral coverage	LGD Class	LRS 1	LRS 2	LRS 3
Minimal Risk	>=100% cash	A	<LGD%>	<LGD%>	<LGD%>
Fully Mitigated	>=100% securities/cash	B	<LGD%>	<LGD%>	<LGD%>
Excess Coverage	>=150%	C	<LGD%>	<LGD%>	<LGD%>
Excess Coverage	>=120%, <150%	D	<LGD%>	<LGD%>	<LGD%>
Fully Covered	>=90%, <120%	E	<LGD%>	<LGD%>	<LGD%>
Partial Coverage	>=70%, <90%	F	<LGD%>	<LGD%>	<LGD%>
Partial Coverage	>=50%, <70%	G	<LGD%>	<LGD%>	<LGD%>
Partial Coverage	>=30%, <50%	H	<LGD%>	<LGD%>	<LGD%>
Partial Coverage	<30%	I	<LGD%>	<LGD%>	<LGD%>
Unsecured	(Large Corporates)	K	<LGD%>	<LGD%>	<LGD%>
Unsecured		L	<LGD%>	<LGD%>	<LGD%>
Subordinated	(Structural)	M	<LGD%>	<LGD%>	<LGD%>
Subordinated	(Contractual)	N	<LGD%>	<LGD%>	<LGD%>
Zero Recovery		Z	<LGD%>	<LGD%>	<LGD%>

It is seen that the current model assigns an LGD rating to each facility, ranging from A1 to Z3. The rating consist of a class (the letter) and a legal risk score (number). The class distinguishes the facilities from secured facilities (A to I) to unsecured facilities (K and L) to subordinated facilities (M to Z). Secured facilities are segmented in facilities which are covered by:

- financial collateral
  - "A" for fully cash collateralized
  - (typically) "B" for collateralization with securities and cash – but not fully cash collateralized
- Other collateral (classes C to I), or
- No collateral (unsecured classes K to Z)

The collateral coverage is calculated from the ratio between the net collateral value and the exposure at default (EAD). The net collateral value is the market value of the collaterals multiplied with the fixed recovery rates of the collaterals. The recovery rates of the different kind of collaterals are displayed in Table 6.

Table 6: Recovery rates Collaterals [17]

Collateral	Recovery rate (%)
Aircraft	50
Cash and deposits	100
Commodities	80
Corporate real estate	77
Gold accounts (daily monitoring)	97
Intangibles	10
Inventory / stock	35
Machinery and equipment (including 'Vehicles / Cars')	39
Marketable securities	64
Marketable securities (daily monitoring)	97
Non marketable securities	10
Other real estate	77
Other receivables	30
Other tangible fixed assets (other assets + sec.other)	30
Precious metals (not being commodities)	50
Residential real estate	80
Ships (BUN incl.Yacht)	56
Ships (ocean going)	55
Trade receivables	38

Another thing that should be mentioned is that the LGD here refers to the part of the exposure that is not guaranteed by a third party. This means that the total EAD is split into two parts:

- A part allocated to the counterparty, which is called the "Exposure to the counterparty" and is referred as Net EAD. So the LGD is here the LGD related to the exposure to the counterparty. This LGD is typically based on the ratio of the net collateral value related to the counterparty and the Net EAD.

- A second part is allocated to the guarantor, which is called the "Exposure to the guarantor" and is referred as guarantee amount. So the LGD here is related to the guarantee amount. This LGD is thus based on the ratio of the net collateral value related to the guarantee and the guarantee amount.

The consequence of this is that the guarantees are extracted from the exposure. So the LGD ratings are based on the non-guaranteed part of the facility. For the capital requirements, the guaranteed part of the facility is separately taken into account, using a substitution or double default method. These methods are out of the scope of this thesis and are therefore not discussed. This way of modeling gives a handle to the business to mitigate risk. Additional guarantees lower the risk, because additional guarantees decrease the exposure to the counterparty.

Next to the class label the rating also consist of a number. This is the legal risk score (LRS). The LRS make distinctions between facilities based on legal risk. It reflects the possibility to enforce the collateral rights.
To apply the LRS the Rule of Law indicator of the World Bank has been used, which can be seen in Table 7.

Table 7: LRS based on the Rule of law indicator [17]

Rule of Law indicator		Legal Risk Score
Scale Low	Scale High	
-2	0	**1**
0	1	**2**
1	**2**	**3**

The definition of the rule of law can be described as follows [18]:

"The extent to which agents have confidence in and abide by the rules of society, including the quality of contract enforcement and property rights, the police, and the courts, as well as the likelihood of crime and violence."

For the collateralized facilities the LRS is determined by the jurisdiction where the majority of the collateral is located. For non-collateralized facilities the LRS is determined by the country of residence of the borrower, assuming that the majority of the tangible assets are in the country of residence of the borrower. If the majority of the tangible assets is not in the country of residence of the borrower, then the country where the ABN AMRO has to exercise its rights over these assets is taken as country for the LRS.

To determine the different values of the LGD for the different LRS classes, there is an add on of 5% from going to LRS 1 to LRS 2, and also 5% added from LRS 2 to LRS 3. So for a given class if the average LGD is x, then for the facilities in LRS 1 the LGD is x, for LRS 2 x+5%, and for LRS 3 x+10%. These add-ons are based on experts' opinion and not on historical data, because of the lack of data available.

The last thing that should be mentioned here is that in the current model the ABN also makes a distinction between Corporates NL, Corporates Non-NL, Private Individuals (≤ 1 mln exposure), and Private Individuals (> 1 mln exposure). So all these different segmentations have an own LGD table with their own percentages per rating class.

The advantage of using this rating model is that it is intuitive and transparent for users, because all the values are mapped into discrete rating classes with a logical rank ordering. Besides, it gives handles to the users to mitigate risk. More collateral is a lower risk.

The model for the in-default LGD takes the LGD model for performing loans as a starting point and multiplies it with a UCR factor, depending on the UCR rating (the so called UCR factor method). For non-performing loans there are three possible UCRs, namely 6, 7 or 8. UCR 6 is the best rating of these three, followed by UCR 7 and UCR rating 8. Those rating are made by experts within the bank, who look at the financial information of the counterparty.

The current UCR factor method of the ABN multiplies the LGD for UCR 6 rated counterparties with 0.18 and multiplies the LGD for UCR 7 and 8 counterparties with 1.0.

4.2 New model

This section, mostly based on [19] and [20] [21], describes the different techniques that are used to build a new model more in detail. The first paragraph explains the linear regression technique. The second paragraph describes the regression tree technique. The third paragraph explains the model tree technique. In the last paragraph the decision technique is described. This technique is used in the results to predict whether a loss occurs on a loan or not.

A last thing that should be mentioned here is that none of the models will definitely find the global optimum. All the models will find a local optimum that might be the global optimum, but not necessarily is.

The reason for this is that there is no algorithm found up till now that runs in Polynomial time and guaranteed to find the global optimum. An algorithm runs in Polynomial time if the time, as function of the input, that is needed to come to a solution is restricted by a polynomial function.

4.2.1 Linear regression

Linear regression is a technique that is often used to predict a numeric class variable. The idea of this technique is to predict the class variable as a linear combination of other attributes. The output looks as follows:

$$y = w_0 + \ w_1x_1 + w_2x_2 + \cdots + w_nx_n$$

In this formula *y* is the class attribute, $x_1 \ldots x_n$ are the other attributes. $w_1 \ldots w_n$ are the weights that indicate how important the different attributes are for predicting the class attribute. Those weights are estimated from the training data.

For example, assume that the training data contains *m* instances. Now for each instance *j* of *m*, it is tried to predict class value $y^{(j)}$ with attribute values $x^{(j)}$and corresponding weights $w^{(j)}$. This can be written with the following formula $\sum_{i=1}^{n} w_i^{(j)} \ x_i^{(j)}$.

The idea of linear regression is to find those weights, such that the sum of the squared difference between the predicted and the actual value of $y^{(j)}$ is minimized. To write it in formula form [19]:

$$\sum_{j=1}^{m}\left(y^{(j)} - \sum_{i=1}^{n} w_i x_i^{(j)}\right)^2$$

So in the training step the weights are chosen in such a way that this equation is minimized.

Linear regression is a good, simple technique for numeric prediction. A disadvantage of this technique is that it assumes linearity. It assumes that the class variable can be written as a linear combination of the other attributes, which is not always the case. So linear regression tries to predict the distribution of the LGD with a linear function.

To implement this model, the linear regression function of WEKA is used. This function uses the M5 algorithm to select the attributes. This algorithm works quite the same as the pruning process of regression and model trees, which is explained in the next paragraph.

4.2.2 Regression tree

Another technique that can be used for numeric prediction is a regression tree. The idea of a regression tree is, as the name already says, to build a tree that predicts the class variable. The leafs of these trees are the predicted class values.

The first step of building a regression tree is to find a root node. The root node should contain the attribute that has the most “predictive power” with respect to the class variable. This root node is split into different branches, and for each branch again the attribute is chosen with the most “predictive power”. This continues until the class values of the instances that reach a node vary very slightly, that is when their standard deviation is less than 5% of the standard deviation of the original set.

Now it is important to know how to measure the “predictive power” of an attribute in each node. The attribute with the most “predictive power” is the attribute that maximizes the expected error reduction. The expected error reduction, also called standard deviation reduction (SDR), is defined as follows [19]:

$$SDR = sd(T) - \sum_i \frac{|T_i|}{|T|} * sd(T_i)$$

Here *T* is the set of instances that reach this node, and $T_1, T_2, \dots$ are the set of instances that are resulted from splitting the node according to the chosen attribute. So at each node the attribute is searched that maximizes this SDR.

In the second step the initial tree is pruned back from each node with the help of a pruning process. The reason for this pruning is that the tree might be too specific built on the training data; such that is does not perform well on the test data. By pruning the tree, the tree will become less complex, which will increase the error on the training data, but will decrease the error on the test data.

In the pruning process the expected error of the test data at each node is estimated. This is done by calculating the absolute difference between the predicted and the actual value of the class variable for all cases that reach a node that has to be pruned. This difference is averaged over the number of training instances that reach that node. Due to the fact that this tree is built specifically on the training data, this average will probably underestimate the expected error for the test set. To compensate for this, the averaged difference is multiplied with a factor $\frac{n+1}{n-1}$. Here *n* is the number of training instances that reach the node. Now a term is dropped if this decreases the error estimation. The procedure goes on as long as the expected error rate decreases.

In the last step of this technique, called smoothing, sharp discontinuities that exist between neighboring leaf nodes of the pruned tree are removed. The appropriate smoothing calculation is as follows [19]:

$$p' = \frac{np + kq}{n + k}$$

Here *p'* is the predicted value backed up to the next higher node, *p* is the predicted value at the current node. Moreover, in this formula *q* is the predicted value at the leaf node below the current node, *n* is the number of training instances that reach the current node, and finally *k* is a smoothing constant, often chosen to be 15 [23].

So regression tree cuts the distribution of the LGD in pieces, and tries to predict the LGD within those pieces with a constant value.

Also this model is implemented in WEKA. It is decided to set the minimum number of instances in a leaf node to 4.

4.2.3 Model Tree

The third technique discussed is the model tree. The model tree is almost the same as a regression tree. The only difference now is that a model tree has a regression model in his leaf node, rather than a value. For the rest the idea of a model tree is the same as regression tree. So first an initial tree is build, but now with a linear model in the leaf nodes. The linear model has the following form: $w_0 + w_1 a_1 + \cdots + w_n a_n$. Here $a_1, a_2, \ldots, a_n$ are again the attributes, and $w_0, w_1, \ldots, w_n$ are the weights.

The pruning process is also the same as with the regression tree. So at each node the expected error of the test data is estimated. The only difference is that the multiplication factor is in this case

$\frac{n+V}{n-V}$, where *n* is still the number of training instances that reach the node and *V* the number of parameters in the linear model that gives the class value at that node. Again a term is dropped if this decreases the error estimation. The procedure goes on as long as the expected error rate decreases.

The smoothing process is the same as in the regression tree. So sharp discontinuities that exist between neighboring leaf nodes of the pruned tree are removed.

The model tree cuts the distribution of the LGD in pieces, but now it tries to predict the LGD within those pieces with a linear function. So in fact model tree uses a piecewise linear function to predict the distribution of the LGD.

This model is also implemented in WEKA. The minimum number instances in a leaf are 4.

4.2.4 Decision Tree

The last technique that is used in this thesis is decision tree. Decision tree works quite the same as regression tree, only decision tree is used for classification rather than numerical prediction. Also in decision tree the first step of building the tree is to find the root node. Only now the root node is the attribute with the highest information gain. To calculate the information gain, a measure called Entropy is used. Entropy is a measure for the (im)purity of a set of training instances.

If S is a collection of training instances, and c the number of possible class values of S, than the entropy is defined as[21]:

$$Entropy(S) \equiv \sum_{i=1}^{c} -p_i \, log_2 p_i$$

Here $p_i$is the proportion of S belonging to class i.

Now with the entropy calculated it is possible to calculate the information gain of an attribute. The information gain of an attribute A and a collection of training instances S is defined as[21]:

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)}^{c} \frac{|S_v|}{|S|} Entropy(S_v)$$

Here Values(A) are all the possible values for attribute A. $S_v$is here the subset of S for which attribute A has value v. So Gain(S,A) is a measure of the expected reduction in entropy by using attribute A as attribute to split. This goes on till all the instance are correctly classified or all the variables are split.

Again this tree will work quite well on the training data, but might be too specific for the test data, which will result in overfitting. Therefore, this tree is pruned back at each node according to a pruning process.

The idea of the pruning process is to look to the set of instances that reach a certain node. Now assume that the majority class value is chosen to represent that node. This results in a certain amount of errors E of the total number of instances N. Now assume that the true error rate at that node is q, Moreover, assume that the N number of instances is generated according to a Bernoulli process with parameter q. Finally, assume that E of these N are errors[19].

Now it is interesting to get a good estimation of that true error rate q. If we denote with $f = \frac{E}{N}$ the observed error rate, based on the training data, than f is a random variable with mean q and variance $\frac{q(1-q)}{N}$. It can be shown mathematically that for large N the distribution of f approaches a normal distribution.

For a standard normal distributed random variable X a confidence limit z, for a particular confidence level c, can be found with $P(X \geq z) = c$, which can be solved with the help of a standard normal table.

Now if the mean q is subtracted from f and we divide this result by the standard deviation $\sqrt{\frac{q(1-q)}{N}}$, we get a standard normal random variable. For this variable it is again possible to get a confidence limit z, for a particular confidence level c (which is set to 0.25 in the algorithm that is used in this thesis), with $P\left(\frac{f-q}{\sqrt{q(1-q)}} \geq z\right) = c$. With this equation an upper confidence limit for q can be found. This limit is used as an estimate for the error rate at a node[19]:

$$e = \frac{f + \frac{z^2}{2N} + z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}}}{1 + \frac{z^2}{N}}$$

Now at each leaf node this error rate e is estimated. Furthermore, the error rate of a parent node is calculated, by calculating the combined error rate of his children. Finally the error rate of that same parent node is also estimated with the formula above. If this estimated error rate is lower than the combined error rate of his children, the children are pruned away. This process continues till there are no children left that can be pruned away.

This model is also implemented in WEKA with the J48 algorithm. In this function the minimum number of instances in a leaf node is set 2. Furthermore as already mentioned the confidence level c is set to 0.25.

4.3 Quality measures

To measure the quality of models, the dataset is first split in both training and test set. The model is built on the training data, the quality of the model is measured on the test set. The data is split into training and test with the help of a 10-fold cross validation. With a 10 fold cross-validation the total dataset is split into 10 equal size parts. Now for each model technique, 10 models are trained on each combination of 9 out of 10 parts, and is tested on the remaining part. Finally the error rate of all the ten models is averaged to get an estimate of the overall error rate of the build model.

For the error rate both the Root Mean Squared Error (RMSE) as the Mean Absolute Error (MAE) are used.

The RMSE is calculated as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)}$$

Here $\hat{Y}_i$ is the predicted class value of instance i and Y_i the actual class value of the test set of instance i.

The MAE is calculated with the following formula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{Y}_i - Y_i\right|$$

Again $\hat{Y}_i$ is the predicted class value of instance i and Y_i the actual class value of the test set of instance i.

5 Data

In this chapter the data and the preprocessing steps are discussed. In Appendix A all the variables of the data are given and explained shortly.

The original LGD data consist of 7244 instances and 52 variables. As already said the data consist of non-performing loans of the non-retail branch of the ABN AMRO. The oldest loan went in default in 1986, and the newest loan went in default on 15 February 2012. Moreover, 4819 of the 7244 instances are resolved loans, and 2425 of the 7244 instances are non-resolved loans. Resolved loans are defaulted loans, for which the recovery period is finished. For these loans the real loss and the real EAD are known, and so the real LGD is known. The non-resolved loans are also defaulted loans, but for those loans the recovery process is not yet finished. So for the non-resolved loans, the real loss and the real EAD, and so the real LGD are not yet known. The dataset does contain LGD values for those non-resolved loans, but those values are predicted.

Due to the fact that the LGD for the non-resolved loans is predicted, it is decided to delete all the non-resolved instances, and build the model only on resolved instances. A disadvantage of this is that the data now does not contain the most recent loans.

Next, it is interesting to see whether there are anomalies and missing values in the dataset. There are different kind of methods to handle with anomalies and missing values. One possibility is to change those anomalous or missing values to zero. Another method is to calculate the average of all the other instances, and change the anomaly or missing value to that average value. A last method is to delete the whole instance. It depends on the situation which method is the best to use.

The first anomalies are found in "Period". In this variable some negative values are found, which means that the resolve date was before the default date, which is of course not possible. It is decided to delete those instances, because I did not trust them. The reason for this is that either the resolved date or the default date is wrong. Due to the fact that these dates are both used to calculate the Net Present Value of the losses, the calculated LGD might be wrong. Next, there are also negative values found in the variable "Outstanding year", which should not be possible. Those instances are also deleted, because it does not make any sense to set them to zero or to the average value of the other instances. Besides changing the values might only spoils the data, and there is no harm by deleting these small numbers of instances. Furthermore, there are some missing values found in the variables "NET Collateral Final", "Cost Incurred", "LGD_gr", and "Provision". All those missing values are set to 0, because the experts say it is quite assumable that these values should be zero. Besides if we would delete all those instances, too many data is deleted, which is not desirable.

There are also some quite high values found in the write off percentage of the OAD and the coverage ratio, but those seem to be right.

Another interesting thing to see is that in 80,8% of the cases the realized LGD is 0.5%. This means, that in 80,8% of the cases the default is totally recovered, but the losses contain indirect costs. This can be seen well in Figure 6.

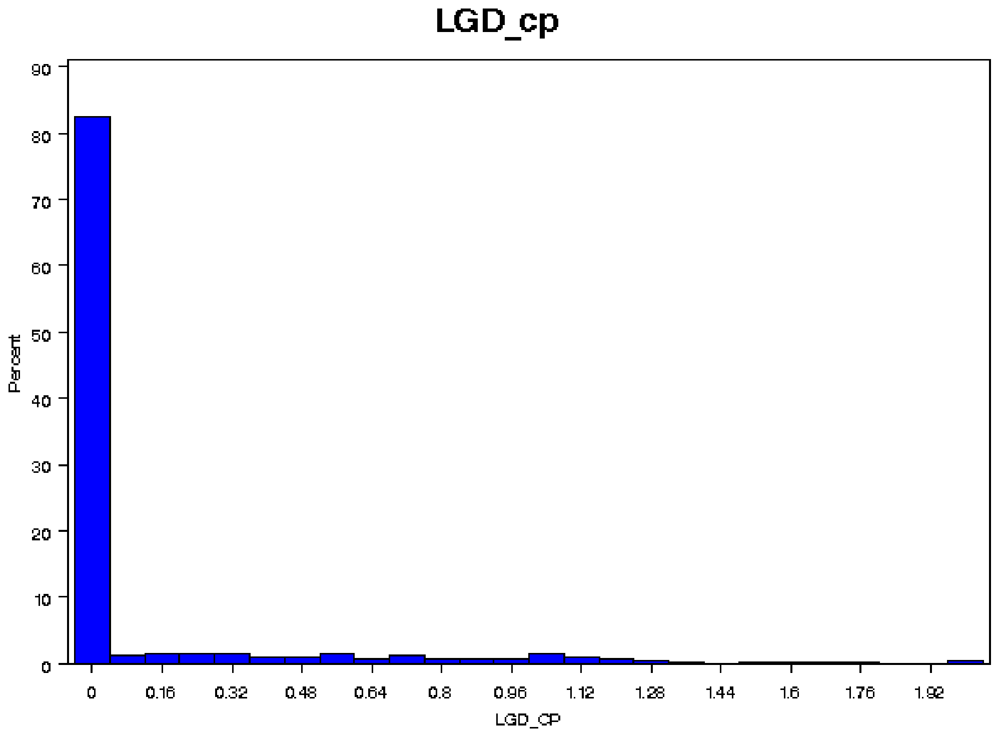


Figure 6: Probability density function of the LGD

It is seen that in our data the distribution of the LGD is not bimodal. However if the LGD is capped between 0 and 1 the distribution looks more bimodal. This can be seen in Figure 7

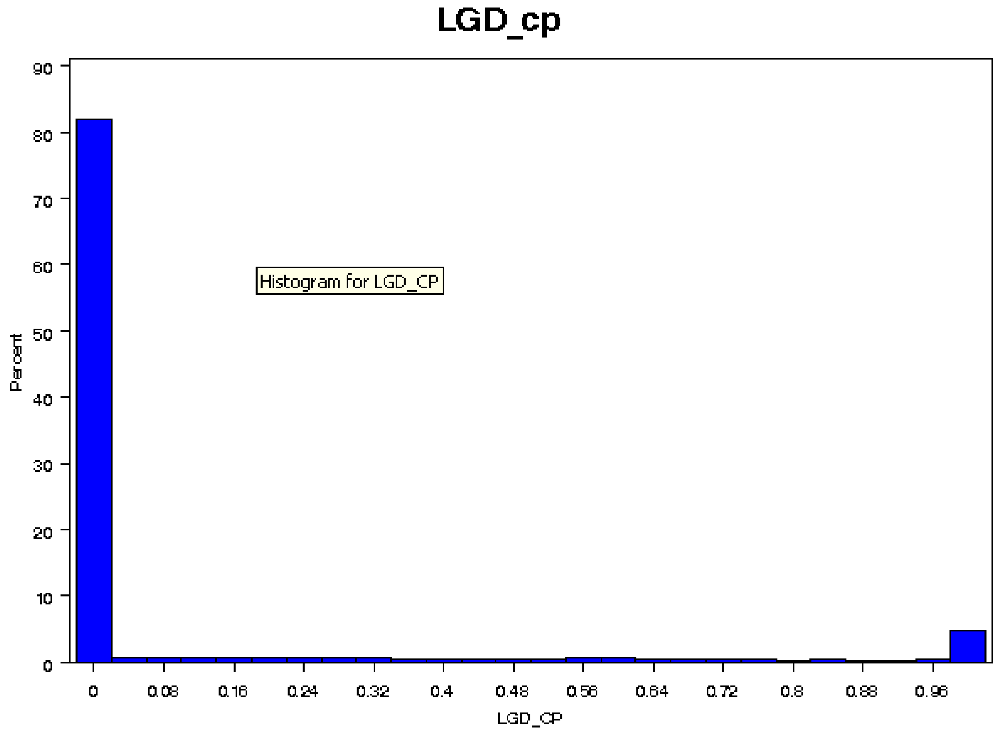


Figure 7: Probability density function of the capped LGD

An explanation for the many totally recovered loans is that the ABN is quite strict in assigning the label “default” to loans. If the bank has only slight the feeling that a company might not fulfil his credit obligations they will label the company loans as “default”. Another explanation is that the ABN helps the defaulted companies in their restructuring process, such that they have a bigger chance to recover. These reasons might explain the relative high number of totally recovered loans in the dataset.

Furthermore, it is decided to replace the AGIC codes by the industry they belong to, because the variable AGIC code has too many different values. By clustering the AGIC codes to 12 different industries, the variable becomes probably more interesting.

The customized LGD data consists of 4791 instances and 52 variables. Now this data is combined with the Probability of Default rating data. This is done to investigate whether there is a correlation between the LGD and PD.

It is only possible to get rating information for only 518 of the 4791 instances of the LGD dataset. The reason for this is that it was hard to combine the LGD dataset with the PD dataset. So for 4273 instances it was not possible to get rating information. It is decided to delete those instances, because it does not make sense to set those missing values to zero or to the average value of the other instances.

Furthermore the variable UCR_after_default is added to the dataset. This variable also contains some missing values. To find a suitable value for these missing values, it is decided to look to the amount of provision that has been taken. The reason for this is that the provision amount is a very good indicator for the expected losses. So if the provision amount is 0, it is expected that no loss will occur.

So if the provision amount is 0, it is decided to set the UCR to 6, otherwise it is set to 7. However all the instances that have a missing UCR have a provision amount of 0, so all the missing UCR values are set to 6. Furthermore there are some UCR below 6. This is strange, because all defaulted loans should have UCR 6,7 or 8. It is decided to set all those values to 6, because in general it holds the lower the UCR the better. So it is quite logic to assign the lowest possible UCR to those defaulted loans, which is 6 in this case.

It is now important to see when the remaining variables are known. This can be seen in Table 8.

Table 8: When are variables known

Variables	Pre-default	Default date	Post default	Resolution
CP_type	x	x	x	x
Ind_country	x	x	x	x
Ind_country2	x	x	x	x
Segmentx	x	x	x	x
Agic_oid	x	x	x	x
CP_id	x	x	x	x
Cst_incr				x
Dt_Resolv				x
Eff_dt_of_dflt		x	x	x
Fac_id	x	x	x	x

Facility Limit	x	x	x	x
Max_wrt_off_dt				x
NM	x	x	x	x
Outstanding_default		x	x	x
Out_quarter		x	x	x
Prvsn		x	x	x
Prvsn_dt		x	x	x
Segment	x	x	x	x
Wrt_off				x
Loss_dt				x
Gram	x	x	x	x
Grpc				x
Outstanding_year	x	x	x	x
Expwrt				x
Expout				x
Resolved	x	x	x	x
Loss				x
Recovery				x
Interest_rate				x
Period		x	x	x
RecoveredPart				x
NonRecoveredPart				x
RecoveredPartDiscounted				x
NonRecoveredPartDiscounted				x
DirectCost_NPV				x
DiscountedLoss				x
Wrt_off_oad				x
Lgd_gr				x
Lgd_cp				x
Num_cp				x
Den_cp				x
Coverage Ratio	x	x	x	x
Country_residence	x	x	x	x
Counterparty_type	x	x	x	x
BU	x	x	x	x
UCR_1y_prior	x	x	x	x
Net_Collatateral_final	x	x	x	x
EAD	x	x	x	x
Ind_ooe	x	x	x	x
Tear	x	x	x	x
Quick ratio		x	x	x
OAAM/TA		x	x	x
RE/TA		x	x	x
TD/EDIBTA		x	x	x
Current Ratio		x	x	x
Net operating Margin		x	x	x
Accounts Payable/COGS		x	x	x
UCR_after_default		x	x	x
Collateral Type	x	x	x	x

It is seen that some variables are only known at the resolved date. Those variables should not be taken into account if we want to estimate the final LGD during the recovery period, which is before the resolve date. This means that those variables can be deleted.

Moreover, the data contains some variables that have the same value as other variables. Besides, there are also variables that do not contain any value. It is not interesting to build a model on same variables or on variables without values, so these variables are also deleted.

The final dataset has 27 variables and 518 instances.

Finally it is decided to change all nominal variables to binary. So for each possible nominal value a binary attribute is made, containing a 1 when an instance contains that value, and a 0 otherwise.

6 Results

This chapter gives the results. The first section describes the results of predicting the in-default LGD. The second section gives the results for predicting whether a loss occurs or not. In the third section I tried to predict the total in-default LGD, which is the LGD of the counterparty and the LGD of the guarantor combined. In the fourth section the in-default LGD is predicted with the cashflow data. In the fifth section I tried to find a new model for the LGD of performing loans.

6.1 Predicting the in-default LGD

The first paragraph describes the results of the univariate analysis. The second paragraph gives the results of the multivariate analysis for the small dataset with the PD variables. The third paragraph shows the results for the multivariate analysis for the large dataset.

6.1.1 Univariate analysis

This paragraph shows only the most interesting results of the univariate analysis. The first interesting variable considered is the provision. In Table 9 it is seen whether provision influence the LGD.

Table 9: Influence of provision on LGD

Provision	N	Mean	Std.
0	503	1%	65%
>0	15	15%	29%

It is seen that defaults with non-zero provision have on average a higher LGD than defaults with zero provision. This is quite logic, because as already mentioned the provision amount is an indicator for the expected losses. So it looks likes that provision influences the LGD.

The second variable considered is the variable industry, which can be seen in Table 10.

Table 10: LGD per industry

Industry	N	Mean	Std.
Retail	90	25%	13%
Technology	10	1%	0%
Travel & Leisure	42	1%	4%

It is seen that that the retail industry has on average a higher LGD than the other two industries, so industry might be a risk driver for the LGD.
The last variable considered is period. This variable can be seen in Table 11.

Table 11: Influence of period

Period	N	Mean	Std.
≤ 0,5	219	1%	0%
> 0,5 and ≤ 1	161	1%	8%
> 1 and ≤ 1,5	68	2%	12%
> 1,5 and ≤ 2	39	2%	6%
> 2	31	6%	21%

This result is less clear than the other cases, but it can be seen that a default with a period longer than 2 years has on average a higher LGD than defaults with a period smaller than 2 years. This is not strange, because you can imagine that the longer the loan is in default the less likely it is that the loan will recover.

6.1.2 Multivariate analysis with small dataset

This paragraph describes the results of the multivariate analysis. In Table 12 gives the result of the current model. As discussed in Chapter 4 the current model is based on the coverage ratio and the UCR factor. In this model there are fixed classes, all containing a fixed LGD. The LGD of these classes are compared to the actual LGD in the dataset to get the RMSE and MAE.

Table 12: Results current model

Model	RMSE	MAE
Current Model	0.0813	0.0278

Now it is investigated if it is possible to get a better model with lower RMSE and MAE. First the linear regression technique is used to find a better model. For the linear regression model the output is given in equation (6.1.1):

$$LGD = 0.0154 * Period + 0.0019 \tag{6.1.1}$$

So according to the linear regression technique the LGD can be written as a linear combination of period. It can be seen that the higher the period, the higher the LGD. This quite logic, as explained in the former paragraph.
Furthermore it can be seen that industry and provision does not have that much influence on the LGD as one would expect after the univariate analysis.

The second model considered is the regression tree model. The output of this model can be found in equation (6.1.2).

$$LGD = 0.0144 \tag{6.1.2}$$

According to the regression tree model the LGD should be 0.0144 for all the defaults. So not a single variable is included in this model. This is quite striking, because we would at least expect, looking to the univariate analysis, that provision and industry would be involved. The reason for the flat LGD is probably the high percentage (97.30%) of instances that has no loss. The regression tree technique is probably not able to come up with a good model in such cases. Maybe other (more complex) techniques that are not used in this thesis will be able to find a better model with risk drivers.

The last model considered is the model tree. In (6.1.3) thee output of the model tree is given.

$$LGD = 0.0111 * Period + 0.0002 * Provision + 0.002 \tag{6.1.3}$$

According to the model tree the LGD can be written as a linear combination of period and provision. Again it can be seen that a higher period and a higher provision leads to a higher LGD.

Furthermore it can be seen that the variable industry is again not involved in this model. An explanation for this can be that if provision and/or period are already in the model, adding the variable industry does not give more information. So that could explain why industry is not in the model, even though it does influences the LGD, as we saw in the univariate analysis.

Another thing that can be seen is that also in this model no PD variable is involved.

In Table 13 both the mean as the standard deviation of the RMSE and the MAE over the 10 fold cross validation are given for the three considered models. This way it can be seen which model is the best to use in this case.

Table 13: Results new model

Model	RMSE	Std. RMSE	MAE	Std. MAE
Linear Regression	0.0830	0.0559	0.0186	0.0084
Regression Tree	0.0844	0.0562	0.0185	0.0077
Model Tree	0.0804	0.0564	0.0168	0.0090

It is seen that the model tree has the lowest RMSE and MAE. However to ensure that the difference is significant a test should be applied. With the help of the correct paired t-test of WEKA we can check if this difference is significant. The significance level is set to 5%. Appendix B describes how the test exactly works.
The results in Weka shows that is not possible to reject H_0, so I cannot say that the difference is significant.

It is striking that linear regression does not give here the same linear model as the model tree. Both models are now just a linear combination of variables, so it would be expected that those models would output the same in this case, which is not the case. It is checked whether adding the variable provision to the linear regression model improves the model, but that was indeed not the case. Adding the variable provision to the linear regression model increases the RMSE and MAE. Unfortunately, it is not possible to come up with a good explanation for this.

6.1.3 Multivariate analysis with large dataset

In this paragraph another multivariate analysis is performed, but now on the whole LGD dataset of 4791 instances. This dataset does contain the PD variables, but these variables have only a value for 518 of the 4791 instances. For all the other instances, these values are just missing in the dataset. Again the first model considered is the linear regression model. In Table 14 one can see which variables, with corresponding weight, are involved in this model.

Table 14: Involved variables with weights

Variable	Weight
$CP\_Type_{Corporates}$	-0.0568
$Period$	$+0.0658$
$IND\_OOE_{\geq 1mln}$	-0.0477
$Quick\ Ratio$	$+0.0059$
$Current\ Ratio$	$+0.0058$
$Industry_{Consumer\ Goods}$	-0.0474
$Industry_{Industrials}$	-0.0240
$Industry_{Technology}$	$+0.1029$
$Industry_{Oil\&Gas}$	$+0.6148$
$Industry_{Telecommunications}$	$+0.3866$
$Provision$	$+0.0001$
Constant	$+0.0536$

The LGD can now be written as a linear combination of certain variables. It can be seen that according to this model Corporates have on average lower LGD than Private Individuals. Moreover, it can be seen that the industries Consumer Goods and Industrials has on average lower LGD than the industries Oil&Gas and Technology. Finally, it can be seen again that also according to this model a higher period and provision leads to a higher LGD.

The second model considered is again the regression tree. The tree build with the regression tree can be seen in Figure 9.

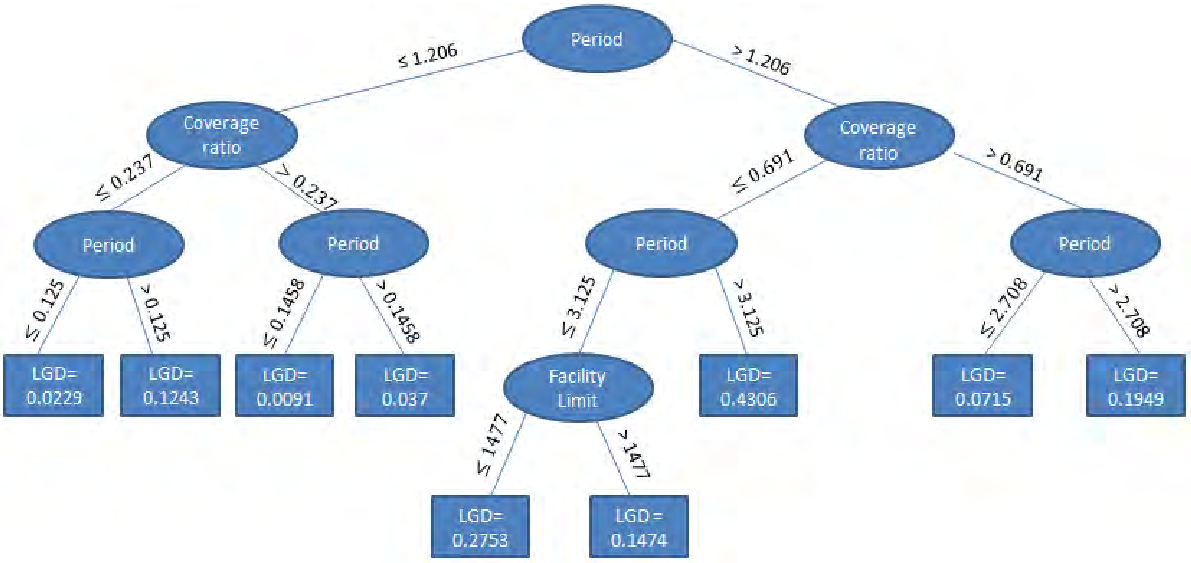
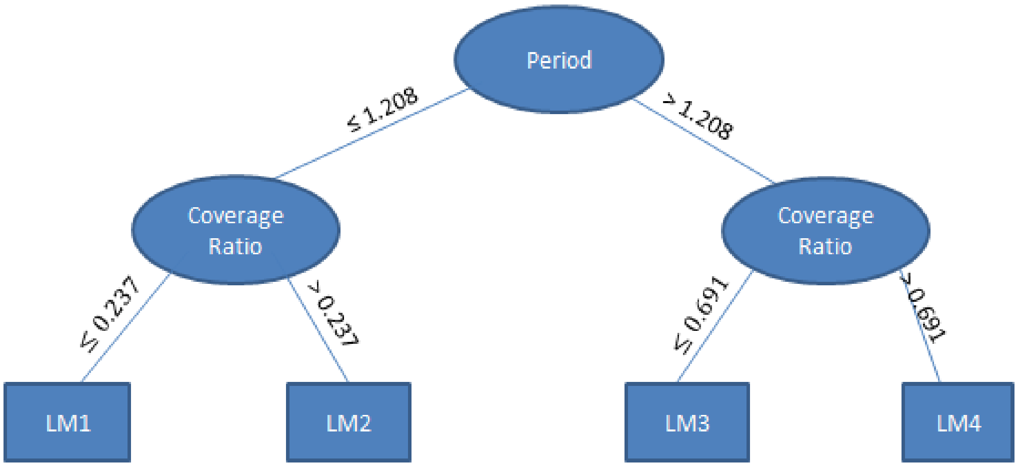


Figure 9: Output regression tree

It can be seen that the regression tree is now a little more complex than in the former paragraph. Instead of a constant LGD, the LGD now depends on three variables. It is seen that according to this model a higher period leads in general to a higher LGD, which was also seen and explained in the univariate analysis in the beginning of this chapter. Moreover, it can be seen that a higher coverage ratio does not lead to a lower LGD in all the cases. A reason for this might be that period influence the LGD more than coverage ratio. So a defaulted loan with a high coverage ratio, but a long period, has on average a higher LGD than defaulted loans with a lower coverage ratio, but also a shorter period.

The last model considered is the model tree. The output of the model tree can be found in Figure 10.



$$LM1:\ LGD = 0.1295 * Period + 0.0367$$
$$LM2:\ LGD = 0.046 * Period - 0.0006$$
$$LM3:\ LGD = 0.0528 * Period - 0.1229 * Coverage\ Ratio + 0.1837$$
$$LM4:\ LGD = 0.0566 * Period - 0.0284$$

Figure 10: Output model tree

It can be seen that almost the same variables are involved in this model as in the model of the regression tree. Only facility limit is not a variable in this model. It is hard to draw any conclusion from model trees, since the variables that are involved are both in a tree as in a linear model, which makes it hard to discover a correlation between the different variables and the LGD.

In Table 15 the RMSE and the MAE are given for both the three considered models as well as the current method. This way it can be seen which model predicts the LGD the best in this case.

Table 15: Results for all the models

Model	RMSE	Std. RMSE	MAE	Std. MAE
Current Model	0.2242	-	0.1688	-
Linear Regression	0.2964	0.0214	0.1711	0.0090
Regression Tree	0.2962	0.0191	0.1675	0.0081
Model Tree	0.2947	0.0204	0.1663	0.0095

It is seen that model tree has again the best result of the new models. However, to know if this difference is significant a correct paired t-test should be applied. The result in Weka shows that it is again not possible to reject H_0, so I cannot say that the difference is significant.

Another thing that can be seen is that the current model is no worse than the new models, so the current model is not that bad at all.

6.2 Predicting the loss/no loss

In this section it is investigated whether it is possible to predict whether there will be a loss on a defaulted loan or not. To do this, a new variable was introduced, indicating whether there was a loss on a loan or not. With the help of the decision tree algorithm it is tried to predict this new variable.

First I tried it with the small dataset. The decision tree algorithm outputs a 0, which means that it is optimal to predict no loss for all the defaults. This results in 97,30% of the instances correctly classified. The accuracy of this model directly gives an explanation why the algorithm outputs always a 0. 97.30% of the instances in this small dataset have no loss, so 504 of the 518 instances have no loss. So it is not quite strange that the model outputs a 0 for all the instances.

Now it is tried with the large dataset. The algorithm now outputs a very large tree with 76 leaves, resulting in 81,73% of the instances correctly classified. If it is noticed that the number instances that does not have a loss is 80,82%, it can be say that the prediction of the algorithm is not very impressive. Due to the fact that the complex tree is hard to implement in the business, it is decided to simplify the model. This results in a tree that outputs always 0, with an accuracy of 80,82%.

Now a model for the Loss Given No Loss (LGNL) is made. No advanced techniques are needed to find a model for the LGNL, since it is always 0.005. So (6.2.1) gives the model of the LGNL, which has a RMSE and MAE of 0.

$$LGNL = 0.005 \tag{6.2.1}$$

Now it is interesting to calculate the overall error rate of this model. Since all the instances are classified as 0, it is easy to calculate the overall error rate. This can be calculated by assigning 0.005 to all the instances, and calculate the RMSE and MAE.

This results in a RMSE of 0.3381 and MAE of 0.1448. So unfortunately, predicting the LGD this way does not give better results than the current model.

6.3 Predicting total in-default LGD

In this section I tried to find a model for predicting the total LGD, which is the LGD of the counterparty and the LGD of the guarantor combined. Again first I tried to do it with the small dataset.

For the small dataset all the three models outputs the formula of (6.3.1).

$$LGD = 0.015 \tag{6.3.1}$$

So all the models outputs a constant LGD, which means that not a single variable is involved. The reason for this is probably the high percentage of instances in data that have no loss.

In Table 16 the error of the different models can be found.

Table 16: Results for all the models

Model	RMSE	Std. RMSE	MAE	Std. MAE
Linear Regression	0.0834	0.0565	0.0196	0.0079
Regression Tree	0.0834	0.0565	0.0196	0.0079
Model Tree	0.0834	0.0565	0.0196	0.0079

It is seen that for all the models the RMSE and the MAE are the same. This is quite logic, because all the models were the same.

Now it is tried with the large dataset. Again first the output of the linear regression method is given in Table 17.

Table 17: Involved variables with weights

Variable	Weight
$Period$	$+0.0596$
$IND\_OOE_{\geq 1mln}$	-0.0669
$Industry_{Technology}$	$+0.0524$
$Industry_{Oil\&Gas}$	$+0.5953$
$Industry_{Telecommunications}$	$+0.3424$
$Provision$	$+0.0001$
$Collateral\ Type_{Real\ Estate\ other\ corporates}$	-0.0835
$Colleteral\ Type_{Real\ Estate\ other\ Residential}$	-0.0245
$Collateral\ Type_{Real\ Estate\ Commercial}$	$+0.0536$
$Collateral\ Type_{Aircraft}$	-0.1876
$Collateral\ Type_{Other\ tangible\ fixed\ assets}$	$\mathbf{+0.5302}$
$Collateral\ Type_{Diamonds\ Jewellery}$	$\mathbf{+0.3838}$
$UCR\ After\ Default_{7}$	$+0.1027$

$UCR\ After\ Default_8$	$+0.1027$
Constant	$+0.0470$

It can be seen that there are more variables involved in this model than in the model of the small dataset. This has probably to do with the fact that in this dataset the percentage of instances that has no loss is smaller than in the small dataset.

Moreover, it can be seen again that a higher period and a higher provision leads to a higher LGD. A striking thing that can also be seen is that Collateral Types “Other Tangible fixed assets” and “Diamonds Jewellery” has positive weights. This means that having those Collateral types leads to a higher LGD. This is strange as we would expect that more collateral leads to a lower LGD. An explanation for this might be that it is better to have other collateral types as coverage than those collateral types. A last thing that can be seen is that counterparties with UCR 7 or 8 have on average higher LGDs than counterparties with UCR 6, which is exactly what should be the case.

The output of the Regression Tree is given in Figure 11

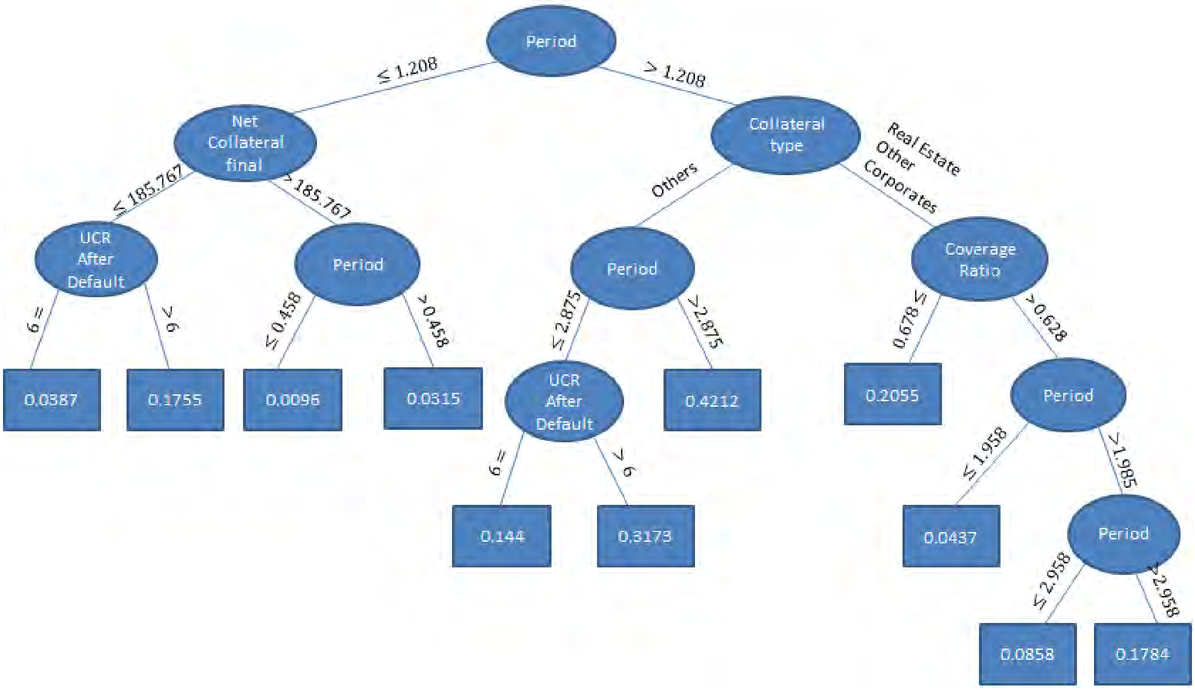


Figure 11: Output **r**egression tree

It is seen that with the regression tree technique also more variables are involved in this model than in the model of the small dataset. Moreover, it is seen again that a higher period leads in general to a higher LGD. Finally, it is also seen that UCR 6 has on average a lower LGD than UCR 7 or 8, which also quite expectable.

The output of the model tree can be found in Table 18.

Table 18: Involved variables with weights

Variable	Weight
$Period$	$+0.0528$
$IND\_OOE_{\geq 1mln}$	-0.0140
$Provision$	$+0.0001$
$Collateral\ Type_{Real\ Estate\ other\ corporates}$	-0.0772
$Collateral\ type_{Both\ Accounts\ Recievable\ and\ Inventory}$	$\mathbf{+0.0209}$
$Colleteral\ Type_{Real\ Estate\ other\ Residential}$	-0.0311
$Collateral\ Type_{Real\ Estate\ Commercial}$	-0.0437
$Collateral\ Type_{Inventory\ Stock}$	$\mathbf{+0.0517}$
$UCR\ After\ Default_7$	$+0.1009$
$UCR\ After\ Default_8$	$+0.1009$
Constant	$+0.0437$

It is seen that the model tree has different variables in his linear model than the regression tree. In this result we see the same phenomena as we saw in the output of the linear regression model. So again there are some positive weights assigned to some collateral types. Moreover we can see here again that UCR 7 or 8 has on average a higher LGD.

In Table 19 the RMSE and MAE of the different models can be found.

Table 19: Results for all the models

Model	RMSE	Std. RMSE	MAE	Std. MAE
Linear Regression	0.2641	0.0194	0.1587	0.0141
Regression Tree	0.2684	0.0164	0.1530	0.0131
Model Tree	0.2652	0.0172	0.1500	0.0139

A first thing that is seen is that it looks like that linear regression has the lowest RMSE. This is striking, because one would say that the model tree would give no worse result than linear regression. However it should be tested if the difference is significant. Again the result in Weka shows that H_0 cannot be rejected, so it cannot be said that linear regression tree has significant better results.

Furthermore it is seen that also in this case the error rates of the new models are higher than the current model. So apparently with this dataset none of the model techniques that are used in this thesis are able to find an equal or better model than the current model. This is on the one hand striking, because one would say that it should be possible to find the current model with the regression tree or the model tree. On the other hand as I already said in Chapter 4, none of these techniques shall definitely find the global optimum, which these results also show.

6.4 Predicting with cashflow data

In this section I tried to predict the LGD with the cashflow data. The cashflow dataset is a different dataset that contains for each facility, quarter snapshots about the outstanding and the amount of provision that has been taken. To this dataset for each facility both the LGD of the counterparty as the total LGD are added. Furthermore, for each quarter snapshot the period is calculated, which is in this case the time in months between the moment of default and the moment that the quarter snapshot is taken. Moreover, the dataset contains the ratio $\frac{Provision}{Observed\ Outstanding}$, denoted as ProvToOut. Finally, also the outstanding at default is added to it.

This dataset is interesting, because the provision indicates the banks expected losses on their loans. Besides, the more closely the quarter comes to the resolve date, the better it should be possible for the bank to indicate what the loss will be. So it should be possible to come up with a better model with this dataset.

6.4.1 Predicting the in-default LGD with cashflow

Again the first model considered is the linear regression model. The result of this model can be found in Table 20.

Table 20: Involved variables with weights

Variable	Weight
$Period$	$+0.0042$
Constant	$+0.1163$

It can be seen that the linear regression technique finds only one interesting variable, namely Period. Strangely enough it is not the variable Provision that is a variable in this model, as one would expect.

Next the regression tree is considered. This model outputs a very complicated tree with 731 leaves.

Finally the model tree is considered. This model outputs a very complicated tree with 147 leaves.

The error results of all the three models can be found in Table 21.

Table 21: Results for all the models

Model	RMSE	Std. RMSE	MAE	Std. MAE
Linear Regression	0.3666	0.0137	0.2525	0.0103
Regression Tree	0.2732	0.0112	0.1529	0.0090
Model Tree	0.3149	0.0124	0.1845	0.0101

It is seen that the regression tree outputs the best results. This is striking, because one would say that the model tree can output the same as the regression tree or better. However, it should be checked if the differences are significant. Weka outputs that Regression Tree is significant better than linear regression at a significance level of 5% . Weka also outputs that it is not possible to say that Regression Tree is better than Model Tree.

6.4.2 Predicting the total in-default LGD with cashflow

In this paragraph I tried to get better results by predicting the total LGD. The linear model of the linear regression algorithm can be found in Table 22.

Table 22: Involved variables with weights

Variable	Weight
$Period$	$+0.0042$
Constant	$+0.1114$

Again only the variable Period is involved and not the variable Provision, which would be expected.

Both the regression tree algorithm as the model tree outputs again a very complicated tree.

The errors of the models can be found in Table 23.

Table 23: Results for all the models

Model	RMSE	Std. RMSE	MAE	Std. MAE
Linear Regression	0.3385	0.0095	0.2387	0.0039
Regression Tree	0.2569	0.0111	0.1484	0.0049
Model Tree	0.2858	0.0105	0.1697	0.0064

It can be seen that of the new models considered, regression tree gives again the best results. However to know this for sure, it should be tested. The results in Weka shows that regression tree is indeed significant better than linear regression at a significance level of 5%, but it cannot be said that the regression tree is significant better than model tree. Furthermore, it is seen that also in this case the error rates of the new models are higher than the current model. So also with this dataset none of the model techniques that are used in this thesis are able to find a better model than the current model so far.

6.4.3 Predicting the loss/no Loss with cashflow

In this paragraph it is tried to predict whether there will be a loss on a defaulted loan or not with the cashflow data. Therefore again a new variable is introduced, indicating whether there was a loss on a loan or not. With the help of a decision tree I tried to find a model that predicts this new variable.

The output of the decision tree can be found in Figure 14.

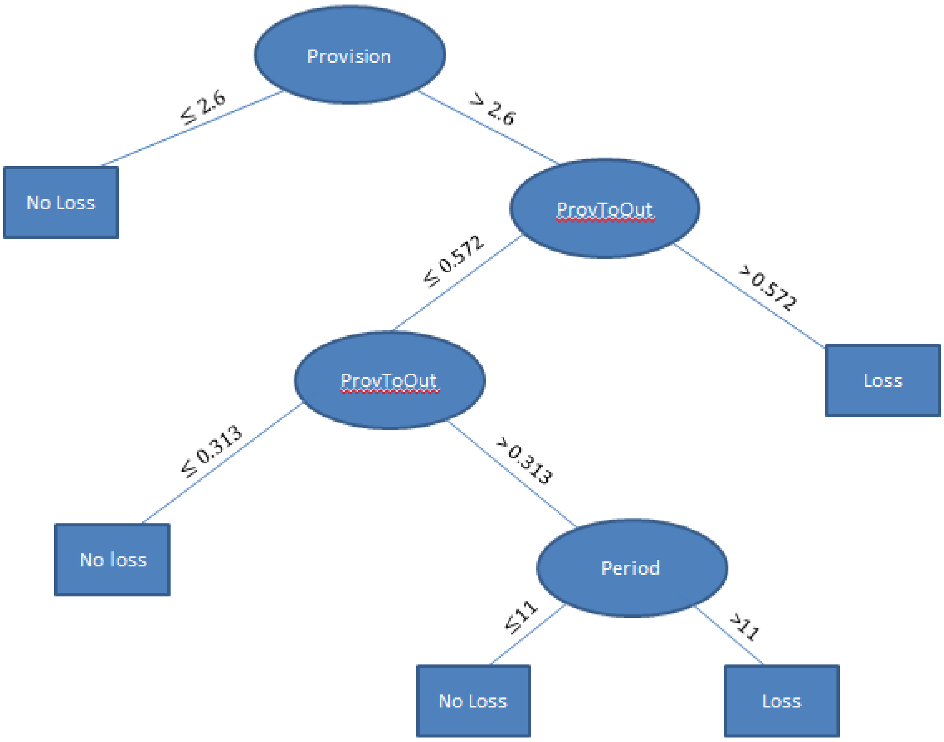


Figure 12: Output Decision tree

The decision tree outputs a model that depends on Provision, $\frac{Provision}{Observed\ Outstanding}$, and Period.

Now it is interesting to see if it is possible to find a good model for predicting the LGL. The output of linear regression model can be found in Table 24.

Table 24: Involved variables with weights

Variable	Weight
$Period$	$+0.0025$
$ProvToOut$	-0.0013
Constant	$+0.5922$

So according to the linear regression method Period and $\frac{Provision}{Observed\ Outstanding}$ are the important variables. A strange thing that is seen here is the negative weight assigned to ProvToOut. So the higher this ratio the lower the LGL. However, the constant is also pretty high, so maybe this ratio has to compensate this somehow.

The regression tree and the model tree outputs exactly the same tree, which can be found in Figure 13.

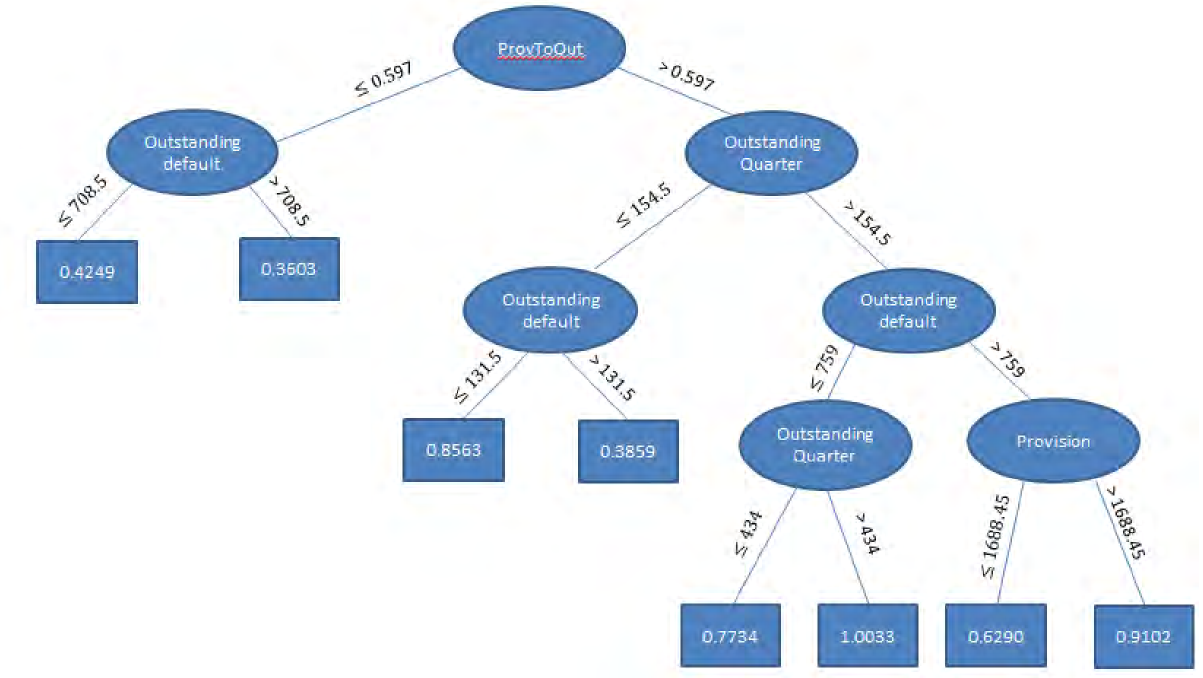


Figure 13: Output Model and Regression tree for LGL

It can be seen that in both models the variables $\frac{Provision}{Observed\ Outstanding}$, Provision, Outstanding at Default, and Outstanding Quarter are involved. In this tree we see the same kind of things as already mentioned in former paragraphs. So a lower provision leads to a lower LGL. Furthermore, it can be seen that in general it holds that a lower $\frac{Provision}{Observed\ Outstanding}$ leads to a lower LGL. This is exactly what we would expect.

It is interesting to see what the error rates of these models are. This can be seen in Table 25.

Table 25: Results LGL for all the models

Model	RMSE	Std. RMSE	MAE	Std. MAE
Linear Regression	0.4084	0.0140	0.3269	0.0067
Regression Tree	0.3584	0.0100	0.2727	0.0084
Model Tree	0.3584	0.0100	0.2727	0.0084

It looks like that regression tree or model tree is the best technique to use in this case. It can be seen that those errors are higher than for predicting the LGD. The output of the corrected t-test tells is that this difference is indeed significant at a significance level of 5%.

However this does not tell you anything yet, because it is not possible to compare the LGL with the LGD. So with these results we have to find the RMSE and the MAE for the LGD.

The model for the LGNL is of course the same as (6.2.1) with a RMSE and MAE of 0.

Now it is interesting to see what the overall error rate is of predicting the LGD this way. Therefore the three models should be combined. So first we should classify whether a loss occurs or not with Figure 12. Then for each record, either the LGNL model ((6.2.1)) or the LGL model (Figure 13) should be applied, depending of the outcome of the decision tree.

The final results of this can be found in Table 28.

Table 28: Results weighted error rate all instances

	RMSE	MAE
Results	0,3083	0,1459

It is seen that the overall RMSE is 0,3083 and the overall MAE is 0,1459. This result is unfortunately not better than the current model.

6.5 Predicting the LGD

In this section it is tried to find a suitable model for the LGD of performing loans. Due to the fact it is now tried to predict the LGD before default, all the variables that are only known from the moment of default are deleted.

In the first paragraph the results of the current model are described. In the second paragraph the results of the new models are described

6.5.1 Current Model

As already mentioned the current model for the LGD for performing loans is only based on the coverage ratio. In Table 29 the results can be found.

Table 29: Results current LGD model

Model	RMSE	Std. RSME	MAE	Std. MAE
Current Model	0.3244	0.0429	0.2189	0.0530

It is seen that the RMSE and MAE for the LGD for performing loans are higher than for the in-default LGD. That is not very surprising, because there is more information available when loans are already in default than when loans are not yet in default.

6.5.2 New Model

Again it is tried to build a new model with the three different techniques. First it is tried with the linear regression technique. The results can be seen in Table 30.

Table 30: Involved variables with weights

Variable	Weight
$CP\ type_{Corporates}$	-0.0762
$Country_{Netherlands}$	$\mathbf{+0.0472}$
$IND\ OOE_{\geq 1\mathrm{mln}}$	-0.0509
$Industry_{Industrials}$	$+0.0224$
$Industry_{Oil\&Gas}$	$+0.5898$
$Industry_{Telecomunications}$	$+0.3049$
$Collateral\ Type_{Real\ Estate\ other\ corporates}$	-0.1286
$Colleteral\ type_{Both\ accounts\ receivable}$	$\mathbf{+0.0570}$
$Colleteral\ type_{Inventory\ (semi\ finished\ goods)}$	-0.0282
$Colleteral\ type_{Non\ Marketable\ securities}$	$\mathbf{+0.0524}$
$Colleteral\ type_{Real\ Estate-residential}$	-0.0394

$Colleteral\ type_{Real\ Estate-Commercial}$	-0.0490
Constant	$+0.1781$

It can be seen that there are many variables involved. It can be seen again the Corporates have on average a higher LGD than Private Individuals. Moreover, it can be seen again that there are collateral types with positive weights. A last striking thing that can be seen is that according to this model counterparties in Netherlands have on average a higher LGD. A good explanation for this result is hard to find.

The next technique that it is tried is again the regression tree. The tree can be found in Figure 14.

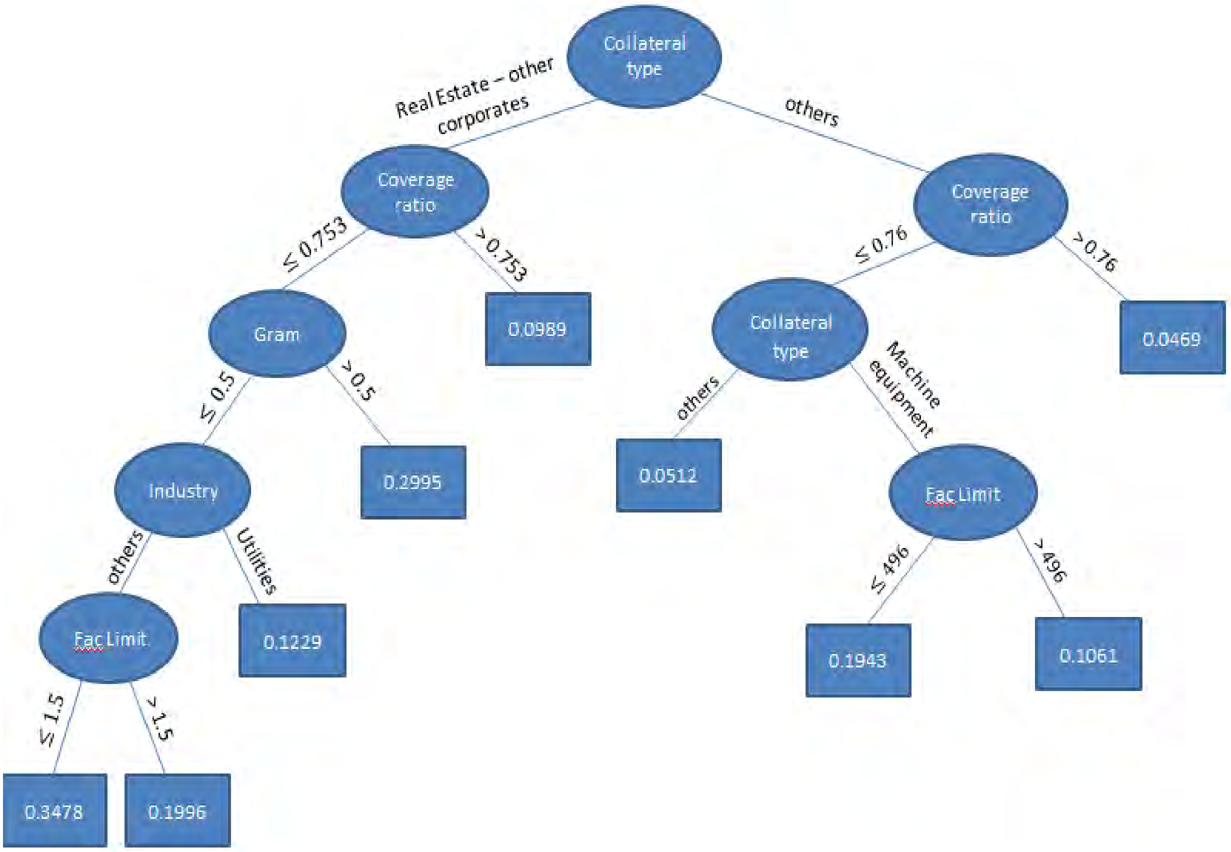
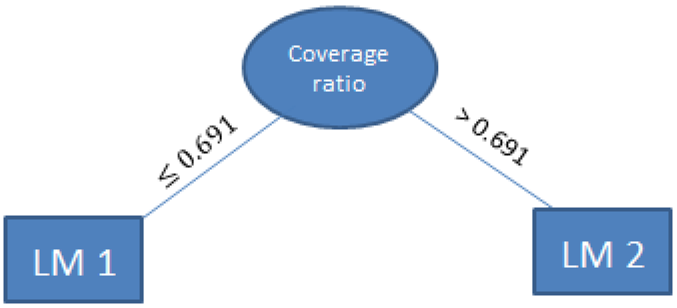


Figure 14: Regression tree for LGD

It can be seen that the variable colleteral_type, coverage ratio, guarantee amount, industry, and facility limit are involved. Furthermore, it can be seen here that now, when the variable period is not available anymore, a higher coverage ratio leads to lower LGDs in all the cases. It can also be seen that a higher facility limit leads to a higher LGD. This is not strange, because a higher limit can result in higher loss, and thus a higher LGD.

The last technique considered is the model tree. The result can be found in Figure 15.



LM 1:

$$LGD = -0.1508 * CP\_type_{Corporate} - 0.076 * Coverage\ Ratio - 0.0813 * IND\_OOE_{\geq 1mln} + 0.1172 * Industry_{Technology} + 0.9441 * Industry_{Oil\ \&\ Gas} - 0.0025 * Industry_{Telecommunications} + 0.2378$$

LM 2:

$$LGD = -0.0414 * CP\_type_{Corporate} - 0.023 * IND_{OOE\ \geq 1mln} + 0.0036 * Industry_{Technology} + 0.0038 * Industry_{Oil\ \&\ Gas} - 0.889 * Industry_{Telecommunications} + 0.0754$$

Figure 15: Model tree for LGD

It is seen that the tree only consist of one variable. However in the linear models in the leaf nodes, more variables are involved.

In Table 31 the RMSE and MAE are given for the different techniques.

Table 31: Results new LGD model

Model	Mean RMSE	Std RMSE	Mean MAE	Std MAE
Linear Regression	0.3063	0.0325	0.1863	0.0118
Regression Tree	0.3070	0.0310	0.1836	0.0113
Model Tree	0.3065	0.0305	0.1836	0.0097

The result in Weka shows that it is not possible to say whether the difference between the models is significant. This is not quite strange, because the difference is very small.

6.6 Predicting the total LGD

In this section it is tried to find a suitable model for the total LGD of performing loans. First the results of the linear regression technique are given, which can be found in Table 32.

Table 32: Involved variables with weights

Variable	Weight
$CP\ type_{Corporates}$	-0.0794
$Industry_{Oil\&Gas}$	$+0.6128$
$Industry_{Telecomunications}$	$+0.3302$
$Collateral\ Type_{Real\ Estate\ other\ corporates}$	-0.1164
$\mathbf{Colleteral\ type_{Both\ accounts\ receivable}}$	$\mathbf{+0.0598}$
$Colleteral\ type_{Real\ Estate-residential}$	-0.0306
$Colleteral\ type_{Real\ Estate-Commercial}$	-0.0393
Constant	$+0.1778$

Again it can be seen that there are many variables involved. Moreover, it can be seen again that Corporates have on average a lower LGD than Private Individuals. Furthermore it can also be seen that there are again some collateral types with a positive weights.

The results of the regression tree can be found in Figure 16.

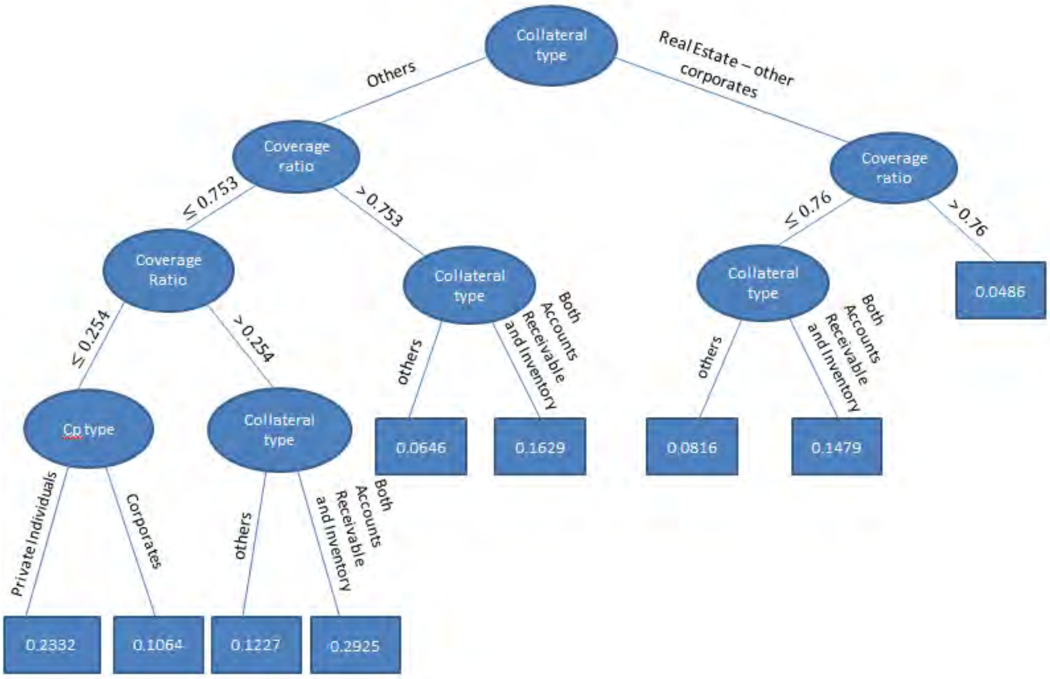
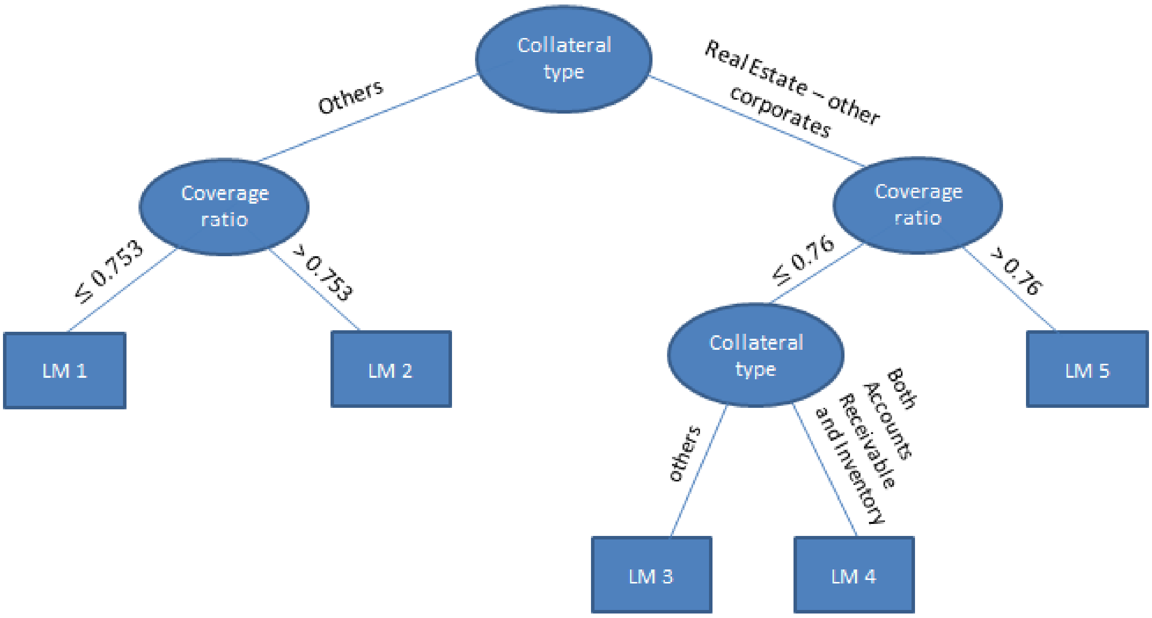


Figure 16: Regression tree for total LGD

It can be seen that the variables collateral_type, coverage ratio, and CP_type are involved in this model. Furthermore, it is seen here that a higher coverage ratio does not always leads to a lower LGD.

The results of the model tree can be found in Figure 17.



$$LM1:$$

$$Total\ LGD = -0.0755 * Coverage\ ratio + 0.968 * Industry_{Oil\&Gas} - 0.0007 * Colleteral\ type_{Real\ Estate\ Other\ Corporates} + 0.1353 * Colleteral\ type_{Both\ Accounts\ Recievable\ and\ Inventory} + 0.1828$$

$$LM2:$$

$$Total\ LGD = +0.0264 * Industry_{Oil\&Gas} - 0.0007 * Colleteral\ type_{Real\ Estate\ Other\ Corporates} + 0.1085 * Colleteral\ type_{Both\ Accounts\ Recievable\ and\ Inventory} + 0.063$$

$$LM3:$$

$$Total\ LGD = -0.0014 * Coverage\ ratio + 0.0037 * Industry_{Oil\&Gas} - 0.0007 * Colleteral\ type_{Real\ Estate\ Other\ Corporates} + 0.0033 * Colleteral\ type_{Both\ Accounts\ Recievable\ and\ Inventory} + 0.0814$$

$$LM4:$$

$$Total\ LGD = -0.1176 * Coverage\ ratio + 66.3003 * Industry_{Oil\&Gas} - 0.0007 * Colleteral\ type_{Real\ Estate\ Other\ Corporates} + 0.0044 * Colleteral\ type_{Both\ Accounts\ Recievable\ and\ Inventory} + 0.1901$$

$$LM5:$$

$$Total\ LGD = 0.0037 * Industry_{Oil\&Gas} - 0.0007 * Colleteral\ type_{Real\ Estate\ Other\ Corporates} + 0.0291 * Colleteral\ type_{Both\ Accounts\ Recievable\ and\ Inventory} + 0.0388$$

Figure 17: Model tree for total LGD

In the output of the model tree of Figure 17, there is something striking about the fourth linear model. Here a weight of 66.3003 is assigned to the variable industry, which is quite high. Probably it has to compensate the negative weight assigned to the coverage ratio.

In Table 33 the results of the different techniques can be found.

Table 33: Results new total LGD model

Model	RMSE	Std.	MAE	Std.
Linear Regression	0.2854	0,0169	0.1778	0,0092
Regression Tree	0.2846	0,0169	0.1741	0,0090
Model Tree	0.2838	0,0180	0.1734	0,0099

It looks like the model tree gives the best result. It has both the lowest RMSE and MAE. However to ensure that the difference is significant a test should be applied. The output in Weka shows that is not possible to reject H_0, so it is not possible to say that the difference is significant, which is of course not very strange seen the small difference.

Besides all the three techniques have a lower error than the current model. Again this should be tested. However, in this case it is much more difficult to test whether the new model is significant better than the old model. This because it is not possible to perform the same cross validations on the current model as on the new model. So in this case I decided to take each record of the dataset and perform a statistical test on all the records. In this case the paired Wilcoxon-test is used, which is explained in Appendix C. The result shows that it is possible to reject H_0 at significance level of 0,05%. So it can be said that the difference is significant.

7 Conclusion

In this thesis I tried to build models that predict the in-default LGD and the LGD for performing loans “better” than the current models. It was tried to build different kind of models with different kind of techniques, to see which model approximates the LGD the best and to give the bank more possible solutions for their LGD problem. The techniques that were used are linear regression, regression tree, and model tree. By building the models it was tried to find a balance between accuracy and transparency.

First a new model for the in-default LGD was build. This was first tried with a dataset that contains both LGD as PD information. However, by combining the PD and LGD dataset 90% of the data was lost, because there was no rating information available for 90% of the instances. This resulted in a dataset that contained only 518 instances, from which 504 instances had no loss.

In the models, based on this small dataset, no PD variable was involved. That’s why it was also tried to build a better model with the large dataset, consisting of 4791 instances. This dataset did contain PD variables, but these variables had only values for 518 of the 4791 instances. For all the other instances, these values were missing in the dataset. However, with this large dataset it was not possible either to find a better model for the in-default LGD than the current model.

Next, I tried to predict whether a loss occurred or not, and then tried to build separate models for the LGNL and LGL. Unfortunately, this did not result in a better model than the current model. Moreover, I tried to get a better model by predicting the total LGD, which is the combined LGD of the counterparty and the LGD of the guarantor. This did also not result in a better model than the current model.

Moreover, it was tried to build an in-default LGD model based on the cashflow dataset. This cashflow dataset contains the quarter snapshots of the provision. Predicting the LGD directly did not result in a better model than we already had. If we first predict whether a loss occurred or not, and then built models on both the LGL and LGNL separately, still no better model was received.

Finally, I also tried to find a model for the LGD for performing loans. Again first a model was built for the LGD directly. This gave already a slightly better RMSE and MAE than the current model. Second, it was tried to build a model for the total LGD again. Both the regression tree technique as the model tree technique gave lower RMSE and MAE than the current model.

So for the in-default LGD no better model was found in this thesis.

For the LGD of performing loans it is optimal to predict the total LGD. Which of the models to use, depends on what the bank really wants. With the regression tree it is possible to hold the same model structure as the current model, but now with different risk drivers involved. Figure 16 shows what the model looks like if we use this technique.

The model tree is the most complex model and outputs a continuous function. The model tree cuts the distribution of the LGD into different pieces and predicts each piece of the distribution by a linear function. This results in a so called piecewise linear function. It might be difficult for the bank to implement a (piecewise) linear function. However, it can be possible to divide the output of this function into buckets. Then we have buckets based on the LGD, rather than on coverage ratio as in the current model.

This looks for example like this:

Output LGD function	Assigned LGD
<0	0,005
0 < LGD < 0.1	0.05
0.1 < LGD < 0.2	0.15
Etc.	Etc.

For further research it would be interesting to add variables like seniority of the loan and the reason of default. Especially the last one can be interesting, as it might explain the high number of totally recovered instances.

For further research I would also recommend to try to get a much larger dataset with both LGD as PD variables. I would recommend investigating why there is no rating information available for 90% of the instances. This, because a dataset of 518 instances is too small to build real confident model on (at least with the techniques that are used in this thesis). Besides, that dataset also contains only 14 non-cured case, which is too less and not representative to build a confident model on.

So if it is possible to get a larger and more representative dataset with both PD and LGD variables, with also the variables seniority and reason of default added to this, then it might be possible to get a much better model with one of the techniques that are used in this thesis.

Bibliography

[1] E. Winands, slides Inleiding Risk Management, *(2009)*

[2] Wikipedia, Basel II

[3] T. Schuerman, What Do We Know About Loss Given Default, *Credit Risk Models and Management (2) (2004)*

[4] Basel Committee on Banking Supervision, History of the Basel Committee and its memberships, *(2009)*

[5] H. Esrel, Basel I and Basel 2: History of an Evolution, *(2011)*

[6] Basel Committee on Banking Supervision, International convergence of capital measurements and capital standards, *(1988)*

[7] B.J. Balin, Basel 1, Basel 2 and Emerging Markets: A Non Technical Analysis, *(2008)*

[8] J. Puts, Bank Balance Sheet Optimization under Basel III, *master thesis (2011)*

[9] Basel Committee on Banking Supervision, International convergence of capital measurements and capital standards, *(2005)*

[10] Basel Committee on Banking Supervision, Basel III: International framework for liquidity risk measurements, standards and monitoring, *(2010)*

[11] Internal Law Office, Basel 3, *(2010)*

[12] Clayton UTZ, Overview of Basel III – Minimum Capital Requirements and Global Liquidity standards, *(2011)*

[13] Norea, rapport Basel II, *(2010)*

[14] B. Stroomer, Predicting downturn LGD for a confidential portfolio, *master thesis (2009)*

[15] Basel Committee on Banking Supervision, Consultative Document - The Internal Ratings Based Approach, *supporting document for the new Basel Accord (2001)*

[16] B. Bode, L. Deborgies-Sanches, L. Sokolova, M. van Beest, G. Chatzis Bank Loan Loss Given Default: A European Perspective

[17] E.P van Klaveren, W. Yip, A. Maciel, A. Praagman, Y. Li, F. Pardoel, Technical Model Documentation*, internal paper (2011)*

[18] http://info.worldbank.org/governance/wgi/pdf/rl.pdf

[19] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, *second edition (2005)*

[20] J. Zurada, A.S. Levitan, J. Guan, A Comparison of Regression and Artificial Intelligence Methods in a Mass Appraisal Context, *journal of Real Estate Research, 33 (3) (2011), pages 349-388*

[21] T. M. Mitichell, Machine Learning, *International edition (1997)*

[22] www.baseliii.nl

[23] J.R. Quinlan, Learning with Continuous Classes, *proceedings AI'92, pages 343-348*

[24] J. Demsar, Statistical Comparison of Classifier over Multiple Data Set, *Journal of Machine Learning Research 7 (2006). Pages 1-30*

[25] A.C Eberhart, William T. Moore, R.L. Roenfeldt, Security Pricing and Deviations from the Absolute Priority Rule in Bankruptcy Proceedings, *Journal of Finance 45 (5) 1990, pages 1457-1469*

Appendix A : Description of variables

Variables	Description
ind_cp_type	Counterparty type (Corporates, Private Individuals)
ind_country	Land of CP (NL, NON-NL)
ind_country2	Land of CP (NL, FR, others)
segmentx	Combination of cp_type and country
AGIC_OID	This code contains information about the industry
CP_ID	Counterparty identity number
CST_INCR	Cost incurred (recovery cost etc,)
DT_RESOLV	Date of resolved
EFF_DT_OF_DFLT	Real date of default
FAC_ID	Facility ID
FAC_LIMIT_YEAR	Limit EAD of facility
MAX_WRT_OFF_DT	Date where the write off took place
NM	Name of Counterparty
OUTSTANDING_DEFAULT	Observed EAD at default, total outstanding (times 1000)
OUT_QUARTER	Outstanding quarter, last outstanding considered for non resolved cases (times 1000)
PRVSN	The amount of Provision that has been taken (times 1000)
PRVSN_DT	Last date when Provision was made, or current date, when running.
SEGMENT	The segment within the bank where the data is from
WRT_OFF	Amount of write off
loss_dt	The date when the bank knows a loss is going to occur
gram	Guaranteed amount, amount that is exposure to the guarantee (times 1000)
grpc	Guarantee proceeds. The part of the GRAM that has been recovered (times 1000)
OUTSTANDING_YEAR	Observed outstanding 1 year before default (times 1000)

expwrt	Expected write off, is only for non-resolved cases
expout	Expected outstanding, is only for non-resolved cases
resolved	Is it resolved or not
loss	The total amount of loss including guarantee (times 1000)
recovery	Amount that is recovered (times 1000)
interest_rate	Interest rate
period	Workout period in years
RecoveredPart	Max(recovery,0) (times 1000)
NonRecoveredPart	The total amount of loss including guarantee (times 1000)
RecoveredPartDiscounted	Discounted recovered part (times 1000)
NonRecoveredPartDiscounted	Discounted non recovered part (times 1000)
DirectCost_NPV	Net present value direct costs (times 1000)
DiscountedLoss	Total discounted loss including guarantee (times 1000)
LGD_perc	LGD percentage
wrt_off_oad	Write off percentage observed EAD
lgd_gr	LGD of the guarantee
LGD_cp	LGD of the counterparty
num_cp	Numerator counterparty, loss of the counterparty (times 1000)
den_cp	Denominator counterparty, EAD counterparty (times 1000)
weight	Denominator counterparty, EAD counterparty (times 1000)
LGD_CLASS_new	LGD class
LGD_CLASS	LGD class
COVERAGE_new	Coverage percentage
COUNTRY_RESIDENCE	Real country
COUNTERPARTY_TYPE	Type of the counterparty
BU	Business Unit
UCR_1Y_PRIOR	UCR 1 year before default

NET_COLLATERAL_FINAL	Net collateral value (times 1000)
ead	Exposure at Default, predicted EAD at default, based 1year before default (times 1000)
ind_ooe	Indicator of outstanding (< 1mln or ≥ 1mln)
year	Year of default
Default_Id	ID default
Quick ratio	Cash+ Marketable Securities + Accounts Receivable / Current Liabilities
OAAM/TA	Own And Associated mean / Total assets
RE/TA	Retained Earnings / Total assets
TD/EBITDA	Total Debt / Earnings before interest, taxes, depreciation and amortization
Current Ratio	Current Assets/Current Liabilities
Net operating Margin	Net operating Margin
Accounts Payable/COGS	Accounts Payable/Cost of goods sold
UCR_After_Default	UCR after default (6,7 or 8)
Collateral Type	Type of Collateral

Appendix B : Corrected t test

This Appendix, based on [19], explains how the corrected t-test of Weka works. The reason for performing a statistical test is that we want to say which technique is the best. To say this, It is not enough to look to the average error rate over the cross validations. The reason for this is that the difference may be caused by an estimation error. Therefore statistical test are used to determine this. A statistical test gives confidence bounds on the error rate of the different techniques.

Now assume that we have unlimited data, that $x_1, x_2, \dots, x_k$ are the errors obtained by the 10-fold cross validation of the first technique and that $y_1, y_2, \dots, y_k$are the errors obtained by the 10-fold cross validation of the second technique. We want to know if the mean of the errors of the first technique, denoted as $\bar{x}$, is significant different than the mean of the errors of the second technique, denoted as $\bar{y}$. It can be shown that if k is large enough, both means have a normal distribution.

Remember form Chapter 4.3.4 that if X is a normally distributed random variable with mean μ and variance σ^2, then $\frac{X-\mu}{\sqrt{\sigma^2}}$ is standard normally distributed. So if the actual mean μ, which is unknown, is subtracted from $\bar{x}$ and standard deviation $\sqrt{\sigma^2}$ is divided from this, we get a standard normally distributed random variable. However the variance is also not known and should be estimated from the errors $x_1, x_2, \dots, x_k$, which is $\frac{\sigma_x^2}{k}$. If we know denote with$\widehat{\sigma^2}$, the estimator of the true variance σ^2, then $\widehat{\sigma^2} = \frac{\sigma_x^2}{k}$.

Now due to the fact that the variance is estimated, $\frac{X-\mu}{\sqrt{\widehat{\sigma^2}}}$ is not standard normally distributed, but has a Student t distribution with k-1 degrees of freedom. So to get confidence limits for the true error rate μ we have to use the confidence table of the t distribution.

Now, we want to check whether the difference between $\bar{x}$ and $\bar{y}$ is significant. If $d_i = x_i - y_i$ is the difference between the i-th error of x and y, assuming that x_i and y_i are received from the same dataset, then the mean of the differences, denoted with $\bar{d}$, is the same as the difference of the means ($\bar{d} = \bar{x} - \bar{y}$). Under the null hypothesis this difference is 0 ($\bar{x}$ and $\bar{y}$ are the same). Under this null hypothesis $t = \frac{\bar{d}-0}{\sqrt{\frac{\sigma_d^2}{k}}}$ has a t distribution with k-1 degrees of freedom. In this formula σ_d^2 is the variance of the differences. If we take as significance level 5%, then with the student's t table with the corresponding degrees of freedom it is possible to get the confidence limit z. Now if the value of t is bigger than this z, we reject the null hypothesis (H_o) and accept the alternative hypothesis (H_1), which means that $\bar{x}$ is significant bigger than $\bar{y}$, so technique 2 is significant better than technique 1.

However up till now we assumed that the data is unlimited, such that it is possible to select independent sets. In practice this of course not the case, often we have a dataset of limited size. Now it is possible to perform different cross validations with different seeds, to get enough samples to get an estimate of the error rate. However these samples are not independent from each other. So a small correction of the standard t-test should be applied in this case. The corrected t-test uses now the following statistic[19]: $t = \frac{\bar{d}-0}{\sqrt{(\frac{1}{k}+\frac{n_1}{n_2})\,\sigma_d^2}}$. Here k is the total number errors computed. If you perform 10 times a 10 fold Cross Validation, which is done in this thesis, then k=100. In this formula *n_1* and *n_2* are respectively the number of instances used in cross validation for training and for testing. So if you use a 10-fold cross validation $\frac{n_1}{n_2} = \frac{0.1}{0.9}$. Finally σ_d^2 is the variance of the 100 computed error rates.

Appendix C : Paired Wilcoxon test

This appendix, based on[24], describes the paired Wilcoxon test. The Wilcoxon is, rather than the t-test, a non-parametric statistical test. The idea of the test is to rank the difference between two techniques, while ignoring the signs. Furthermore, the test compares the ranks of the positive and negative differences.

Let's denote with d_i again the difference between the i-th error of the two techniques. Now these differences are ranked according their absolute values. Let's denote with R^+ the sum of the ranks where the first technique is better than the second techniques and with R^- the sum of the ranks where the second technique is better than the first techniques. The cases where $d_i = 0$ are split evenly across R^+ and R^-. If there are an odd number of cases where $d_i = 0$, one is ignored. In formula form it looks like this[24]

$$R^+ = \sum_{d_i>0} rank(d_i) + 0.5 \sum_{d_i=0} rank(d_i)$$

$$R^- = \sum_{d_i<0} rank(d_i) + 0.5 \sum_{d_i=0} rank(d_i)$$

Now if we denote with T the smallest one of these two, then the statistic can be defined as follows [24]:

$$z = \frac{T - \frac{1}{4}N(N+1)}{\sqrt{\frac{1}{24}N(N+1)(2N+1)}}$$

It is shown that for large number of data this statistic is approximately standard normally distributed. So in this case the null hypothesis is rejected if z is smaller than the confidence limit z, which can be obtained, from the standard normal table, for a given significance level α.