UNIVERSITY OF LJUBLJANA
FACULTY OF ECONOMICS

MASTER'S THESIS

PRACTICAL ASPECTS OF DEVELOPING AND VALIDATING CORPORATE PROBABILITY OF DEFAULT MODEL AND THE USE OF MACHINE LEARNING

Ljubljana, July 2016 ALJOŠA ORTL

AUTHORSHIP STATEMENT

The undersigned Aljoša Ortl, a student at the University of Ljubljana, Faculty of Economics, (hereafter: FELU), author of this written final work of studies with the title Practical aspects of developing and validating corporate probability of default model and the use of machine learning, prepared under supervision of prof. dr. Igor Masten.

DECLARE

1. this written final work of studies to be based on the results of my own research;
2. the printed form of this written final work of studies to be identical to its electronic form;
3. the text of this written final work of studies to be language-edited and technically in adherence with the FELU's Technical Guidelines for Written Works, which means that I cited and / or quoted works and opinions of other authors in this written final work of studies in accordance with the FELU's Technical Guidelines for Written Works;
4. to be aware of the fact that plagiarism (in written or graphical form) is a criminal offence and can be prosecuted in accordance with the Criminal Code of the Republic of Slovenia;
5. to be aware of the consequences a proven plagiarism charge based on the this written final work could have for my status at the FELU in accordance with the relevant FELU Rules;
6. to have obtained all the necessary permits to use the data and works of other authors which are (in written or graphical form) referred to in this written final work of studies and to have clearly marked them;
7. to have acted in accordance with ethical principles during the preparation of this written final work of studies and to have, where necessary, obtained permission of the Ethics Committee;
8. my consent to use the electronic form of this written final work of studies for the detection of content similarity with other written works, using similarity detection software that is connected with the FELU Study Information System;
9. to transfer to the University of Ljubljana free of charge, non-exclusively, geographically and time-wise unlimited the right of saving this written final work of studies in the electronic form, the right of its reproduction, as well as the right of making this written final work of studies available to the public on the World Wide Web via the Repository of the University of Ljubljana;
10. my consent to publication of my personal data that are included in this written final work of studies and in this declaration, when this written final work of studies is published.

Ljubljana, July 18th, 2016 Author's signature: _____

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

In spite of new financial regulations credit risk models (especially rating models) are gaining in importance. In 2014 the International Accounting Standards Board (hereinafter: IASB) published the complete version of the new international financial reporting standards 9 (hereinafter: IFRS 9), which will need to be implemented by banks till 2018. Tata (2015, p. 3) notes that the new expected loss impairment model of the IFRS will require financial institutions to recognize the credit loss from a financial asset at the first reporting date immediately after origination. The expected credit loss model will be a forward-looking approach that will provide greater convergence with the regulatory capital models. It will strongly motivate the use of rating systems[1] for impairment calculations. A valid credit rating system consists of estimates of probabilities of default (hereinafter: PD), loss given default (hereinafter: LGD) and exposure at default (hereinafter: EAD). These parameters form the basis for impairment calculations for individual borrowers and take into account not only historical but also estimated future risks. The heavy use of models has historically been linked to capital requirements calculations under advanced internal ratings-based approaches (hereinafter: IRB). The regulations for this approach are grounded in the capital requirements regulation (hereinafter: CRR) and the directive (hereinafter: CRD) of the European Commission (hereinafter: EC; 2013). Validated internal rating systems often allow banks to decrease their regulatory capital requirements. European Banking Authority (hereinafter: EBA; 2013b, p. 53) reports that banks with IRB approach have on average much lower risk-weighted assets and lower capital requirements for credit risk. In Slovenia, the central bank recently started to stipulate default rate and loss given default calculations with its new guidelines (in force since 27. 11. 2015). This will enforce banks to collect data that can later be used for modelling purposes.

All this creates a strong incentive for the development of rating systems. Besides the regulatory use, there is a variety of other uses, which gives the motivation for developing advanced rating systems. We made a list of potential uses of rating systems based on Samuels (2012, p. 19), Bank for International Settlements (hereinafter: BIS; 2001, p. 48) and Petrov (2015, p. 4):

- credit approval (better and automated allocation),
- capital requirements calculation (Basel IRB),
- impairments calculation (IFRS),
- credit portfolio limit setting,
- risk-based loan-pricing,
- business line capital allocation and business strategy,
- risk-based remuneration,
- reporting of the credit portfolio's risk dynamics to the bank management,
- performing stress tests.

[1] In the EC (2013, p. 94) the rating system is defined as all methods, processes, controls, data collection and IT systems that support the assignment of exposures to rating grades or pools, and the assessment of default and loss estimates. A rating system needs to cover both, borrower and transaction risk (usually through PDs and LGD's).

In this thesis we concentrate on developing a PD rating system for corporate borrowers of a small bank in Slovenia. The main technique that is used in practice and also in many scientific papers is logistic regression. This is also one of the techniques used in this thesis. Some papers that use the logistic regression approach are Altman and Sabato (2007), Hayden (2002) and Cipollini, Dainelli, & Giunta (2009). Examples of papers based on Slovenian data are articles of Jovan (2005) and Ortl and Gorenjec (2015).

Besides the benefits of the different uses of rating systems, there has also been a rapid development in methodologies for estimating probabilities of default in the last years. Lately there has been much written about breakthroughs and the global use of machine learning techniques in different fields of science, including finance. We will show that these algorithms are not only useful as an alternative to estimating probabilities of default by the common method of logistic regression, but also as complementary methods at various stages of rating model development. In this thesis we apply random forests (hereinafter: RF) and multilayer artificial neural networks (hereinafter: ANN). Random forests are also used for variable selection problems, similar as in Ortl and Gorenjec (2015). Brezigar-Masten and Masten (2012, pp. 10–14) similarly use classification trees. They augment the logistic regression with dummies that are created based on the final nodes of classification trees and confirm the improved predictive power when compared to the basic logistic regression. They find out that classification trees in general do not overperform logistic regression on the database of Slovenian firms, however, they can be used as a good complementary tool. For setting optimal grading structure we use a genetic optimisation algorithm called differential evolution (hereinafter: DE).

Falavigna (2006, p. 21), Angelini, Tollo, & Roli (2008, p. 747), Zhang, Hu, Patuwo, & Indro (1999, pp. 20–22) and Atiya (2001, p. 930) did some research on relevant articles and found that the results of ANN compared to logistic regression are mixed for the small and medium-sized enterprises (hereinafter: SME) portfolios. While most authors find out the superiority of ANNs, there are some articles where logistic regression outperforms ANNs. Atiya (2001, p. 930) and V. Jagrič, Kračun, & T. Jagrič (2011, p. 385) note that most authors use multilayer networks with a backpropagation training algorithm for predicting corporate defaults. Zhang et al. (1999, pp. 20–22) discuss also some other research where other ANN algorithms were tested.

Two comprehensive articles that compare standard logistic regression with machine learning algorithms are for instance Lessmann, Seow, Baesens, & Thomas (2015) and Falavigna (2006). Lessmann et al. (2015) is the most comprehensible study to our knowledge on the comparison of different machine learning algorithms for default prediction. They test 16 individual classifiers (i.e. logistic regression classification trees and multilayer ANNs), 8 homogenous ensembles and 17 heterogeneous ensembles[2] on 7 different real-world retail credit scoring datasets.

[2] Homogenous ensembles combine the predictions of multiple base models (i.e. random forests or boosting combine classification trees), whereas heterogeneous ensembles use different algorithms (example is HCES).

Lessmann et al. (2015, p. 31) identify the best models (taking into account 4 different performance measures) to be HCES (Hill-climbing ensemble selection), RF, boosting, ANN, support vector machines. While logistic regression still performs well, it lags behind aforementioned algorithms in their research. Interestingly enough, the new methods, which are even more advanced than ANN, perform worse. Lessmann et al. (2015, p. 42) actually propose the RF to be the new benchmarking model due to its simplicity of implementation and its good predictive accuracy.

Angelini et al. (2008, p. 744) use ANN with two hidden layers on the corporate data of Italian firms. As an alternative model they present ANN where they group input neurons by three (based on observations for one borrower for three different years) in a four-layer network. Jagrič et al. (2011, p. 384) implement the learning vector quantization (LVQ) neural network on a database of Slovenian retail borrowers, which resulted in an outperformance of the logistic regression model, especially, because it can better handle the properties of categorical variables. Zhang et al. (1999) compared ANN and logistic regression on a sample of manufacturing firms. The ANN significantly outperformed logistic regression.

In this thesis we do not want to concentrate only on building a PD rating model and the comparison of results for different methods for estimating PDs. We want to discuss and introduce also the implementation of the whole rating system, including the calibration of obtained PDs, structuring optimal rating grades, validating a model, monitoring the rating system through on-going validation and to propose some approaches for the actual implementation. We provide further references to the literature that deals with regulatory aspects and practical issues regarding development of such rating systems. Good examples of such books are Löffler and Posch (2007), Anderson (2007), Balthazar (2006) and Engelmann and Rauhmeier (2011). The latter consists of many articles of top-tier practioners in this field. The somehow more advanced topics can be found in Christodoulakis and Satchell (2008). However the regulations sometimes differ. In the European Banking Federation (hereinafter: EBF; 2012)[3] report they show different practices of European supervisors on some topics. Several EBA reports from 2013 (2013a[4]; 2013b[5]; 2013c[6]) also discuss the shortcomings of the current credit risk models regulation. Especially EBA (2013b) reports major differences across different supervisory jurisdictions. In future EBA (2015) will provide guidelines to cope with these issues as stated in its report. Examples of older technical standards or working papers in this field are Bank of England (hereinafter: BoE; 2013) in England, Oesterreichische Nationalbank (hereinafter: ONB; 2004) in Austria and Deutsche Bank (hereinafter: DB; 2003) in Germany. A very important document in this field is BIS (2001) and of course the currently valid CRR directive (EC, 2013). The hardest part for the supervisors is the question of how to properly

[3] Results are based on a questionnaire completed by 42 banks and 14 banking associations across Europe.
[4] This report discusses the modelling approaches for low default portfolios. Results are based on a survey involving 35 banks using the IRB approach for at least one of their low default portfolios.
[5] The report is based on data provided by 43 large EU banks for three different portfolio segments (residential mortgages, SME retail and SME corporate).
[6] This is a summary report on the comparability and pro-cyclicality of capital requirements under the IRB approach. It combines the results of five previous reports, including EBA (2013a) and EBA (2013b).

validate a rating system. It seems that the toughest question remains how to calibrate the predicted PDs to long-term default rates. BIS (2006) discusses the inappropriateness of current calibration testing techniques in case of emerging markets. However, BIS (2005) still serves as a starting paper when one is interested in model validation. Due to their importance and difficulties such tests are an interesting topic for doctoral or master dissertations and scientific papers. Examples are Wu (2008), Clavero-Rasero (2006), Maarse (2012) and also Rauhmeier and Scheule (2005). A great article from a practioner's point of view is Joseph (2005). It is accompanied by excel spreadsheets that cover most of the common validation tests.

As can be seen from this introduction, we are interested in all aspects of developing a PD rating system. The main research questions will however be limited to the estimation of the default probabilities with different algorithms. Our hypothesis is that machine learning techniques can provide better results compared to logistic regression, if we feed them with large quantities of data. Due to our data limitations we believe machine learning techniques will perform similarly well as logistic regression, but will not outperform them. The true usefulness of machine learning in our thesis will be in complementing logistic regression at different stages.

The structure of this thesis is closely related to the **flowchart in Appendix A** (note that the chapter numbers relate directly to the flowchart). The flowchart consists of all important steps in rating system development. In **Chapter 1** we discuss differences between models (quantitative versus qualitative) and expert opinions of loan officers. Each type has its own strengths and weaknesses. Not all approaches are appropriate for all loan segments (retail, corporate, institutions etc.). We will give the theoretical foundations for logistic regression, random forests and neural networks. **Chapter 2** is devoted to data preparation. First we discuss the different definitions of default, time horizons, dealing with cured borrowers etc. It is important to clean the data and treat the missing values and outliers. For logistic regression there are several different data transformations available. A very important aspect is a valid preparation of data samples. Before modelling we need to decide which variables have good predictive power for the problem at hand. The possible approaches are discussed in **Chapter 3**. In more detail we present the method of random forests variable importance. We will use it also later in the empirical part. **Chapter 4** discusses how to decide on the best model among all estimated alternatives. We need to make sure that the variables have economic interpretation, and that the model shows good discriminatory power out-ofsample. This means that good borrowers are assigned low PDs and bad borrowers are assigned high PDs. The PDs that we obtain from the model may have good discriminatory power, but they need to be calibrated to the default rates (hereinafter: DRs) that are expected for the underlying portfolio. Related to this are two rating philosophies. The standard is the point-in-time (hereinafter: PIT) philosophy that is closely related to one year PDs. The IRB approach requires banks to calculate long-term PDs, which is feasible with the through-the-cycle (hereinafter: TTC) approach or with hybrid approaches. All this is discussed in **Chapter 5**. Now that we have estimated our best estimates of PDs, it is a common practice to divide borrowers into different rating grades. In **Chapter 6** we show how to do this optimally with the use of a differential evolution optimisation algorithm. The whole system also needs to be properly implemented in the bank's information systems and

integrated into the processes. In **Chapter 7** we present the regulator's expectations and some good practices. After the model is implemented, we need to monitor its performance and stability. **Chapter 8** is devoted to the practical implementation of many different validation techniques. We discuss both, calibration and discrimination tests. Finally in **Chapter 9**, we perform all eight steps in an empirical analysis on real data.

1 MODELLING DECISIONS AND APPROACHES

1.1 Motivation for models

IRB requires a two-dimensional rating system. One dimension measures borrower specific risks and the other measures transaction/facility specific risks. With PDs we represent the borrower dimension. Transactional/facility specificities are usually incorporated in the LGD model (it reflects collateral related to the facilities). There are in general three ways to estimate PDs (Balthazar, 2006, p. 178; BIS, 2005, p. 18):

- In the **historical method**[7] the pooled PD for a risk grade is estimated using historical data on the average default rate among borrowers assigned to that grade over previous years (we assume that the best forecast for the future PDs are the historical default rates).
- The **statistical method** derives predicted probabilities of default with a statistical model.
- In the **mapping**[8] procedure we link the model rating with external benchmarks for which historical default rates are available (i.e. rating agency ratings).

Balthazar (2006, p. 179) notes that in some cases, none of these approaches will be possible. Then we need to apply expert opinions and conservatism, i.e. the qualitative approach. In practice, banks often use a combination of both systems. We will discuss this in detail later in the thesis. For the regulators, it is important to build transparent, understandable and acceptable rating models with high predictive power.

We have already discussed in the introduction that the use of rating models is motivated by the regulators. The motivation for statistical models is that they demonstrate superior discriminatory performance compared to heuristic models. In theory, the individual credit risk assessment by a credit analyst (loan officer) should always provide better credit score assessment because models cannot incorporate all data and are weak at interpreting qualitative data. On the other hand, also humans have limitations and biases. There exists rich "behavioural finance" literature that uses documented psychological biases to explain anomalies. These biases include for instance (Balthazar, 2006, p. 118; Falkenstein, Boral, & Carty, 2000, pp. 16–17):

- people tend to overestimate the precision of their knowledge,

[7] The common method is the migration matrices approach. For backtesting see ONB (2004, p. 124).
[8] We can try to replicate external ratings by using an ordered logistic regression (multinomial logit/probit) that gives the probabilities of belonging to n ratings and not only to two categories (default or not-default).

- they recall information related to their successes more easily than information related to their failures,
- they cannot systematically (at least not over time) analyse such a big quantity of data (for instance, if a company recently defaulted because of environmental problems, the loan officer will put too much negative weight on these issues in the near future),
- they have subjective biases,
- their judgement is focused on exceptions.

1.2 Segments

A good PD model is applied to a homogenous data pool, hence the importance of defining segments is huge. For a bank applying for IRB, it is to be assumed that different rating systems will be applied to corporate exposures, sovereign exposures (governments and the public sector), financial institution exposures, and to retail exposures[9]. Bigger banks would have much deeper granularity, as presented for instance in ONB (2004, pp. 17–28). Other possible segments are also small business owners, international companies, non-profit organizations and specialized lending (start-ups, project finance, real estate, commodity etc.). International banks need also deeper granularity with respect to different country risks, sometimes banks make sub-models for different industries. Note that the article 148 of the CRR directive (EC, 2013) requires IRB banks to eventually cover all segments.

Samuels (2012, p. 14) criticizes banks for too little granularity in risk modelling. It is hard to determine the "right" number of specific credit models[10]. ONB (2004, p. 11) and Balthazar (2006, p. 126) warn that the degree of granularity depends heavily on how homogenous the portfolios are and how much data is available. A big issue are the portfolios that exhibit low rates of default (e.g., lending to large corporations, financial institutions, sovereign portfolio, public sector enterprises, etc.). These segments rarely generate enough internal default data to support a robust statistical analysis or validation. Samuels (2012, p. 14) mentions a number of answers to this problem. One possibility is better external data, or the "expert judgment-based" models. A common statistical approach for modelling low default portfolios was developed by Pluto and Tasche (in Engelmann & Rauhmeier, 2011, p. 75). EBA (2013a, pp. 8, 17, 41) reports that banks often use mixed approaches that take into account internal and external data as well as expert judgement. They state that in their survey that 23 out of 35 banks use the standardised approach for low default portfolios. In the EBA report (2015, p. 49) it is recognised that there exist huge differences in estimated PDs and in approaches among different banks.

BIS (2001, p. 101) and article 148 of the CRR directive (EC, 2013) state that it is expected from the bank to cover most of the exposures under the IRB approach within a reasonably short period

[9] Retail is often segmented further into current accounts, consumer loans, credit cards, residential loans and private banking. For details on retail models see Porath (in Engelmann & Rauhmeier, 2011, p. 24) and Anderson (2007).
[10] A study from KPMG (2003, in Balthazar, 2006, p. 126) revealed important differences between the US and EU banks: in the US there were on average five non-retail scoring models and three retail, while in Europe the average was ten non-retail models and eight retail.

of time. They want to prevent so called "cherry-picking" opportunities during the transition from partial coverage to full implementation of IRB. Article 150 of the CRR (EC, 2013) allows banks to partially use the standardised approach only for non-significant business units and immaterial portfolios (i.e. for portfolios of institutions, central governments etc.). EBA (2013b, p. 71) reports that despite the regulation, many IRB compliant banks still have more than 50 % of their exposure under the standardised approach. Observations differ by banks and jurisdictions. EBA (2013a, p. 41) will in future reconsider to allow banks to apply permanent partial use of the standardised approach for portfolios where data limitations cannot yield robust results (i.e. low default portfolios).

In this thesis we focus on a PD model for corporates. Because of a rather small data sample we do not develop special models for different industries or for large corporates.

1.3 Qualitative approaches

In the previous chapter we have mentioned that it is especially hard for models to incorporate qualitative data. Balthazar (2006, p. 179) notices that fully automated statistical models may especially be useful for retail counterparties, where margins are small and volumes are high. On the other hand, for low default portfolios (hereinafter: LDP) and for large exposures, the amounts lent justify a more detailed analysis of each borrower. Scoring models are a very good supportive tool but in some cases may not be sufficient. A good rating process should integrate also qualitative aspects.

Examples of qualitative information are opinions of loan officers about the quality of the management of a company, quality of accounting systems, quality of organization structure, geographical location, views on the trend of the sector, the history of the banking relationship etc. Such information is not easily integrated into statistical models. It may be that a bank develops separate partial scoring functions for quantitative and qualitative data. ONB (2004, p. 52) discusses two approaches to combine qualitative and quantitative models: horizontal linking and vertical linking. Linking horizontally[11] means that the final score is a direct combination of quantitative and qualitative models. The other possibility is defined as "vertical linking". It means the incorporation of "overrides". In line with the approved methodologies such "overrides" allow credit experts to correct the model-based rating. Facts that are already used in a statistical or causal analysis module should not serve as the basis for later modifications by credit analysts. Adding overrides is important, as some data (most relevant data, future prospects, qualitative data) are only known to the credit analyst and cannot be covered by the quantitative module. Note that article 172 of the CRR directive (EC, 2013) requires that any manual modification of the automated rating proposal is possible only if the overrides are defined within clearly defined limits, documented and justified individually. Banks must record all the data used to give a rating (including override rating and pure model rating) to allow

[11] ONB (2004, p. 84) discusses optimisation and heuristic weighting approaches and link to fuzzy logic system.

backtesting (Balthazar, 2006, p. 116; Blochwitz and Hohl (in Engelmann & Rauhmeier, 2011, p. 259); ONB, 2004, p. 53).

Balthazar (2006, p. 179) notes that many banks use "qualitative scorecards"[12]. These are semi-automated formalized checklists/questionnaires of the qualitative elements. Credit officers answer a set of questions that have different weights. This produces a qualitative score for each company.

It is important to monitor the performance of overrides. In a good rating system we would expect a high match between suggested ratings and final ratings. The excessive use of overrides may indicate a lack of understanding or a bad rating model (ONB, 2004, p. 53).

1.4 Quantitative approaches

Balthazar (2006, pp. 119–122) and ONB (2004, pp. 32–50) provide a broad overview of quantitative models used in the PD modelling. In general we can distinguish between statistical models (i.e. discriminant analysis, regression models, i.e. logit, panel logit), machine learning (classification trees, random forests, neural networks, support vector machines etc.) and causal models (Merton structural model, hazard approach, cash flow simulations). Falavigna (2006, p. 19) discusses also some other (exotic) interesting approaches to default prediction.

Some methods are more appropriate for certain credit segments than others. For instance, to employ Merton like structural models one would need stock market prices, which is only feasible for large corporates and banks, hence their use is limited to listed companies. Cash flow (simulation) models are especially well suited to the specialized lending segment. Models differ in interpretability. Statistical approaches are much more interpretative than most of the machine learning techniques (excluding classification trees). The differences are also in economic theory foundations. With statistical approaches one verifies the economic meaning by checking signs of parameters. For machine learning approaches one needs to perform sensitivity analysis. Structural models are developed based on economic theory (option theory), so there is no need for testing it.

An interesting fact is that most of the methods are cross-sectional, because all covariates are related to the same period. It is possible to expand the cross-sectional input data to a panel dataset. Such models can integrate macroeconomic variables into the model and as such are useful for TTC PD estimates[13]. With macroeconomic variables we can improve the model because many macroeconomic data sources are more up-to-date than the borrowers' financial ratios. On the other hand, the oil price is available on a daily frequency. Macroeconomic variables can also be used in intuitive stress testing scenarios. The disadvantage is that introducing the time correlation (feature of panel models) in the estimation procedure is

[12] An example of a qualitative scorecard is in Balthazar (2006, p. 178).
[13] An example of random effects panel model for developing TTC PDs is a paper of BIS (2006).

cumbersome for discrete outcome variables (Hayden & Porath in Engelmann & Rauhmeier, 2011, p. 7).

Hayden and Porath (in Engelmann & Rauhmeier, 2011, p. 11) comment that the Basel requirements do not have any preference for any certain method. However, some models would need some modifications, if they are to be approved. BIS (2001, p. 47) emphasises the importance of a clear understanding of the key model inputs and the model's methodology. "Black box" models that are not well understood by bank personnel might be questionable (i.e. neural networks).

In the empirical part of this thesis we implement logistic regression, neural networks, and random forests algorithms. Now we provide theoretical foundations for these models.

1.5 Logistic regression

Logistic regression is the most used method in corporate default prediction. The main data inputs are financial ratios. We present the basics of this method mainly from the text of Löffler and Posch (2007, pp. 1–13) and also Hayden and Porath (in Engelmann & Rauhmeier, 2011, pp. 5–7).

Standard scoring models linearly combine dependent and independent variables. Let x denote the K explanatory variables[14] and b the weights (coefficients or parameters) attached to them. The representation of the model of regression is:

$$y_i = \beta' x_i + u_i \tag{1}$$

The variable y_i corresponds to defaults and takes binary values 0 or 1. The variable takes value 1, if the company defaulted the year following the one, for which we have collected the explanatory values, and zero otherwise. The overall number of observations is denoted by N.

$$y_i = \begin{cases} 1 & default \\ 0 & no\ default \end{cases} \tag{2}$$

After estimating the model, it returns scores. These are not yet default probabilities and not yet 0/1 votes.

$$Score_i = b_1 x_{i1} + b_2 x_{i2} + \ldots + b_K x_{iK} = \boldsymbol{b}' \boldsymbol{x_i}$$
$$\boldsymbol{x_i} = \begin{bmatrix} x_{i1} \\ \ldots \\ x_{iK} \end{bmatrix} \boldsymbol{b} = \begin{bmatrix} b_1 \\ \ldots \\ b_K \end{bmatrix} \tag{3}$$

[14] In machine learning vocabulary these are called features or attributes. In statistical language these are called independent variables.

To obtain default probabilities we need to link scores (linear predictors) to default probabilities first. In case of logistic regression this is done by applying a link function Λ over scores. If we use logistic distribution, then the resulting probabilities of default would be constrained to the interval 0 to 1. Also a monotonic relationship between $e^{b'x_i}$ and PD_i is assured because $\Lambda(.)$ represents a cumulative distribution function. If we apply the logistic distribution to our problem we get probability of default:

$$PD_i = Prob(Default_i) = \Lambda(Score_i) = \Lambda(\boldsymbol{b}'\boldsymbol{x}_i) = \frac{e^{b'x_i}}{1+e^{b'x_i}} = \frac{1}{1+e^{(-b'x_i)}} \quad (4)$$

For probit (it is the quantile function of a normal distribution) we would have to apply normal distribution: $\Phi(\boldsymbol{b}'\boldsymbol{x}_i)$. If we do the calculations, we see that higher scores correspond to a higher default probability. The scoring model should predict a high default probability for those observations that defaulted, and a low default probability for those that did not.

The parameters of logit models are usually estimated by the maximum likelihood method (ML)[15]. They are chosen in a way that the probability (likelihood) of observing the given default behaviour is maximized. As the method is very cited, we do not go into the details. We refer an interested reader to the references above.

One disadvantage of logit models is that they are prone to collinearity problems. It is a linear method, hence capturing non-linearities (although with proper transformations) is not an easy task. These two issues are an easy task for machine learning techniques such as random forests and neural networks.

1.6 Neural networks (ANN)

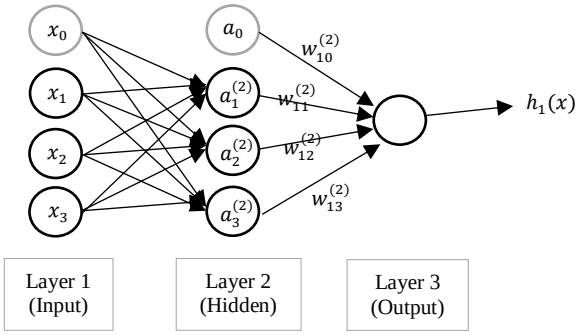
In the empirical part we use the multilayer artificial neural network (ANN). Atiya (2001, p. 930) emphasises the power of ANN when estimating nonlinear saturation effects. For instance, if the EBIT/Assets ratio changes from -0,1 to 0,1, the effect on the prediction of default might be much bigger compared to the change from 1,0 to 1,2. Also the effect of a firm with negative cash flow is being amplified when it has large liabilities (harder to finance for leveraged firms). ANNs are very good at modelling such nonlinearities and multiplicative effects.

The multilayer feedforward ANNs are based on the forward propagation and backpropagation algorithms with the sigmoid/logistic link function ($h=\sigma(z)=\frac{1}{1+e^{-z}}$). We present the algorithm on a three layer neural network as presented in Figure 1. We have $n=1,\ldots,N$….training examples /

[15] Hayden and Porath (in Engelmann & Rauhmeier, 2011, p. 6) note that it would be possible to use also the ordinary least squares estimator (OLS), but the coefficients would be heteroscedastic without the use of the weighted least squares estimator (WLS). The estimation of parameters would be inefficient, standard errors of the estimated coefficients would be biased.

observations, which consist of $N <x_n,y_n>$ pairs. The x_n is a feature vector for observation n (financial ratios for borrower n) and y_n is the output value for this observation (default or non-default event of borrower n). This chapter is based on three main courses/articles: Ng (2015), Hinton (2012) and Mitchell (1997). We combine their notations and make our own.

Figure 1. Neural network topology



Source: modified from A. Ng, *Machine Learning*, 2015.

$n=1,...,N$….training examples / number of observations (we have $N <x_n,y_n>$ pairs of data

$x_k=a_k^{(1)}$… input feature/unit k for observation n in input layer 1

$x_0=a_0^{(1)}$… bias unit in input layer 1 (similar to constant in standard regression techniques)

$k=0,1,...,K$…number of input units/features in input layer 1 (i.e. $K=3$)

$a_h^{(2)}$…activation unit h in layer *2* (by activation we mean the value which is computed and output by unit h in layer *2*)

$a_0^{(2)}$… bias unit in hidden layer 2 (similar to a constant in standard regression techniques)

$h=0,1...,H$…number of hidden units in hidden layer 2 *(i.e. $H=3$)*

$h_1=\hat{y}_1$…predicted output unit (in our classification example there will be only one output unit[16]). Do not mix this with h which is defined above.

$w^{(l)}$… these are the weights / parameters that control the function mapping from layer l to layer $l+1$[17]. l is the layer we are moving from

$l=1,...,L$ … layer (i.e. $L=3$)

$w_{hk}^{(1)}$… weight that controls the function mapping from unit k in layer 1 to unit h in layer 2 (k is the origin neuron and h is the destination neuron in the hidden layer)

$w_{32}^{(1)}$ … weight that maps from input unit 2 in layer 1 to activation unit 3 in layer 2 ($x_2 \rightarrow a_3^{(2)}$)

$w_{1h}^{(2)}$… weight that controls the function mapping from unit h in layer 2 to unit *1* in layer 3 (h is the origin neuron and *1* is the destination neuron in the output layer)

[16] In multi-classification problems there would be multiple output units.

[17] If ANN has n_l units in layer l and n_{l+1} units in layer $l+1$, then $w^{(l)}$ has dimensions $[n_{l+1} \times n_l + 1]$ (we add +1 because of an additional bias unit in layer l). In our example $w^{(1)}$ is a matrix of dimensions $[3 \times 4]$ because it is linking 4 input units in layer 1 to 3 activation units in layer 2. The $w^{(2)}$ is a row vector of length 4 mapping 4 activation units in layer 2 with 1 output unit in layer 3. Note that every input/activation goes to every neuron in the following layer.

n_l… number of units in layer l (H in layer 2, K in layer 1 or I in layer 3)

Note that the unit (input, hidden or output units) can also be referred to as node/neuron. We use the terms "unit" and "neuron" throughout this text. As can be seen from Figure 1 and the discussion below, the main characteristics of ANN are (Lantz, 2013, pp. 208–217): an activation/link function, a network topology and the training algorithm.

An **activation/link function** transforms a neuron's input signal into a single output signal. The main difference between different activation functions[18] are the output values. In our empirical part we use a sigmoid/logistic activation function. The output for sigmoid is in the range [0, 1] (for input values below -5 or above 5 the output is transformed to 0 or 1). Because of this feature it is often referred to as the squashing function. Hinton (2012) notes that logistic neurons give real-valued output that is a smooth and bounded function of the total input. The logistic function is differentiable, which is a nice feature as will be shown in the next paragraphs. Examples of sigmoid, linear and threshold activation functions are shown in Figure 2.

Network topology, which describes the number of neurons and layers (and the ways how they are connected). The main questions here are: the selection of number of layers, number of units within each layer[19] and whether the information in the network is allowed to travel backward. In general, by using a more complex topology, one can identify more complex patterns. The topology is usually kept constant, but it is also possible to define it dynamically. We use feedforward multilayer networks that are connected (every neuron is connected to the next layer). Feedforward[20] means that the input signal is fed continuously in one direction only. The larger number of hidden layers and the number of units inside, in general increase the predictive capability and the computational complexity of an ANN. If there is more than one hidden layer, we call the ANN a deep neural network. However, adding more layers might also cause an overfitting problem, hence calibration is needed.

The **training algorithm** specifies how connection weights are set. Lantz (2013, p. 216) describes that in the forward phase, one first sets the initial weights (randomly). Then the neurons and activations are calculated in sequence from the input layer to the output layer. In the last layer the output signal (prediction) is computed. For the backward phase we use the backpropagation algorithm[21]. The network's output signal (h_I) resulting from the forward phase

[18] Examples of other activation functions are: linear, threshold, hyperbolic tangent, saturated linear, rectified linear, Gaussian, stochastic binary neurons, etc. (Hinton, 2012; Lantz, 2013).

[19] The number of input units is determined by the number of features. The number of units in the hidden layer is left to the user to decide, before training the ANN. Lantz (2013, p. 214) warns that the optimal number of neurons in the hidden layer depends on many characteristics, i.e. on the number of input neurons, the amount of training data, the amount of noisy data, and the complexity of the learning problem. Hence it needs to be calibrated and tested well.

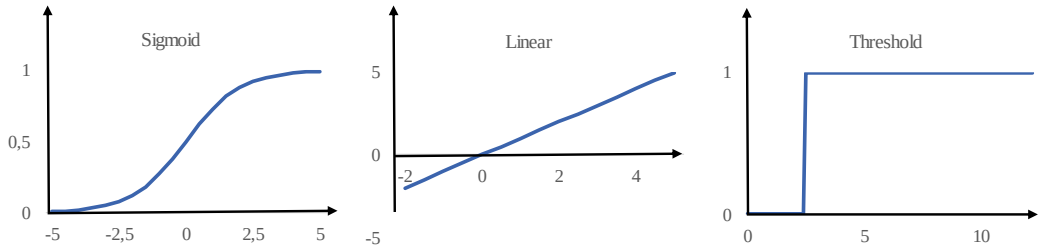
[20] In contrast, a recurrent network allows signals to travel in both directions using loops. This could be used for instance for stock market prediction. However these ANNs are much more complicated and rarely used in practice as noted by Lantz (2013, p. 214). Hinton (2012) describes recurrent networks as networks that have directed cycles in their connection graphs. Special cases of recurrent nets are Hopfield nets, Boltzmann machines etc.

[21] Angelini et al. (2008, p. 743) mentions alternatives such as genetic algorithms and simulated annealing.

is compared to the true value for this observation (y_1). The difference between them is the error that is propagated backwards in the network. It is used to modify the connection weights between neurons and reduce future errors. The backpropagation uses the derivative of each neuron's activation function to identify the gradient in the direction of finding minimum error. The gradient suggests how steeply the error will be reduced, if a weight changes. The algorithm will change the weights that result in the greatest reduction in the error. The steepness is moderated by setting the learning rate parameter. The higher the learning rate, the faster the algorithm will attempt to descend down the gradients (this reduces training time but increases the risk of overshooting). In the following paragraphs we will describe both phases in more detail.

Lantz (2013, p. 211) advises to transform variables prior to training them with ANN. He would transform inputs such that the feature values fall within a small range around 0 (i.e. by standardizing or normalizing the features). This would prevent large valued features to dominate small valued (because of the squashing effect of logistic activation function). Additionally the algorithm would be faster to train. On the other hand there are some articles that claim the contrary, for instance Zhang et al. (1999, p. 25) use raw data and obtain better results.

Figure 2. Activation functions for ANN (sigmoid, linear and threshold)



Source: G. Hinton, *Neural Networks for Machine Learning*, 2012.

For non-normally distributed features it is advised to standardize to a 0–1 range with equation:

$$\frac{x\text{-}min(x)}{max(x)\text{-}min(x)} \tag{5}$$

Angelini et al. (2008, p. 749) do not use the Min-Max formula because there are many outliers in the data and lots of information is lost. They use logarithmic formula instead: ($\hat{x} = log_m (x+1)$. They set *m* near to the actual maximum of the corresponding variable. Jagrič et al. (2011, p. 391) scaled all the variables between -1 and 1 by using a similar transformation.

Tucker (1996, p. 4) proposes splitting the ratios and allowing ANN to determine its own internal representations. Another possibility is to split positive and negative values of variables as these may have completely different effects on the outcome.

Lantz (2013, p. 216) and Hayden and Porath (in Engelmann & Rauhmeier, 2011, p. 9) discuss the advantages and disadvantages of the ANN algorithm. The **advantages** of ANN are that it is one of the most accurate modelling approaches, it models highly complex, nonlinear relationships very well. It is also free from any distributional assumptions about the underlying relationships of the data (it is a non-parametric method). The **disadvantages** are that it is computationally intensive, hard to train (calibration procedure), easy to overfit or underfit and it is a black-box algorithm which means, it is not easily interpretable. Also ANN is hard to calibrate since there is no formal procedure to determine the optimum network topology for a specific problem, i.e. the number of the layers or number of neurons. Hidden units are the reason why ANN is so strong at modelling nonlinearities.

Hayden and Porath (in Engelmann & Rauhmeier, 2011, p. 9) note that ANNs can be used for all rating segments. ANNs require a substantially larger quantity of data in the development stage as other statistical models. Instead of being constrained by the original input features (as in logistic regression), a neural network can learn its own features (neural network with no hidden layers is the same as logistic regression) (Ng, 2015).

Now we present the forward propagation and backpropagation procedures of the ANN algorithm.

a) Forward propagation

For the network in Figure 1 we now describe how to calculate all neurons in a **forward propagation step**. This subchapter is based mainly on the course of Ng (2015). The calculations for N observations are done iteratively for one observation at a time. First (for observation n) we calculate each of the layer 2 activations based on the input values x_{hk} (with added bias term).

$$\begin{aligned} a_1^{(2)} &= \sigma(z_1^{(2)}) \\ a_2^{(2)} &= \sigma(z_2^{(2)}) \\ a_3^{(2)} &= \sigma(z_3^{(2)}) \end{aligned} \tag{6}$$

Here σ corresponds to the logistic function, which is applied over the linear combination of inputs for each unit h in layer 2 ($z_h^{(2)}$). The $z_h^{(2)}$ is the weighted sum of inputs from layer 1.

$$\begin{aligned} z_1^{(2)} &= w_{10}^{(1)} x_o + w_{11}^{(1)} x_1 + w_{12}^{(1)} x_2 + w_{13}^{(1)} x_3 \\ z_2^{(2)} &= w_{20}^{(1)} x_o + w_{21}^{(1)} x_1 + w_{22}^{(1)} x_2 + w_{23}^{(1)} x_3 \\ z_3^{(2)} &= w_{30}^{(1)} x_o + w_{31}^{(1)} x_1 + w_{32}^{(1)} x_2 + w_{33}^{(1)} x_3 \end{aligned} \tag{7}$$

We then calculate the final prediction (this is the single neuron in layer 3) similarly as above. The only difference is that now the input is not the input values (x_{hk}), but the activation values from the preceding layer ($a_h^{(2)}$).

$$h_1(x)=a_1^{(3)}=\sigma(z_1^{(3)}) \tag{8}$$

Again the sigmoid function is applied over the weighted sum of inputs (in this layer the inputs are activation values).

$$z_1^{(3)}=w_{10}^{(2)}a_0^{(2)}+w_{11}^{(2)}a_1^{(2)}+w_{12}^{(2)}a_2^{(2)}+w_{13}^{(2)}a_3^{(2)} \tag{9}$$

$z_1^{(3)}$… the weighted sum of activations from layer 2. In vector notation we have:

$$\begin{aligned} z^{(2)}&=w^{(1)}x \\ a^{(2)}&=\sigma(z^{(2)}) \\ z^{(3)}&=w^{(2)}a^{(2)} \\ h_1&=\sigma\left(z^{(3)}\right)=\sigma(w^{(2)}a^{(2)}) \end{aligned} \tag{10}$$

Where $z^{(2)}$ and $a^{(2)}$ are 4×1 vectors, and $\sigma(.)$ applies the logistic function element-wise to each member of the $z^{(2)}$ vector.

In summary, to implement the forward propagation on real data, one starts with one observation from N data pairs, which are given by the problem $<x_n,y_n>$. In the first run of the algorithm one sets small initial values for the weights. With forward propagation one pushes the input through the network and computes the activation of each layer (sequentially) and the predicted value h_1. Then we continue with the backpropagation step.

b) Backpropagation

As Ng (2015) describes, in the **backpropagation step** one first compares the predicted value h_1 (resulting from the forward propagation) with the observed value y_1 for given observation. Then the errors are computed and propagated backwards through the net (until the input layer is reached, where there is no error in this layer, since the activation equals the input). These error measurements for each observation are used to calculate the partial derivatives, which are needed to minimize the cost function by updating the initial weights. The main goal of backpropagation is to obtain the gradient of total error with respect to all weights which are used in the optimisation algorithm. The larger the errors, the more the weights change in the next iteration of the algorithm. After changing the weights (in the backpropagation algorithm we try to change them optimally to find the optimum) one reruns the forward propagation and calculates new predictions. These steps are then repeated several times until the stopping criterion is reached.

We now discuss how to optimally calculate new weights with the backpropagation algorithm. This subchapter is based mainly on the course of Ng (2015), whereas the derivations are based also on Mitchell (1997) and Hinton (2012).

Note again that we focus here on one single observation *n*. The algorithm will loop over all observations, which we describe later. First, one needs to calculate the partial derivative terms: $\frac{\partial E_1}{\partial w_{ij}^{(l)}}$. Here *i* and *j* refer to different units regardless in which layer they are. We want to calculate the partial derivatives with respect to single parameters. Partial derivatives are real numbers (not a vector or matrix). Our goal in backpropagation is to change weights before running the next loop of forward propagation. If we do this smart, the errors should be smaller in the next round.

$$w_{ij} \leftarrow w_{ij} + \Delta w_{ij} \tag{11}$$

where:

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}}$$

Note that the weights will not change after looping through each observation, but rather after the loop through all observations is done (this depends on the implementation of the algorithm[22]). To obtain the true gradient of error, one would sum up the delta values (and before that multiply by activations, as we will see in the next paragraphs) before altering weight values.

Where η is the learning rate. Mitchell (1997, pp. 88, 92, 98) notes that the learning rate is usually set to some small value (e.g. 0,1 or 0,05). To prevent overshooting the optimum, it is advisable to gradually reduce the value of η, as the number of gradient descent steps grows. *E* is the error function and y_1 is the true outcome value for observation *n*.

We need to calculate Δw_{ij} for all layers. The equation for the output layer differs from the equation of hidden layers, hence we derive them separately. First we start with the **output layer**: In the backpropagation the weight that leads to the output layer will be affected by the error in this last output layer. Our job is to define the cost (error) function to apply to the last layer of our neural network. Errors are then backpropagated through the network.

We first define the squared error cost function, then we define also the alternative logistic error. It is defined as the squared error term. Note again that we are deriving everything for an individual observation *n* (later everything will be looped through all observations and accumulated).

$$E(w) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{2} \left(h_1(x^{(n)}) - y^{(n)} \right)^2 \tag{12}$$

We prefer to use the logistic representation which is well suited to our classification problem. We can define the error function for logistic regression as (Ng, 2015):

[22] In the case of individual updating, we would have stochastic approximation to gradient descent approach, which is used in "big-data" problems. There is tradeoff between the two approaches of prediction accuracy and computational efficiency.

$$E(w)=\frac{1}{N}\sum_{n=1}^{N} Cost\left(h_1(x^{(n)}),y^{(n)}\right) \tag{13}$$

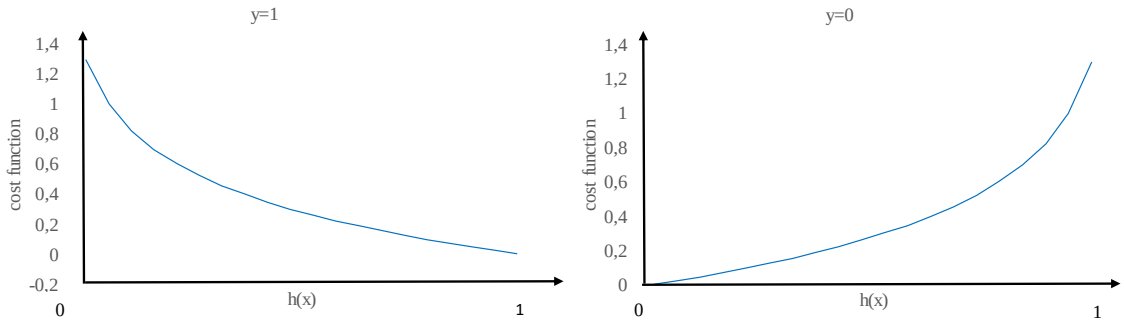
$$Cost(h_1(x),y))=\begin{cases} -\log(h_1(x)) & if\ y=1 \\ -\log(1-h_1(x)) & if\ y=0 \end{cases}$$

We plot the function in Figure 4. When we are right, the cost function is 0. Else the cost function slowly increases (as we become more wrong). On the x-axis is what we predict, and on the y-axis is the cost associated with that prediction. This error function is convex. We can write the error for logistic regression more compactly in a single line (Ng, 2015).

$$E(w)=-\frac{1}{N}\sum_{n=1}^{N}\left[y^{(n)}\log\left(h_1(x^{(n)})\right)+(1-y^{(n)})\log\left(1-h_1(x^{(n)})\right)\right]+\underbrace{\frac{\lambda}{2N}\sum_{k=1}^{K} w_k^2}_{regularization\ term} \tag{14}$$

The regularization term is not necessary, but it can help to solve the problem of overfitting (high variance), which can be caused by having too many features that lead to a cost function of exactly zero. Having too many features, the algorithm tries too hard to fit the training set, but fails to generalize (apply to new observations). To deal with this, we can reduce the number of features, or use regularization (keep all features, but reduce the magnitude of weights w). Regularization works well when we have a lot of features, which are all good predictors. λ is the regularization parameter that controls the tradeoff between fitting the estimation sample while keeping the weights small. If λ is very large we penalize all the weights, so all of them end up being close to zero (the result is underfitting). On the other hand with a small λ we do not solve the problem of overfitting. Hence it is important to choose λ carefully (Ng, 2015).

Figure 3. The cross entropy error function



Note the similarity of the general form with the entropy function $(-\sum_i p_i \log(p_i))$. Our logistic error function is sometimes referred to as cross entropy or deviance (Mitchell, 1997, p. 170). Another reason why we use this cross entropy error function is that this function can be derived from statistics using the principle of maximum likelihood estimation (see chapter on logistic regression and Löffler and Posch (2007) for derivation). However, Hastie, Tibshirani, & Friedman (2008, p. 395) note that for classification problems one can use either, squared error or this latter cost function.

We can then find the parameters w to minimize E with the use of gradient descent (or any other advanced optimisation, i.e. conjugate gradient, BFGS, etc.). The error function for ANN is similar, but it differs in the regularization term:

$$E(w)=-\frac{1}{N}\sum_{n=1}^{N}\left[y^{(n)}\log\left(h_1(x^{(n)})\right)+(1-y^{(n)})\log\left(1-h_1(x^{(n)})\right)\right]\underbrace{+\frac{\lambda}{2N}\sum_{l=1}^{L-1}\sum_{i=1}^{n_l}\sum_{j=1}^{n_{l+1}}(w_{ji}^{(l)})^2}_{\text{regularization term}} \tag{15}$$

We do not sum over the bias terms (starting with 1 for the summation). We sum for each observation $n=1,\ldots,N$ in the first term. The second term is referred to as weight decay and is a massive regularization summation term. In contrast to the logistic regression, here there must be sums over weights in different layers (Ng, 2015).

As we will see later the algorithm cycles over each observation, so the error for one observation without the regularization term can be written as:

$$E_1(\vec{w})=-(y_1\log(h_1)+(1-(y_1))\log(1-h_1)) \tag{16}$$

Our goal is to get the derivative of error with respect to all weights $\nabla E(\vec{w})\equiv\left[\frac{\partial E}{\partial w_0},\frac{\partial E}{\partial w_1},\ldots\right]$. The negative of this gradient[23] gives us the direction of the steepest descent along the error surface (for intuition see Figure 4). Note that we have weights in two layers, hence we need to solve our problem in two steps. Now we have the error function, how do we minimize it? Following the idea of changing weights we first need (Mitchell, 1997, p. 91):

$$\Delta w_{1h}^{(2)}=-\eta\frac{\partial E_1}{\partial w_{1h}^{(2)}} \tag{17}$$

This equation tells us that the steepest descent is achieved by altering each weight in proportion to the partial derivative. In the next paragraphs we present equations for the cross-entropy based error. To get it, we need to first derive (for derivation see the Appendix B-a):

$$\frac{\partial E_1}{\partial w_{1h}^{(2)}}=\frac{\partial E_1}{\partial z_1^{(3)}}\frac{\partial z_1^{(3)}}{\partial w_{1h}^{(2)}}=\ldots=a_h^{(2)}\delta_1^{(3)}=a_h^{(2)}(y_1-h_1) \tag{18}$$

Where $\delta_1^{(3)}$ is the error term of the 1 unit in the output layer: $\delta_1^{(3)}=y_1-h_1$. Then we could calculate the change for particular weight that maps from unit h in layer 2 to unit 1 in layer 3:

$$\Delta w_{1h}^{(2)}=-\eta\frac{\partial E_1}{\partial w_{1h}^{(2)}}=-\eta a_h^{(2)}(y_1-h_1) \tag{19}$$

[23] The gradient $\nabla E(\vec{w})$ is a vector, whose components are the partial derivatives of E with respect to each of the weights (Mitchell, 1997, p. 91).

Now we continue with the **hidden layer** similarly as above:

$$\Delta w_{hk}^{(1)} = -\eta \frac{\partial E_1}{\partial w_{hk}^{(1)}} \tag{20}$$

For that we need to derive this (for derivation see the Appendix B-b):

$$\frac{\partial E_1}{\partial w_{hk}^{(1)}} = \frac{\partial E_1}{\partial z_h^{(2)}} \frac{\partial z_h^{(2)}}{\partial w_{hk}^{(1)}} = \ldots = x_k^{(1)} \delta_h^{(2)} = x_k^{(1)} w_{hk}^{(2)} \left(y_1 - h_1\right) \left(a_h^{(2)} \left(1 - a_h^{(2)}\right)\right) \tag{21}$$

Where $\delta_h^{(2)}$ is the error term of hidden unit h in layer 2. Since observations provide target values y_1 only for network outputs, no target values are directly available to indicate the error of hidden units' values as noted by Mitchell (1997, p. 99). Thus the $\delta_h^{(2)}$ is a weighted sum[24] of the next layers delta values, weighted by the parameter associated with the links:

$$\delta_h^{(2)} = \left(w_{hk}^{(2)}\right)^T \delta^{(3)} \left(a_h^{(2)} \left(1 - a_h^{(2)}\right)\right) = w_{hk}^{(2)} \left(y_1 - h_1\right) \left(a_h^{(2)} \left(1 - a_h^{(2)}\right)\right) \tag{22}$$

These errors are calculated for each hidden unit. In vector notation (with dimensions) this is:

$$\frac{\partial E_1}{\partial w^{(1)}} = x^{(1)} \delta^{(2)} = x^{(1)} \left(w^{(2)}\right)^T \left(y_1 - h_1\right) * \left(a^{(2)} * \left(1 - a^{(2)}\right)\right) \tag{23}$$

$$\underbrace{\delta^{(2)}}_{[4\times 1]} = \underbrace{\left(w^{(2)}\right)^T}_{[4\times 1]} \underbrace{\left(y_1 - h_1\right)}_{[1\times 1]} * \underbrace{\left(a^{(2)} * \left(1 - a^{(2)}\right)\right)}_{[4\times 1]}$$

Where * is the element wise multiplication between two vectors (Ng, 2015). By doing backpropagation and computing all the partial derivative terms we can compute the gradient $\nabla E(\vec{w})$ which is needed for the optimisation.

c) Summary of the backpropagation

To summarize, the algorithm is as follows (Ng, 2015). First we define the network architecture. We need to choose (calibrate) the number of input units (features), number of output units (classes) and the number of hidden units.

For a training set of N observations we have pairs $\{(x_1, y_1), \ldots, (x_n, y_n), \ldots, (x_N, y_N)\}$. First we set the small initial weights[25] and zero delta values:

[24] Weights characterize the degree to which the hidden unit h is "responsible for" the error in output unit 1 (Mitchell, 1997, p. 99).

[25] Ng (2015) recommends picking random small initial values of weights, i.e. values between 0 and 1. One effective strategy in his opinion is to choose values uniformly in the interval $[-\epsilon_{init}, \epsilon_{init}]$ where a good choice is $\epsilon_{init} = \frac{\sqrt{6}}{\sqrt{n_l + n_{l+1}}}$ where n_l and n_{l+1} are the number of units in the layers adjacent to $w^{(l)}$. Mitchell (1997, p. 98) proposes

$$\Delta_{ji}^{(l)} = 0 \text{ for all } i,j,l \tag{24}$$

These values will be used to compute the partial derivatives (as accumulators for computing them). Next, we loop through the training set (each observation pair $<x,y>$):

$$For\ n=1\ to\ N$$

For each observation we run the forward propagation, calculate the error and then backpropagate it back to the input layer. The computed gradient for this observation is then used to update all weights in the network. This gradient descent step is then iterated (often thousands of times, using the same observations multiple times) until the network performs acceptably well (Mitchell, 1997, p. 99).

We first perform forward propagation to compute $a^{(2)}$ for the hidden layer 2 and $a^{(3)}=h_1$ for output layer 3 for every observation n. Then we use the output value y_1 for each observation and compare it with the predicted value h_1. Next, we calculate the delta error for the output layer $\delta^{(3)}$. By using backpropagation we move back through the network from layer l down to layer $l-1$ etc. and calculate the delta error for a hidden layer $\delta^{(2)}$. Finally we use Δ to accumulate the partial derivative terms for all layers.

$$\Delta_{ji}^{(l)} = \Delta_{ji}^{(l)} + a_i^{(l)} \delta_j^{(l+1)} \tag{25}$$

Where l is layer, i is neuron in that layer and δ_j is the error of the affected neuron in the target layer. In vectorised form we have:

$$\Delta^{(l)} = \Delta^{(l)} + \delta^{(l+1)} (a^{(l)})^T \tag{26}$$

Finally, after executing the body of the loop we exit the loop and compute[26]:

$$D_{ji}^{(l)} = \frac{1}{N} \Delta_{ji}^{(l)} \tag{27}$$

for instance values between -0,05 and 0,05. Hastie et al. (2008, p. 400) note that with normalized variables it is typical to use random uniform weights over the range -0,7 and 0,7. Angelini et al. (2008, p. 750) choose initial weights randomly in the interval range from -1 to 1.

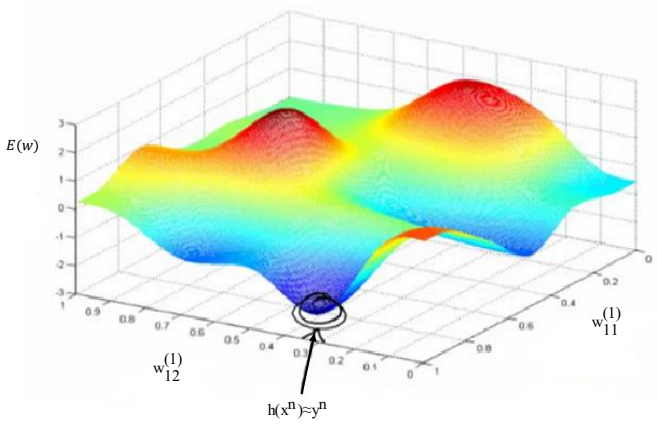
[26] Ng (2015) notes that if we add a regularization term the equation would be (when i=0 we have no regularization term):

$$\mathrm{D}_{ji}^{(l)} = \frac{1}{N} \Delta_{ji}^{(l)} + \lambda \mathrm{w}_{ji}^{(l)} \quad if\ i \neq 0$$
$$\mathrm{D}_{ji}^{(l)} = \frac{1}{N} \Delta_{ji}^{(l)} \qquad if\ i = 0$$

Now we have calculated all the D terms (each D is a real number and not a vector or matrix) $\frac{\partial E}{\partial w_{ji}^{(l)}} = D_{ji}^{(l)}$. These are the partial derivatives for all weights respectively. We can then use all these partial derivatives in the gradient descent, or one of the advanced optimisation algorithms instead. This means that we update weights (with the steepest descent) and then iterate again from (b) until we obtain optimal weights (smallest error). This would be a so called weight update loop which might be iterated thousands of times in a typical application, as noted by Mitchell (1997, p. 99). There may be different termination conditions to halt the procedure (i.e. a fixed number of iterations, the error falling below some threshold, etc.). The choice of the termination condition is important, because too few iterations can fail to reduce the error sufficiently, on the other hand too many can lead to overfitting issues.

By doing this we try to minimize E(x) as a function of weights. Ng (2015) and Mitchell (1997, pp. 92, 104) note that the problem might be susceptible to getting stuck in a local minimum, however in practice this is not a huge problem, as the ANN should find a good local optimum at least (see Figure 4)[27]. To remedy one can use optimisation algorithms that are more appropriate for searching global optimums, another possibility is to start with different starting values or to use a momentum modification of the backpropagation[28]. Hinton (2012) notes that the convergence of weights to their correct values is slower, if we have highly correlated inputs.

Figure 4. Graphical presentation of the error function for ANN



Source: A. Ng, *Machine Learning*, 2015.

The vertical axis in Figure 4 indicates the value of the error function. A gradient descent search determines a weight vector that minimizes the error by starting with an arbitrary initial weight

[27] Mitchell (1997, p. 104) discusses the advantages of stochastic approximation to gradient descent over the original one in case of a local optima. The other possibility is to rerun the ANN with different initial weights.

[28] The idea is to alter the weight update rule by making the weight update on the n-th iteration depend partially on the update from the previous iteration (n-1): $\Delta w_{ji}(n) = \eta \delta_j x_{ji} + \alpha \Delta w_{ji}(n-1)$ where $0 \leq \alpha < 1$ is a constant called the momentum. The first part of the equation is the same as before. The effect of α is to prevent the algorithm to stop in a flat area of local optimum (Mitchell, 1997, p. 100).

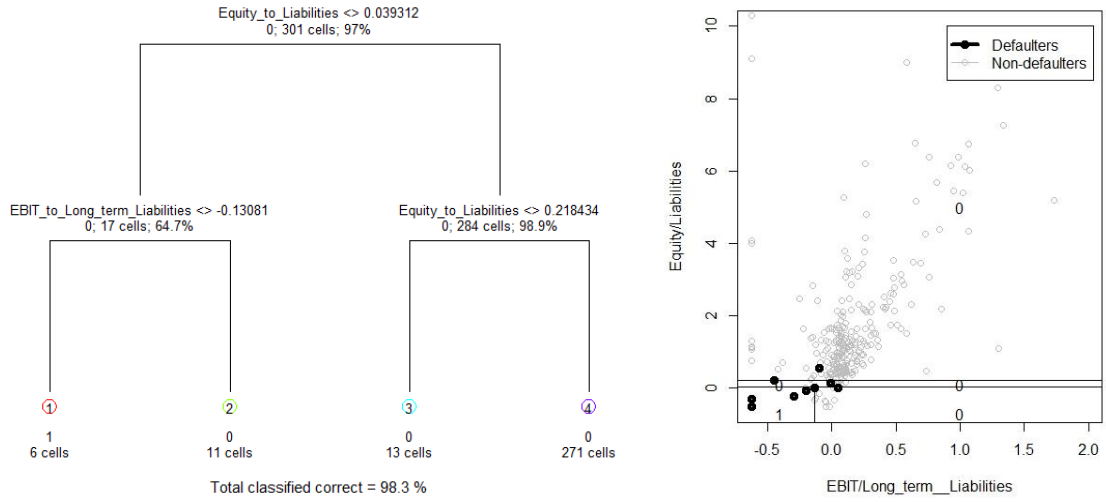
vector. Then it modifies it repeatedly in small steps. At each step, the weight vector is changed in the direction that produces the steepest descent along the error surface (in a direction of the minimum) (Mitchell, 1997, p. 91).

ANN requires a lot of input data (with more data it only gets better). However, this slows down the algorithm. For an introduction to large scale machine learning and ANN approaches see Ch. 17 in Ng (2015)[29]. New advances allow us to use modified algorithms that update the parameters in each step through data, instead at the end of each loop through all the data. These algorithms allow us to use several computers (processor power and memory) in parallel (i.e. map reduce platforms, for instance Hadoop).

1.7 Classification trees and random forests (RF)

Because the random forests method is built on classification trees we discuss them first. Classification trees[30] are not dependent on a functional form or multicollinearity between explanatory variables. They are a non-parametric method, hence they need no distributional assumptions. There are several algorithms for building decision trees. The most popular for binary problems are Classification trees popularized by professor Breiman in year 1984. **Classification trees** are based on splitting rules, which allow us to segment the variable space into smaller regions. The advantage of classification trees is its intuitive and simple interpretation (Ortl & Gorenjec, 2015, p. 23).

Figure 5. Classification tree logic



Source: A. Ortl & R. Gorenjec, *Corporate default prediction with the Random Forest algorithm*, 2015, p. 23.

[29] See Mitchell (1997, pp. 93, 101) for stochastic gradient descent. There, the weights are updated for each observation and not for all together.

[30] Lessmann et al. (2015, p. 17) distinguish between classification regression trees (CART) that are based on the Gini-coefficient and the J4.8 algorithm that is a decision tree based on entropy measure (we discuss these two measures later).

Splitting can be represented graphically as in Figure 5. The example is taken from Ortl and Gorenjec (2015, pp. 23–24). The tree has four terminal nodes. Terminal node 2 tells us that a company with an equity/liabilities ratio of less than 0,039312 and EBIT/long-term liabilities with values of more than -0,13081 will be classified as a non-defaulter (in the sample we had 11 companies with such numbers that have not defaulted). On the right side of Figure 5 one can see its graphical representation. The splitting rules split the space in four regions, consistent with the tree structure. The big black dots in the bottom-left corner represent defaults, and the lines are the splitting rules.

In the following paragraphs we present the classification tree algorithm. The derivation is based on Ortl and Gorenjec (2015, p. 24), Hastie, Tibshirani, James, & Witten (2013, pp. 303–316) and Kastrin (2015, pp. 82–86).

Define the training data as pairs (x_i,y_i), $i=1,\ldots,N$, where x_i is a vector with $k=1,\ldots,K$ explanatory variables (attributes) and y_i is the dependent variable that is in function with the explanatory variables $y_i=f(x_i)$. Define M regions as $R_1,\ldots,R_M$. Each region represents one terminal node. As a splitting criteria or impurity measure we use the Gini index[31].

$$G=\sum_{c=1}^{C=2}\hat{p}_{mc}(1-\hat{p}_{mc}) \tag{28}$$

Note that $\hat{p}_{mc}=\frac{1}{n_m}\sum_{x_i\in R_m} I(y_i=c)$ and m is the node that represents the region R_m that consists of n_m observations (cases). The $\hat{p}_{mc}$ is defined as the share of observations that are in the class c (classes are 0 or 1 in our problem) in the region R_m. The Gini index has a low value for the cases where $\hat{p}_{mc}$ is close to 0 or 1. Then we say that the region is "pure" and mostly consists of observations that correspond to the same class c of the dependent variable. The observation in the node m is then voted to the more represented class in that region (vote class) $c(m)=\arg\max_c(\hat{p}_{mc})$. The algorithm does not try all combinations but is based on recursive binary splitting. In the first step the splitting point for two regions is defined as a minimization problem for which the objective function is the Gini index (we want a low Gini index for both regions after the split). The optimisation problem yields the variable and the splitting point of that variable at each step. In our case that first splitting point would be 0,039 for the equity/liabilities ratio variable. In the next step we have splitting into three regions etc. The algorithm stops, when it satisfies the stopping criteria (i.e. the minimum number of observations in a region or the number of terminal nodes).

The danger of building a full tree is the problem of overfitting. The predictive power of the tree might be bad due to a high degree of complexity. The standard method to remedy this issue is the pruning method, which is described for instance in Hastie et al. (2013, p. 309) and we do not discuss it here.

[31] Other possible measures for our classification problem would be cross entropy or classification error rate.

The other possibility to avoid overfitting is to use the method of **bagging** which is a basis for the construction of the Random forest method. Bagging is based on bootstrapping T random samples with replacements. These samples become the new learning samples for estimation on which we build T classification trees (without pruning, but to the maximum depth). By this procedure we solve the problem of classification trees being unstable. The method we present is based on Ortl and Gorenjec (2015, p. 24) and Hastie et al. (2013, pp. 316–324).

For building one tree we use approximately 2/3 of cases from the original data, the rest of data is duplicated. The cases that are not bootstrapped to the particular tree are kept out as **out-of-bag** cases and are denoted as OOB_t. They can be used as unbiased estimate of the classification error $errOOB_t$, which is defined as ratio between wrongly classified and all classified OOB_t cases. If we average them, we get the estimate of the classification error for our model (Kosič, 2012).

Hayden and Porath (in Engelmann & Rauhmeier, 2011, p. 10) note that the advantage of classification tree is also that it captures interactions between variables. Whether a specific variable or category becomes relevant depends on the categories of the variables in the upper level/node. However, strong variables always appear in the upper nodes not allowing others to show their importance. Such trees would be highly correlated. This disadvantage is solved by applying the other type of randomness in the random forest algorithm.

The **random forest** builds on the idea of bagging, however, randomness is added also in the selection of candidate variables. Now we randomly select m variables (out of K variables) for each subtree. By this we eliminate the influence of strong variables to be selected in every model. Lowering the values of selected variables into each subtree m decorrelates the trees but also lowers the predictive power of these trees, hence we need to calibrate for the optimal value (Kastrin, 2015; Kosič, 2012).

It is important to note that a random forest requires calibration of different parameters (T, m and the number of observations left in the final node). Originally the last parameter was set to 1. Kosič (2012) suggests using 1000–5000 trees. A large number of trees is needed to stabilize the classification error. As number of variables authors usually try $m \approx \sqrt{K}$. However, these parameters should be calibrated to obtain best results. Hayden and Porath (in Engelmann & Rauhmeier, 2011, p. 10) present the advantages and disadvantages of trees:

- easy modelling of nonlinear relationships,
- identifying interactions between variables (parametric methods can model interactions only to a limited extent by introducing dummy variables),
- as with neural networks, decision trees are free from distributional assumptions,
- the output is easy to understand and interpret (not so much for random forests),
- probabilities of default have to be calculated in a separate step, as the output is binary (0,1) and not a continuous score variable,

- decision trees only calculate default probabilities for the final node in a tree and not for individual borrowers,
- the stability of the model cannot be assessed with statistical procedures - we work with k-fold validation.

2 DATA PREPARATION

2.1 Explanatory variables

From a regulatory point of view the bank must integrate all available information. Article 174 of the CRR directive (EC, 2013) requires a bank to use data in such a way that it forms a representation of the population of the institution's actual borrowers. The sources of the data may be internal or external. EBA (2013b, p. 72) reports that most banks use internal data. It is possible that an external rating (given by a rating agency) is also a basis to form an internal rating. A **catalogue of variables** can be very lengthy. Chen and Shimerda (1981, in Falkenstein et al. 2000, p. 27) provide a list of over hundred variables that were cited in the financial distress literature. That is more than anyone has time to analyse systematically. Hence the **variable selection** problem is one of the main problems of financial ratios. We try to cope with this with an automated approach later in this thesis. Falkenstein et al. (2000, p. 28) define good variables, as those that have a conditionally monotone relationship with default probabilities. This means that the relationship is always increasing or always decreasing. However, there are some variables that are good predictors and do not possess this property (i.e. sales growth).

Quantitative variables can provide aggregate (macroeconomic or industry level variables) or borrower-specific (usually financial ratios or payment/default history, age) information. BIS (2005, p. 12) warns that combining aggregate and borrower-specific variables can be dangerous. Macroeconomic variables and financial ratios are often highly correlated (most corporate borrowers will experience higher revenues when gross domestic product (hereinafter: GDP) growth is high and lower revenues when GDP growth slows). Using both types of information to forecast default may be redundant, but on the other hand macroeconomic variables are often used to produce a distinction between TTC and PIT models, which we will discuss later on (macroeconomic variables are also useful for stress testing). BIS (2005, p. 12) defines both approaches (with or without aggregate variables) as valid.

Financial ratios are the most commonly used proxy for creditworthiness. Samuels (2012, p. 14) criticises that too much emphasis is given to financial ratios. In his opinion one should make at least use of industry sector information (such as a company's competitiveness within that sector) and management quality (especially in case of micro companies). These other risk factors have a complicated relationship to financial risk. For instance, a company with a very stable industry risk profile might have a relatively weak financial profile. Later on we will complement our model with the idea of overrides, which is one alternative for incorporating such risks. In the modelling part we will use mainly **financial ratios**. For a discussion on the most important

groups of financial ratios see Balthazar (2006, pp. 156–168) and Falkenstein et al. (2000, pp. 32–40).

On the other hand, **qualitative variables** may be much harder to collect or treat within a model. Because they can improve the model substantially, it is important to take them into consideration as well. Usually a model consists of a few qualitative variables that are easy to treat within a model as dummy variables (such as geographic location or industry sector[32]). For including some other variables from questionnaires (i.e. related to management quality) banks often combine the statistical model with some expert system.

A very important step after constructing the database is to organize a first meeting with credit analysts to define what the possible candidates as explanatory variables are. They should feel included at the earliest stage of the project. ONB (2004, p. 61) notes that it is important to provide representativeness of the data (for the underlying segment), data quantity (for statistical significance) and data quality (garbage in – garbage out). Irrespective of the data source employed, Basel standards (BIS, 2001, p. 47) require PD estimates to be developed using a minimum historical observation period of at least 5 years[33]. EBA (2013b) reports that different length of time series (for estimation or for calibration) bring huge discrepancies between banks. EBA (2015, p. 57) notes that the lengthening of time series may not always be a perfect solution due to structural changes in the market and changes in the bank's policies. Tata (2015, p. 3) notes that there is no such specific definition of the observation period for IFRS 9 impairment calculation needs (see ONB, 2004, pp. 65–72).

2.2 Defaults (dependent variable)

Basel rules **define the occurrence of default**, when one or more of certain events have taken place (EC, 2013, article 178 of CRR directive; BIS, 2001, p. 30): (a) the borrower is unlikely to pay its debt obligations in full[34] or (b) a credit loss was associated with any obligation of the borrower (i.e. charge-off, specific provision, distressed restructuring involving forgiveness or postponement of principal, interest for fees) or (c) the borrower is past due more than 90 days on any credit obligation[35] or (d) the borrower has filed for bankruptcy or similar protection from creditors.

[32] This leads to the danger of integrating specific temporary situations. For instance, imagine that the utilities sector had gone through a heavy deregulation phase over previous years or is now losing implicit state support? This could be a dangerous situation, if credit analysts did not launch a warning signal and this is not incorporated in overrides. In such cases people could rely on the model too much (Balthazar, 2006, p. 148).

[33] The EBF (2012, p. 18) study in 2012 has shown that most of the banks participating in their survey were using historical data starting from 2005.

[34] See article 178 of CRR directive (EC, 2013) for the indications of unlikeliness to pay (i.e. bankruptcy procedure, recognition of impairments, non-performing definition, etc.)

[35] EBA (2013a) defines that considering only filing for a bankruptcy would be a "late" definition of default that would lead to a smaller number of defaults (but with a larger loss amount). BIS (2001, p. 31) states that it may be possible to develop a model by using definitions of default, which are not consistent with the reference definition (but should be calibrated to the reference definition). Hayden (2002, p. 97) provides empirical evidence that models estimated on data without 90-days definition of default are comparable to the ones that include these defaults in terms of discriminatory power.

In the EBA (2013a, p. 40) report is stated that the application of 90 days past due seems to be general practice among banks. In practice, two third of the banks start counting the days past due when the first non-payment occurs and one third when a non-payment materiality threshold is reached[36]. EBF (2012, p. 2) states that differences appear due to the fact that some countries use the 180 days past due definition (for instance in France, Italy and in the United Kingdom). Additionally it is important how one treats aggregations. The Basel rules allow firms to specify, if their default is on a "per borrower" or "per product/facility" level. EBA (2013b, p. 84) reports that differences across countries are huge. Differences appear in the calculation of the observed default rate (which is needed for calibration). Usually it is defined as in EBA (2013b, p. 83) or EBA (2015, p. 24): "only new defaults observed during the next 12 months among the borrowers that were non-defaulted at the reference date". A big issue stated by EBA (2015, p. 23) is also how to treat (multiple) defaults that appear several times in a year (separately or only on the first date of default).

Banks also have a different treatment of transfer to the recovery unit. Some firms include internal transfer to recovery as part of the definition of default. Almost 75 % of banks surveyed, return bad firms to the "good book" after they are defined as cured or returned to an "in order" status. The rest assign a waiting period (for around 5 % of banks this period is more than 1 year). The regulator in the United Kingdom defines government support schemes as default, which is not the case in most other countries (EBF, 2012, pp. 12–13).

Such differences place some banks at a competitive disadvantage or advantage with respect to IRB capital requirements or impairment calculations. EBF notes that further guidance from regulators should be given to establish more consistent measures. It is interesting to read that EBF (2012) discusses that some banks are subject to different home and host supervisory criteria. This only shows the complexity of the issue.

Basel requires the calculation of one year default probabilities. An important decision is related to the **time horizon**, which is connected to the definition of the dependent variable (default). Usually one would take the values of explanatory variables for non-defaulters at time *t* and observe the companies in the year after *(t,t+1)*[37]. Defaulted borrowers would be assigned a binary value of 1 and others would be assigned the value 0.

The new IFRS impairment model is based on an earlier recognition of losses. It is based on a "three-stage" model for impairment, based on changes in credit quality since initial recognition. PWC (2014, p. 1) describes that in the first stage one would need to calculate 12-month PDs.

[36] EBA (2013b, p. 83) and EBA (2015, p. 20) recognize that banks differ in the definition of materiality threshold. Some banks use absolute values of around 200 EUR for retail exposures and 500–10.000 EUR for corporate exposures. Some banks define relative as percentage of exposures (usually 2 %).

[37] For purposes of long-term PDs one could extend the period of observing default to several years, taking into account the data availability. Ideally all models correspond to a time period of a average maturity of the credit as noted in Balthazar (2006, p. 147). The drawback is that this lowers discriminatory power of the model as predicting one year probability of default is easier when compared to predicting default in eight years.

For the last two stages[38] one needs lifetime PDs for each year of the loan maturity. However, note that the availability of information decreases, as the forecast horizon increases. The estimates of lifelong PDs are not expected to be detailed for periods that are far in the future. Management may extrapolate projections from available information. The standard however, is not specific on how to extrapolate projections from available information. PWC (2014, p. 13) warns that this is a highly judgmental area that could have a large impact.

2.3 Sample definition

First we need to define the sample that will be used for estimation. It must be based on the segment to which the model predictions will be applied (it must be representative of the true bank's segment). In this thesis we focus on PD estimation for corporates. Hence it is feasible to exclude all banks, financial companies, holdings, sole proprietors, start-up companies and other groups whose balance sheets and P&L statements have a different nature from classical corporate ones. Balthazar (2006, p. 154) notes that when constructing estimation sample one should remove subsidiaries. When several companies belong to the same group in the dataset, we should work only with the top mother companies (if support to subsidiaries is expected in case of distress).

A valid procedure is to divide our sample into an estimation and a validation subsample. The estimation sample is used to estimate and calibrate the model while the backtesting sample helps us to check, whether we have problems with over-fitting and to verify that the model discriminates well. One usually uses two-thirds of the whole sample for the estimation and one-third for the validation sample. The selection can be done randomly (for instance inside each rating grade) for the performing borrowers rating dataset, and separately for the default dataset (Balthazar, 2006, p. 174).

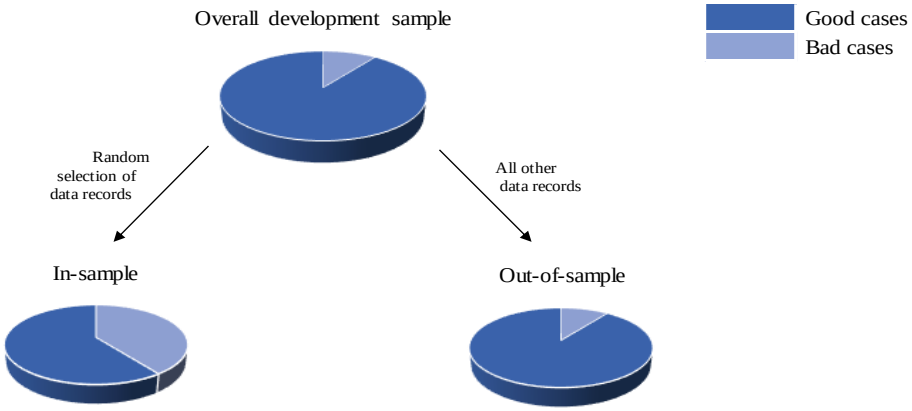
The approach to divide the actual data is preferred. However, in cases where only insufficient data are available ONB (2004, pp. 72–73) proposes the use of a bootstrap (resampling) method. In order to avoid bias due to subjective division, the sample should be split up by random selection as shown in Figure 6. It is also very important that the borrowers included in the estimation sample are not used in the validation sample.

Balthazar (2006, p. 176) alternatively proposes the "leave-one-out" process. There, one uses the whole sample but one company, then records results and repeats by leaving out another company… A popular approach to define samples in terms of validation is the k-fold cross-validation approach. Zhang et al. (1999, p. 17) note that the random sampling into estimation and validation part may introduce bias, in that the characteristics of the validation may be very different from those of the estimation. Refaeilzadeh, Tang, & Liu (2008, pp. 1, 5) describe that the data is first partitioned into k equally sized segments or folds. Then k iterations of training

[38] In the second and third stage there are assets with significant increase in credit risk since initial recognition and the non-performing assets.

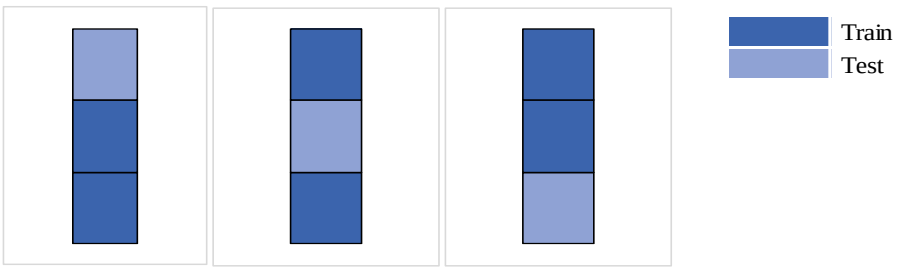
and validation are performed such that within each iteration a different fold of the data is held-out for validation, while the remaining k-1 folds are used for learning. Cross-validation can be applied in the context of performance estimation, model selection, and tuning learning model parameters. Refaeilzadeh et al. (2008, p. 5) discuss that often one wants to estimate the accuracy of a classifier. After applying k-fold cross validation, one averages performance results and obtains a robust measure of their performance. The resulting average accuracy will likely underestimate the true accuracy, when the model is trained on all data and tested on unseen data, but the estimate is supposed to be more reliable. Alternatively cross-validation may be used to compare a pair of learning algorithms or to tune the parameters before training some classifiers, which are parameterised and the parameter tuning can lead to their improved performance. This type of validation is used for instance in Zhang et al. (1999) and in Lessmann et al. (2015). See Lessmann et al. (2015, p. 28) for a combination of the Nx2-fold cross-validation.

Figure 6. Sample construction



Source: ONB, *Rating models and validation*, 2004, p. 73.

Figure 7. Example of a 3-fold cross validation



Source: P. Refaeilzadeh, L. Tang, & H. Liu, *Cross-validation*, 2008, p. 1.

With regard to **weighting good and bad cases in the estimation sample**, two different procedures are proposed by ONB (2004, pp. 72–73). A proportion of bad cases can be representative of the rating segment to be analysed. In this case, the result of logistic regression

can be used directly as a PD without further calibrating or rescaling. This approach is advisable whenever a sufficient number of bad cases can be collected. In practice, banks usually use approximately one fourth to one third of the estimation sample of bad cases. By doing so we maximize the reliability of identifying differences between good and bad borrowers. However, this approach requires the calibration and rescaling of calculated default probabilities. The out-of-sample should be similar to the real bank's data on which the predictions will be used.

Some authors use a matching approach. For instance, Brezigar-Masten and Masten (2012, pp. 4, 16) use a choice-based sampling approach. Their goal is to obtain a more balanced share of defaulted and performing firms in the sample, which is based on size of a firm, industry, year of financial statements and year of default. Then they assign 75 % of observations to the estimation sample and the rest to the out-of-sample. They find out that balancing group shares in the estimation sample in favour of bankrupt firms increases the prediction accuracy of potentially bad risks. However, this is not the case in reality. They note that using choice-based sampling leads to over-rejection of potentially good risks, which is acceptable if the aim is more concentrated on risk limiting than on profit maximization.

2.4 Data cleaning

A very important step before modelling is to be aware of **dangerous traps** when calculating financial ratios. One would expect that low or negative net income and low or negative equity would lead to higher default rates. However, when calculating the profitability ratio "net income/equity", it could happen that we have negative values for the denominator and for the numerator respectively. Hence, it is possible that we get the same ratio for a very bad company (both values negative) and for a good company (both positive). In this case Falkenstein et al. (2000, p. 33) prefer net income/assets ratio. Balthazar (2006, p. 153) proposes forcing the negative equity value to zero.

Another example is equity/debt ratio. In this case only the numerator can take negative values and a higher denominator means more risk. For positive equity values we expect that the higher ratio is connected to lower risk (low PDs). If a company has -10 equity and 100 debt, then the ratio is -10 %. Suppose that such company deleverages and now has the same amount of equity (-10) but lower debt (90). Its ratio would now be more negative -11 %, which is counterintuitive. We would expect a higher ratio, because the firm is now in better shape. Balthazar (2006, p. 153) suggests that we set ratios with a negative numerator at a common negative ratio value.

Sometimes it is important to put a variable that can have negative values in the numerator. For instance, if we have a debt/EBIT ratio, we can get wrong ordering related to risk. Suppose our company has 100 units of debt. If its EBIT is -100, then the ratio is -100 %. If the company succeeds to have less negative EBIT, then the ratio for EBIT -5 would be -2000 %. If EBIT equals 1200 then the ratio is 8 %. Naturally we would expect that when a company has higher EBIT, then the value of the ratio would be smaller and there would be fewer defaults. Our

example shows that this might not be the case at all times. A remedy for this issue is to use the EBIT/debt ratio instead.

The computation is not possible, if the denominator has a zero value (i.e. debt). If possible, we should use this variable in the numerator. Balthazar (2006, p. 153) proposes to correct it by adding a small amount to the denominator (i.e. 1 EUR of debt). However, we should be careful with this approach, as it makes the ratio value extremely huge. In modelling such values would be called outliers and they lead to distorting the estimation of parameters.

Ratios can have extreme values that do not make sense. A common first approach is to limit maximum and minimum values. By **truncation** (or winsorisation) one sets all the values above or below certain quantiles of distribution (for instance 2nd and 98th quantiles) at the values of these quantiles. By doing this we solve the problems of outliers, non-normal skewness and kurtosis[39].

An important decision one has to make is with respect to the treatment of **missing values**. First we need to identify the causes for missing values. It is important to distinguish whether they are a consequence of our calculation (i.e. a zero denominator) or they present a bias in the database (worse firms might not report some accounting variables). Then we need to evaluate the number of missing values per financial variable. If the percentage of such values is too high, we should consider excluding financial ratios that use this variable. If the percentage of such values is small, usually one would replace them by sample median values for this type of data (Balthazar, 2006, p. 151).

Four common approaches to deal with missing values are:

- define a minimum level of availability (i.e. 15 %) and exclude variables that do not satisfy this condition (more than 15 % missing values),
- include missing values as a separate factor category (possible for qualitative variables),
- replace them with estimated values specific to each group,
- replace the missing values with predictions from the regression model as in Jovan (2005) (or the classification tree), where the dependent variable is the variable with missing values, and explanatory variables are other accounting variables that are correlated with it.

ONB (2004, p. 78) proposes the combination of step 1 with step 3, where the suitable group-specific estimates for the estimation sample include the medians (averages can be heavily dominated by outliers) for the groups of good and bad cases. It is important that the median of

[39] A good indicator for the existence of outliers is the excess kurtosis. The normal distribution has excess kurtosis of zero. A positive excess kurtosis indicates that there are relatively many observations far away from the mean. Another indication is skewness. If the variables are skewed it means that extreme observations are concentrated on the left (if skewness is negative) or on the right (if skewness is positive) of the distribution (Löffler & Posch, 2007, p. 16).

the indicator value for all cases is applied in the validation sample, otherwise we would overestimate the discriminative power of a model.

2.5 Transformations

After having selected the candidate variables, the next step is to check whether the underlying assumptions of the logit model apply to the data. Hayden (in Engelmann & Rauhmeier, 2011, p. 17) stresses the importance of the linearity assumption. The score in the logit model is defined as:

$$Score=Log\ odd=ln\left(\frac{P_i}{1-P_i}\right)=\beta' x_i \tag{29}$$

This equation implies a linear relationship between the log odd and the input variables. This assumption can be easily tested by dividing the indicators into groups that all contain the same number of observations[40]. Then we calculate the historical default rate and the empirical log odd within each group. We can quickly see the relation between the ratio and empirical log odds. We can also estimate a linear regression of the log odds on the mean values of the ratio intervals and compare the regression fit statistics.

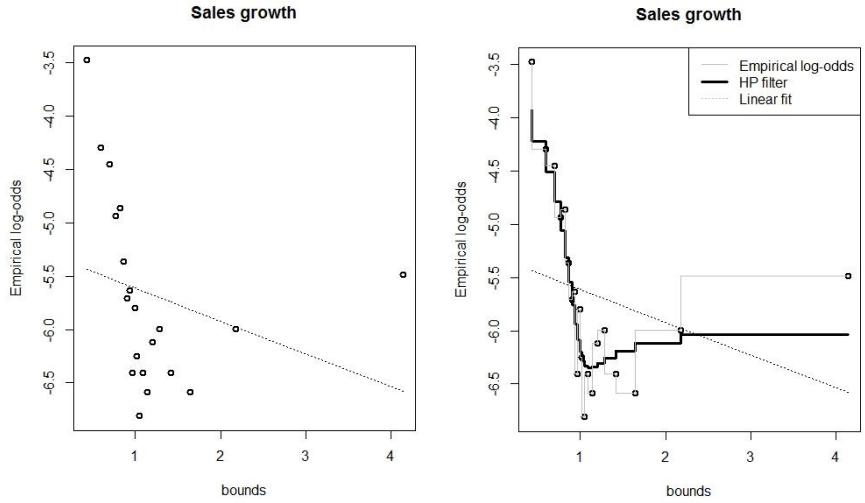
An often cited example of a variable that does not posses a linear and not even a monotone behaviour is "sales growth". Intuitively at low levels of sales growth, a company has a high risk and has weak prospects. On the other hand, high sales growth implies the firm is expanding too fast and the management might not cope well with this growth. Historical data usually shows that the least defaults are observed for firms with moderate growth. The relationship between the original ratio sales growth and the default event cannot be correctly captured by a logistic regression model without transforming/linearizing it. To capture this (U-shaped) non-monotonicity it is possible to enter the square of sales growth together with sales growth itself. Hayden (in Engelmann & Rauhmeier, 2011, p. 19) proposes to use the points obtained from dividing the variable sales growth into groups and plotting them against the respective empirical log odds (or the one-year historical default rate)[41].

Then one uses a smoothing filter, for example the Hodrick and Prescott filter, to reduce noise. Finally, substitute the original input values of sales growth with the log odds according to this smoothed relationship. The whole idea of transformation of the sales growth variable is presented graphically in Figure 8. This non-parametric procedure is also used in our empirical analysis.

[40] Hayden uses 50 groups on the Austrian data. For smaller datasets we may require only 10 or less. Too many groups may lead to overfitting and data-mining problems.

[41] Löffler and Posch (2007, p. 22) note that we need to set the minimum default rate to an arbitrarily small value, for instance to 0,0000001. Otherwise, we could not apply the logit transformation in cases where the default rate is zero.

Figure 8. Transformation of sales growth variable



More technically, we generate T groups. Denote the variable to be transformed with $y_{t\{t=1,\ldots,T\}}$. It represents the value of the categorized value of empirical odds. The Hodrick-Prescott filter is an optimisation problem:

$$\min_{\tau} \left(\sum_{t=1}^{T} \left(y_t - \tau_t\right)^2 + \lambda \sum_{t=2}^{T-1} \left[(\tau_{t+1} - \tau_t) - (\tau_t - \tau_{t-1})\right]^2 \right) \tag{30}$$

Where y_t corresponds to the categorized empirical odds. After applying a smoothing filter we obtain τ_t which is the growth component. Jovan (2005) defines λ as the smoothing parameter (positive number). This parameter decreases the variability in the input data. The higher the parameter, the more smoothed the new transformed data is. The first part of the equation lowers the noise in the cyclical component, whereas the second part lowers the noise in the trend component. Without smoothing the relation between data on the y and x axis in Figure 8 it would be jumpy and have the form of a staircase. It is clear from the figure that only the linear line does not represent the relation between sales growth and empirical odds. However, every method has some disadvantages. Falkenstein et al. (2000, p. 45) note that the commercial loans default in a cyclical fashion (there might be for instance 8 years with below average defaults).

However, this kind of transformation is not the only possibility. Authors often use also other types of transformations. The simplest transformation is truncation that we have already discussed in the previous chapters. Logarithmic transformation is an effective way of treating extreme observations. This is also useful, because we get better performance with the logit model, if the ratios have a distribution close to normal, as noted by Balthazar (2006, p. 151). However, before using the logarithm, we have to transform zero and negative values into small positive ones (by replacing them by 1 or by adding a large enough constant). One common method of transformation proposed by Falkenstein et al. (2000, p. 55) is to standardize the data (remove the mean and divide by standard deviation). Yet, this still leaves very asymmetric and fat-tailed explanatory variables. The nice feature of this transformation is that when we build

the model the betas are comparable. Other common transformations are also categorization, sigmoid functions, polynomial expansions and spline methods. If we have done transformations well and cleaned the dataset, we should expect a better fit and predictive power of our final model. In the end of this chapter it is important to stress the warning of Löffler and Posch (2007, p. 23): "In the case of transformations one should be aware of potential data-mining problems. The transformation introduces a data-driven flexibility in our analysis, so we may end up fitting the data without really explaining the underlying default probabilities."

3 UNIVARIATE ANALYSIS AND VARIABLE SELECTION

After obtaining a variable catalogue, dealing with missing values, cleaning the dataset and transforming the variables, we need to analyse the predictive power of potential variables and then select the best variables to be used as inputs to our competing models. A first method to check variables is to generate default frequency graphs. Secondly, for each indicator, it is necessary to set a working hypothesis on what relation between default rate and ratio value we expect. Only those indicators for which a clear working hypothesis can be given are useful. The univariate analysis then serves as a method to verify whether the hypothesis agrees with the empirical values. It is also common to compare univariate Area under the curve measure (hereinafter: AUC) (ONB, 2004, p. 76).

ONB (2004, p. 77) proposes to calculate the medians of each indicator separately for good and bad borrower groups. We compare whether the median values differ for good and bad groups of cases. If the hypothesis does not hold, we exclude the variable from further analysis. We also need to check the significance of variables (their p-values) after running univariate logistic regressions. Usually we would use a 1 or 5 % rejection level (99 or 95 % confidence level). However, we have to be careful before eliminating ratios, because some ratios that have a weak discriminatory power can perform well in a multivariate model.

To summarize, probably the easiest method for variable selection, is to perform some univariate tests on each of the variables while relating it to the predicted outcome, and selecting the best ones. This is simple, but also univariate, hence it does not take into account the positive effect of correlations between variables. Falkenstein et al. (2000, p. 28) proposes two main methods for multivariate selection of variables in logistic regression, these are the forward and backward selection. In the latter one starts with all variables and reduces all insignificant variables. The objective might not necessarily be the significance of variables, but also the results of the Likelihood ratio test (hereinafter: LR test)[42], as noted by Balthazar (2006, p. 129).

A good alternative is to choose the variables by using an automatic variable selection algorithm. In Ortl and Gorenjec (2015) we obtained the best results with the random forest variable importance method. We build on the definitions and derivations from the previous chapter. We

[42] If adding a new ratio does not increase the likelihood by a confidence interval of 95 %, we stop the selection process.

define the variable k_j for which we are interested to know its the importance for predicting the dependent variable. We assume we have built a random forest model, hence we know its classification errors $errOOB_t$ and the $\widetilde{OOB}_t$ for each subtree t. Now we randomly permute the values of our variable k_j inside the $\widetilde{OOB}_t$ sample. By doing so we obtain new $\widetilde{OOB}_t^{k_j}$ and also new errors $err\ \widetilde{OOB}_t^{k_j}$. The variable importance can be calculated as in Kosič (2012, p. 9):

$$Variable\ importance_{k_j} = \frac{1}{T}\sum_t^T (err\ \widetilde{OOB}_t^{k_j} - errOOB_t) \tag{31}$$

This approach is called mean decrease in accuracy. Another possibility mentioned by Hastie et al. (2013) is to calculate the mean decrease in node impurity (by summing all decreases of the Gini index after each split connected to the variable k_j – the more important variable has a bigger summation value of Gini decreases). The other alternatives are for instance factor analysis discussed in Falkenstein et al. (2000, p. 29) or hierarchical cluster analysis discussed in ONB (2004, pp. 79–80)[43].

The variable selection issue is also related to the issues emphasised by Balthazar (2006, p. 131). He notes the importance of avoiding the classical trap of regression analysis: *over-fitting*. With regression models, for instance, adding a variable always increases the *R*-squared. Over-fitting means calibrating the model too much on the available data. The model will show a high performance on such data, but the results will be unstable on the new data. The other issue is the transparency of the model. A model with only a few ratios will allow users to have a critical view of it. Also, the identification of cases when the model fails will be easier. Although there is no recipe for the absolute optimal number of variables, Balthazar (2006, p. 133) and Falkenstein et al. (2000, p. 29) recommend using around 4–8 variables and Hayden (2002) proposes around 7–15 variables.

4 MODEL SELECTION

Now that we have candidate variables to include into a model, our job is to find the optimal combination of these candidate variables with respect to several goals. We can set our objective to obtain the best fit in the sample or the best AUC out of the sample. In this chapter we will discuss statistics derived from the logistic regression model to measure its fit. In the next chapter we will present measures of discriminatory power. In addition, we also need to make sure that the variables are not correlated, as this would induce multicollinearity problems in logistic regression. For ANN and RF algorithms, the topics of this chapter are in general not so important.

[43] We create groups (i.e. clusters) of indicators which show high levels of correlation within the clusters but only low levels of correlation between the various clusters. The variables that return high AUC values are selected from the clusters.

4.1 Model statistics (logit)

Balthazar (2006, p. 136) notes the distinction between statistical tests and economic performance measures. **Statistical tests** are designed to test each ratio in the model for significance and to measure the fit of a model to the underlying data. On the other hand, **performance measures** are designed to evaluate the discriminatory power of the model. We present them in the next chapter. Balthazar (2006, p. 129) recommends the use of performance measures rather than relying solely on statistical tests that are more abstract and usually lead to selecting more ratios.

Löffler and Posch (2007, pp. 245–250 and pp. 8–11) describe the most common statistics to assess fit and precision of our estimated model. The most common statistical tests are t-ratio test, Pseudo-R^2, likelihood ratio test (LR). As these tests are a very standard tool in credit modelling literature we do not discuss them in detail and refer an interested reader to the references above.

4.2 Model selection

Another step that we need to perform in logistic regression (not needed in RF or ANN) is correlation analysis. Balthazar (2006, p. 171) notes that two ratios can show good performance individually, but using them both in the same model worsens the results, if they are correlated. One potential problem is that we obtain a wrong sign for one of the correlated variables (opposed to what is expected from financial theory, a wrong sign is assigned to the weaker variable). The other problem is that they increase the standard errors of the estimated parameters, which leads to a lower predictive power of the model out-of-sample. ONB (2004, p. 80) suggests to calculate rank order correlation (i.e. Spearman correlation). This is especially important for untransformed variables that are not normally distributed and for small samples. In comparison to the more commonly used Pearson correlation, it uses only the ranks of variable values, not the variable values themselves. Balthazar (2006, p. 171) recognizes that there is no absolute rule regarding the level above which this becomes a problem. As a rule of thumb he suggests that we should take care when correlation is above 70 %. ONB (2004, pp. 79–80) is even more conservative and indicates correlation of 30 % as potentially dangerous. Hayden (in Engelmann & Rauhmeier, 2011, p. 19) pools correlation values of above 50 % into one group. From each correlation subgroup she selects the best predictor measured with AR.

In the last step of model building Balthazar (2006, p. 129) recommends to check again whether the signs of estimated coefficients are in line with the working hypotheses. If we use a (uniform) transformation to default probabilities, all coefficients must bear the same sign. Also important is to obtain significant individual coefficients. ONB (2004, p. 81) prefers to decide on the best model based on the discriminatory power of the scoring function, the stability of the discriminatory power and at the same time to include all variables that cover all relevant credit risk information categories (profitability, leverage, efficiency, liquidity …).

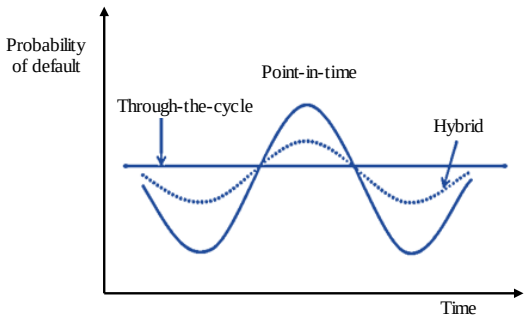
5 MODEL CALIBRATION

5.1 Rating philosophy

A very important distinction to be made before validating a credit default model is a distinction between point-in-time (hereinafter: PIT) and through-the-cycle (hereinafter: TTC) PDs, which are the two main rating philosophies. An additional philosophy is the hybrid approach. Hamilton, Sun, & Ding (2011, p. 9) and BIS (2005, p. 14) emphasise that the key difference between PIT and TTC approaches lies in their information content. A PIT measure uses all relevant borrower-specific information about a firm's creditworthiness, but includes also recent macroeconomic information[44], whereas a TTC measures the firm's long-term credit quality. TTC PDs are free of cyclical variation in the macroeconomic environment. Hence, we can look at it as a problem of signal extraction. We can ask ourselves: "Given an observed change in a PIT PD, how much of it is attributable to a change in the long-run credit trend and how much is due to the cyclical factors?" In other words, PIT measures an idiosyncratic risk specific to the client and TTC measures a systematic risk specific to all borrowers in the economy.

Aguais, Forest, Wong, & Ledezma (2004, p. 6) and Novotny-Farkas (2015, p. 19) describe that a PIT PD measures the default risk over a short horizon, often considered a year or less. It is based on all available information and often includes macroeconomic variables. On the other hand a TTC PD measures default risk over a horizon long enough for the business-cycle effects mostly to be neutralised (over a period of five or more years). A TTC PD should be independent of the cycle and therefore one should not use explanatory variables that are correlated to the cycle for estimation. A TTC approach provides stable and less cyclical ratings. In contrast, PIT PDs vary significantly during a turbulent period. The hybrid approach is a combination of TTC and PIT models.

Figure 9. PIT vs TTC vs hybrid approaches



Source: Wolters Kluwer (in Z. Novotny-Farkas, *The significance of IFRS 9 for financial stability and supervisory rules*, 2015, p. 19).

[44] Macro shocks (shifts in monetary policy, variation in aggregate output, and the supply of credit) tend to affect all firms' credit conditions at the same time. Credit ratings then fall or rise at the same time. These changes are often temporary, for instance think about the impact of economic news. On the other hand they occur more frequently than shocks to the firms' underlying credit quality (Hamilton et al., 2011, p. 9).

Article 180 of the CRR requires banks to estimate PDs by borrower grade from long-run averages of one-year default rates to avoid unnecessary fluctuations in regulatory capital. Ingolfsson and Elvarsson (2007, p. 1) describe that "long-run" should be taken as at least sufficiently long to cover a full credit cycle. However, not all banks have sufficiently reliable data on which to train a credit scoring model for such a long period. The problem is how to construct a model that predicts not only a short term PD (reflecting the current credit environment), but the required long-run PD. Ingolfsson and Elvarsson (2007, p. 6) note that for a logistic regression model based on data that covers a one-year period there is the problem that the model picks up the economic cycle reflected in the dataset. As a result, it predicts lower or higher average PDs, depending on the current state of the economy. Models naturally tend to be PIT because they are usually based on the last available year of financial statements.

Hamilton et al. (2011 p. 8) note that there is still no general consensus about basic definitions. Novotny-Farkas (2015, p. 19) notes that the CRR does not require a specific rating philosophy, but clarifies that PD estimates should reflect the long-run average of one-year default rates in order to ensure that they are relatively stable over time. Hence, it seems that TTC or hybrid approaches are more consistent with the capital adequacy framework. Another big issue stated by EBA (2015, p. 24) is the question of what to do, when there is no data available that would cover the whole economic cycle. Usually an additional margin of conservatism is to be adopted. The second question is a how long time series to use for calibration.

The bank chooses an approach depending on the intended use. Different philosophies have different advantages and disadvantages. There is a wide range of possible applications of default models. Different uses have conflicting requirements. For instance, Basel seems to prefer TTC conservatism for capital calculation[45], but for pricing, loan origination and stress testing we would prefer the PIT approach. For IFRS purposes one needs to use the PIT approach. In a downturn, IFRS 9 impairments might be larger than regulatory expected loss, because the PIT will be larger than TTC PDs (Novotny-Farkas, 2015, p. 21; PWC, 2014, p. 8).

An important question is how to calculate TTC PDs. They can be calculated from PIT PDs or modelled separately. Hamilton et al. (2011, p. 10) use the fact that PIT PDs contain all the information that we need to extract TTC PDs. Thus to use external macroeconomic variables to identify the cyclical effect is not needed. Their approach is based on filtering out the cyclical TTC effect from PIT estimates, which is easier than estimating where in the cycle we are and to predict its future direction. Their approach is developed on the basis of a Merton like default model.

[45] During recessions credit models might overstate default risk. As a consequence capital requirements will increase, which might lead banks to tighten their credit standards. This is an undesired effect. One way to deal with this is to use TTC PDs instead of PIT PDs. Saurina and Trucharte (2007) discuss also other possibilities to adjust capital requirements calculation.

One simple approach in the context of logistic regression model that is used for instance in the United Kingdom is the variable scalar approach[46]. Different variants of this approach have been developed. Begin and Thomas (2012, p. 14) basically convert the PIT PD to a TTC PD via a scalar that varies through the credit cycle. In a period with low credit losses, the scalar will adjust the PIT PD upwards to the TTC PD. In a period with high credit losses, the adjustment will be downwards. The scalar adjustment should be calibrated at risk grade level.

The second approach consists of building a separate rating system for TTC PD rather than adjusting an existing PIT model. Begin and Thomas (2012, p. 15) note that under this approach, TTC PDs are determined by including macroeconomic variables (this could also be on a regional or industry level as long as they can be interpreted as cycle variables). One approach that is often cited in the literature is the approach of Saurina and Trucharte (2007, p. 18) who use the value of the GDP to calculate the TTC. They use the GDP value for the most negative year (worst recession) in the sample. To compute the PD they keep the GDP variable constantly at the worst level, while allowing for changes in other significant variables included in the model. They also provide an acyclical PD, which is a reestimation of the model without including the common factor variable (GDP). The cyclically corrected PD is obtained by substituting in the model the value of GDP fixed at its sample average value and allowing changes in other variables.

The rating philosophy has an important impact on backtesting/validation. Both PIT systems as well as TTC systems are in general fit for the IRB. However, the validation of TTC-ratings is extremely challenging. For backtesting we also need data that covers the whole economic cycle (Blochwitz & Hohl in Engelmann & Rauhmeier, 2011, p. 262).

Agency ratings are assumed to be TTC, whereas most bank internal systems (i.e. in Germany) are PIT, as stated by Blochwitz and Hohl (in Engelmann & Rauhmeier, 2011, p. 262). The EBA (2013a, p. 43) report on IRB banks shows that the TTC approach has been adopted by more than half of the banks in their sample (19 out of 35). EBA (2013b, p. 79) notes that sometimes they could not find large differences in the method of rating calibration between two banks, even if one of them defined itself as TTC and the other as PIT.

Balthazar (2006, p. 143) is very critical about these approaches. He notes that it is impossible to give a rating TTC because covering an economic or business cycle is not as easy as scaling PDs over three to five years. It is important to note that a business cycle may differ from one sector to the other. Hence the TTC approach can lead to a great risk. Rating agencies have been highly criticized because of their slow reactions in decreasing ratings before recessions (due to their TTC approach).

[46] The Financial Services Authority (hereinafter: FSA) allows banks to use scalar methodologies. For the Kalman filter approach see Ingolfsson and Elvarsson (2007).

Table 1. Rating philosophy

	Point-in-Time (PIT)	**Through-the-Cycle (TTC)**
Measure	Unconditional PD	PD conditional on the state of the business cycle
Grade allocation (stability of borrower's rating grade over the cycle)	Pro-cyclical (more rating migration). Rating improves during expansions and deteriorates in recessions.	Stable (less/no rating migration). Grade assigned not dependent on economic cycle.
Rating grade unconditional PDs (stability of borrower rating grade's unconditional PDs)	Stable (unconditional PDs of rating grades do not change). Borrower's higher unconditional PDs during recession are accounted for by migrations to lower rating grades and vice versa.	Pro-cyclical. PDs improve during expansions and deteriorate during recessions.
Observed default rates by grade	Stable (actual default rates in each grade remain unchanged) because of ratings migration. Borrower's higher default rate during recession are accounted for by migrations to lower rating grades and vice versa.	Pro-cyclical (actual default rates in each grade change) because there is no ratings migration. Default rates improve during expansions and deteriorate during recessions.
Credit score distribution (Long term PDs)	The overall portfolio average PD changes. It is represented trough migration between grades.	The overall portfolio average PD remains stable through cycle.
Capital requirements	Very volatile because the overall portfolio PD quickly changes in the credit cycle (also rating migration). The cyclicality is then eliminated by the use of a cyclical scalar.	On average there is no rating migration, so capital requirements remain constant at least with respect to the lowered effect of volatility that is a consequence of the economic cycle*.

Note. * It will still change because the portfolio composition changes the underlying risk.

Source: B. Begin & S. Thomas, *Basel II retail modelling approaches PD Models*, 2012, p. 10; U. Erlenmaier (in Engelmann & Rauhmeier, *The Basel II Risk Parameters*, 2011, p. 47).

In addition, the award-winning paper BIS (2006, pp. 20, 29) found out that the strict Basel rules might need some improvements. Their paper is based on the emerging economy data. They have shown it is not feasible to construct a TTC rating system, when the period includes years of crisis. For developing countries that have gone through many recessions it is difficult to have a large database with financial information stable enough to construct a TTC rating system[47]. The EBF (2012, pp. 44, 19) report shows that even supervisors do not have clear rules on what approach to use[48]. A few of them impose the mandatory use of a TTC approach for regulatory capital reporting. The procyclicality approaches can take different forms, e.g. variable scalar or acyclicality approaches. The most common approaches in practice are variable scalar methodologies, add-on capital requirements or increase of PD or use of longer historic data series.

[47] For more discussion and proposed solutions (i.e. 2 year data, weighting years) see BIS (2006, p. 29).
[48] According to EBF (2012, p. 19), some guidelines have been prepared by the Bank of Spain, The Bank of England and also an Italian regulator are working on it.

5.2 Model calibration

The final step is to assign a default probability to each score obtained from logistic regression (it may be an individual score or rating). Even though the output of logistic regression[49] is a (sample-dependent) probability of default, it may not be calibrated for the portfolio on which the model will be used for the following reasons (Balthazar, 2006, p. 129; ONB, 2004, p. 85):

- the default rate in the estimation sample may not be representative of the default rate of the bank portfolio on which the model predictions will be used (it may be that we have used a higher proportion of defaults in the estimation sample),
- the default definition in the estimation sample may not be the same as the Basel default definition or as the definition in the bank's portfolio (i.e. we estimate on a sample where default is defined only as bankruptcy filing, but a bank defines default also as a 90-day overdue),
- to map a model rating-scale to a bank's master-scale[50].

Thus the PDs given by the model must be adjusted. In the empirical part we scale sample-dependent default probabilities (determined by logistic regression) to the average portfolio default probability of this segment. Sample default rates are not scaled directly by comparing the default probabilities in the sample and the portfolio, but indirectly using relative default frequencies (hereinafter: RDF). ONB (2004, p. 86) describes RDFs as the ratio of bad cases to good cases in the sample. If we want to calibrate PDs we can use the following procedure of ONB (2004, p. 86):

- calculate the average sample default rate resulting from logistic regression (using a sample of the current non-defaulted portfolio) and convert it into RDF_{sample},
- calculate/estimate the average portfolio default rate and convert it into $RDF_{portfolio}$,
- calculate the representation of each default probability resulting from the logistic regression model as $RDF_{unscaled}$,
- multiply $RDF_{unscaled}$ by the scaling factor specific to the rating model,
- convert the resulting scaled RDF into a scaled default probability.

$$RDF_{scaled} = RDF_{unscaled} \cdot \frac{RDF_{portfolio}}{RDF_{sample}} \tag{32}$$

Where $RDF = \frac{PD}{1-PD}$ *or* $PD = \frac{RDF}{1+RDF}$. RDF_{sample} is calculated from the average predicted probability of default, which comes from our implementation sample. $RDF_{portfolio}$ reflects the market/portfolio average default rate. Lastly, the $RDF_{unscaled}$ is calculated from the individual default probabilities resulting from logistic regression (ONB, 2004, p. 86).

[49] For calibration of other statistical, heuristic and option based models see ONB (2004, p. 85).
[50] A separate segment-specific calibration is necessary for each rating model even if we use master-scales because each segment is specific (ONB, 2004, p. 84).

Calibration can also be done by benchmarking with an external model. If logistic regression is only one of several modules in the overall rating model (e.g. in hybrid systems) and the rating result cannot be interpreted directly as a default probability, refer to ONB (2004, p. 85). EBF (2012, p. 19) warns about the use of continuous master-scale[51], the period of time to be considered when calibrating PDs[52], the methodology for calculating cycle adjustments, the size of PDs by the imposition of add-ons[53], the frequency of recalibration[54] etc. EBA (2013c, pp. 19–20) reports that banks use different approaches for calibration. The length of time series used for calibration is sometimes shorter than five years. As stated by the EBA (2015), guidelines will be given in the future on the definition of an economic cycle, the identification of stressed years and on how to cope with absence in the available time series covering the stressed period.

6 OPTIMAL RATING STRUCTURE

Now that we have obtained probabilities of default it is a common approach to transform them to rating grades. The motivation for using rating grades is that for people it is easier to think in a few discrete values instead of a continuous scale. Rauhmeier (in Engelmann & Rauhmeier, 2011, p. 315) notes that this is especially relevant when combining machine ratings with manual overrides (by upgrading or downgrading ratings). With many grades, the pricing of loans is better tailored to individual risk. BIS (2005, p. 10) notes that an IRB bank must usually first allocate borrowers to "risk buckets", where borrowers with the same credit quality are allocated to the same bucket. Article 170 of the CRR (EC, 2013) forces banks to use at least seven rating grades for non-defaulted companies and one for defaulted[55]. A regulator may require even more grades for banks, depending on risk concentration in certain grades. The bank must calculate a "pooled PD" for each bucket (for instance by using medians or averages). Löffler and Posch (2007, p. 215) show that more grades lead to better discriminatory power and less capital requirements, which is an important feature to note.

Tasche (in Christodoulakis & Satschell, 2008, p. 177) notes that the use of rating grades simplifies the validation process (also because of the efficiency properties of averages). Note that if there are too many grades, the backtesting might not be possible, because it may happen that there are no observations in some grades. In the EBA (2013b, p. 77) report it is shown that more than 70 % of large EU banks use more than 10 rating grades for SME portfolios. In EBA (2013c, p. 18) they state that the number of grades usually ranges from 9 to 30, with varying PD

[51] Some regulators allow continuous master-scales. Among banks who use buckets, the EBF report shows, that they use between 8 and 50 buckets. Article 169 of CRR (EC, 2013) allows a continuous grading scale, if the bank can provide probabilities of default for all borrowers directly from the model. This leads to more accurate, but lower capital requirements because the capital requirement function is concave (for further discussion see Löffler and Posch (2007, p. 215).

[52] In some countries regulators require that the years used in the calibration are scaled up to incorporate certain years, where there was a crisis (i.e. historical scenario). If data series are not available, banks need to calculate cycle adjustments (EBF, 2012, p. 21).

[53] According to the EBF (2012) questionnaire, 16 out of 42 banks have additional add-ons imposed by the regulators.

[54] Some require annual recalibration while other expect regular model monitoring (EBF, 2012, p. 43).

[55] The BIS (2001, p. 9) survey has shown that banks use on average 10 rating grades for performing loans and 3 for non-performing.

levels and floors. However, the differences between banks are huge. Not only that they use a different number of rating grades (different granularity), but also the distributions of exposures in the different rating grades are very different. This may lead to very different capital requirements. Rauhmeier (in Engelmann & Rauhmeier, 2011, p. 315) warns that there are also disadvantages when using rating grades, especially when one relies solely on grades and no more on PDs. The grades do not take into account the exponential relationship of grades and corresponding PDs (downgrading from grade 8 to grade 4 might mean nearly a tenfold increase in the forecast PD and not doubling it). When constructing a rating grade structure Balthazar (2006, p. 147) discusses an important issue of the distribution of the available ratings. First it is important that the rating distribution matches the rating distribution of the bank portfolio on which the model will be used. The distributions of the exposures can be "bell-shaped" with few exposures on the left and on the right side of the graph. To obtain the same average error over all ratings it is possible to set a distribution with an equal number of observations in each rating grade. Another possibility is to allocate a higher number of observations into low ratings, because good performance on the borrowers is very important. However, article 170 of the CRR (EC, 2013) requires that a distribution is meaningful and has no excessive concentrations in certain grades. BIS (2001, p. 43) proposes that no more than 30 % of the gross exposures fall into any single borrower grade. Balthazar (2006, p. 147) recommends to obtain the highest concentration on average quality borrowers where most of the exposures usually lie. Now we come to the question on how to construct these rating grades optimally. Defining a grading structure is an important step because such a decision is typically meant to last for several years. One possibility is a machine learning approach called K-means cluster analysis (see BIS, 2006, p. 28).

We use another possibility. First assume that the average expected number of defaults in each rating grade will equal the actual number of defaults in that grade. Then we can use the optimisation algorithm to calculate border PDs for rating grades by maximizing the overall expected AUC. Löffler and Posch (2007, pp. 218–222)[56] achieved this by running Monte Carlo simulations. We use a heuristic[57] global optimisation algorithm Differential Evolution (hereinafter: DE). DE is a population based method, which was developed by Storn and Price in 1997. The algorithm can be summarized as in Gilli, Maringer, & Schumann (2011, p. 347) and Ortl (2014).

With the DE minimization algorithm we try to set optimal upper bounds, i.e. PDs that define rating grades. By optimal we mean that such upper bounds are chosen with the objective of achieving the maximum discriminatory measure of the rating structure that is measured by AUC. Our rating grade allocations will thus lead to the rating structure with the best possible discriminatory power. The algorithm is based on a population matrix P with dimensions $d \times n_P$. Columns represent candidates for an optimal set of bounds ($i=1,\ldots,n_P$)), while rows are upper

[56] See also Anderson (2007, pp. 419–432).
[57] Heuristic methods are not deterministic, they have some stochastic elements which help them not to get stuck in a local optimum. Getting stuck in a local optimum is a problem when we do not have a convex problem (Ortl, 2014).

bounds ($j=1,\ldots,d$). Later in our empirical part we will have d=7 upper bounds (for 8 rating grades), which need to be set optimally by maximizing AUC and at the same time not violating any constraint (we will define this constraint in the empirical part of this thesis). Each column/candidate i in the population matrix $P_{.,i}$ has 7 upper bounds. With 0 we define the initial population matrix, $P_{j,i}^{0}$, i.e. starting population. $P_{j,i}^{1}$ is the population in the next generation/iteration ($k=1,\ldots,n_G$). In each iteration the algorithm evaluates each column in the population matrix. The columns that give good results have more chance to be chosen to the next population matrix in the next generation/iteration. This choice also depends on the learning parameters, which can be tuned to obtain the best tradeoff between "not getting stuck in a local optimum" and fast convergence of the algorithm. Parameters have the effects of mutation (randomness to prevent getting stuck) and crossover to the next generation. These parameters are:

F ... scalar factor of the difference vector for mutation. High value F means slower convergence of the algorithm but less chance of getting stuck in a local optimum (mutation effect).
CR ... crossover probability parameter, its domain is [0,1]. A higher value means higher crossover. If the value is too low, we basically just reproduce the initial column in the starting population matrix.
n_P ... size of population (we need a big enough population).
n_G… number of generations/iterations.

Figure 10. DE algorithm

1: set n_P, n_G, F and CR
2: randomly generate initial population $P_{j,i}^{(1)}$, $j=1,\ldots,d$, $i=1,\ldots,n_P$
3: for $k=1$ to n_G do
4: $P^{(0)}=P^{(1)}$
5: for $i=1{:}n_P$ do
6: randomly generate $r_1 r_2, r_3 \in 1,\ldots,n_P$, $r_1 \neq r_2 \neq r_3 \neq i$
7: compute $P_{.,i}^{(v)}=P_{.,r_1}^{(0)}+F\times(P_{.,r_2}^{(0)}-P_{.,r_3}^{(0)})$
8: for $j=1$ to d do
9: if $rand<CR$ then $P_{j,i}^{(u)}=P_{j,i}^{(v)}$ else $P_{j,i}^{(u)}=P_{j,i}^{(0)}$
10: end for
11: if $f\left(P_{.,i}^{(u)}\right)<f(P_{.,i}^{(0)})$ then $P_{.,i}^{(1)}=P_{.,i}^{(u)}$ else $P_{.,i}^{(1)}=P_{.,i}^{(0)}$
12: end for
13: end for
14: return best solution

Source: M. Gilli, D. Maringer, & E. Schumann, *Numerical Methods and Optimization in Finance*, 2011, p. 347.

7 IMPLEMENTATION ISSUES

Now that we have a good working model, we also have to integrate it into our process and into our systems. There are many issues in the implementation. The first thing is how to implement the model to all relevant processes in a bank. We cannot cover all issues in this chapter, but we can name a few.

Managers and loan officers are often curious about the effect of some variable in a logistic regression on the overall rating grade. If we have used a normalisation as a transformation procedure we already have comparable parameters on the same scale. Because in our case the relationship is not linear (we use the logistic link function to obtain probabilities) we can use sensitivity analysis or compute marginal effects as in Löffler and Posch (2007, p. 24) to obtain the impact of individual variables. If this is not the case, then the comparison is a bit harder. Falkenstein et al. (2000, pp. 61–63) propose the use of numerical derivatives or the use of relative contributions, where they calculate the PD for an average/median company and then recalculate separate variables (one by one) by leaving other variables unchanged.

As we discussed in the previous chapter it is a common banking practice to use rating grades. Rauhmeier (in Engelmann & Rauhmeier, 2011, pp. 315–316) notes that banks often use a bank wide rating scale called a "master-scale". All rating systems are mapped into this scale[58]. With a master-scale it is easier to generate reports because rating grades of the master-scale mean the same across different segments and across different years. EBA (2013b, p. 76) reports that the majority of the large EU banks use only one master rating grade scale for the calculation of the capital requirements (some use up to six for different regulatory portfolios). BIS (2005, p. 100) provides an illustrative example of such a master-scale. Consider a bank's SME lending portfolio that is modelled using a 10-grade rating scale, while the corporate portfolio uses a 20-grade rating scale. If the master-scale has less grades (i.e. 10), we will lose some information from the more granulated corporate portfolio and also the discriminatory power will be lower. For instance, the South African Reserve Bank defined an internationally comparable 25 point master rating scale with predefined border PDs and predefined mapping of external ratings (Moody's, Fitch and S&P). Exposures are mapped to this master-scale according to their calibrated segment's rating grade PDs (Adams, 2009, p. 6).

Another important issue is how to cope with structural issues such as parental support, exposure to holding companies, new subsidiaries, joint ventures, relationships to government, etc. Samuels (2012, p. 15) notes that the effects can be positive (e.g. parental support) or negative (e.g. cash flows might be allocated between subsidiaries). He is worried that statistical models rarely cope well with such complexities. Erlenmaier (in Engelmann & Rauhmeier, 2011, p. 70) proposes to rate both companies on a standalone basis. The borrower's rating is then usually

[58] Engelmann and Gruber (in Engelmann & Rauhmeier, 2011, p. 341) notice that the best rating grades in such rating scales are usually intended mainly for sovereigns, international large corporates and financial institutions with excellent creditworthiness (and only a few small business clients). However, this observation depends on the types of exposures a bank has.

calculated by weighted average of the associated PDs, (the weight should depend on the degree of influence as determined by the loan manager). In the report of EBF (2012, p. 19) it was found that only one half of banks cope with this problem. Some banks use the ratings of the parents or they use the best, worst or a combination of the two ratings.

Another important issue is to recognize that the model cannot assess all risks and does not take into account all relevant data. The regulation requires banks to update their ratings at least every 12 months, or as soon as any new relevant information becomes available. We have already discussed the possibilities of combining quantitative and qualitative information (with overrides). Complementing the quantitative model with human judgment is required also by article 174 of the CRR (EC, 2013). Human judgements shall take into account all relevant information and weaknesses of the model. Rauhmeier (in Engelmann & Rauhmeier, 2011, p. 313) proposes implementing a modular approach. The first module is often called "machine rating", typically based on statistical models. This module can also contain qualitative inputs from a loan officer. The second module is called "expert-guided adjustments". An analyst can make adjustments of the rating subject to borrower characteristics that are not sufficiently reflected in the first module. Usually this is done in a predefined standardised way. In the second module, one can upgrade or downgrade the borrower rating. The third module is the "supporter logic" and captures effects, if a parent company supports its subsidiary and similar. Here the potential supporter ratings are used, again with the help of predefined rules. The fourth module is in place because it is impossible to foresee every important characteristic that may lead to default. This module is called "manual override". Note that overrides should not be used too often and must be well documented and approved by a senior management board. As required by article 172 of the CRR (EC, 2013) our IT system should track all changes and allow for later validation of such overrides[59].

Also interesting is the treatment of cured companies. EBF (2012, pp. 13, 21) reports that more than a half of the surveyed banks do not penalise cured borrowers with a higher PD. On the other hand, 20% of the banks assign a higher PD for some period. The remaining banks assign a higher PD depending on the type of default only. Additionally, a past default is often used as an additional factor in the models (this is required for instance in Spain and Germany). The United Kingdom has an especially strict definition of forbearance, because some borrowers need to be treated as being in default for regulatory capital purposes, even if they have been "cured".

There are many more issues that we haven't discussed (i.e. how do we ensure data quality, feedback from model users, staff training, IT implementation, audit trail, foreign currency lending, stress testing etc.)[60].

[59] To compare the discriminatory power of two models (rating model versus final rating model with overrides) we can use the Redelmeier test (see Rauhmeier (in Engelmann & Rauhmeier, 2011, p. 325)).
[60] For details see EBF (2012, p. 37), ONB (2004, p. 130) or Löffler and Posch (2007).

8 VALIDATION

To obtain confidence in the model results and to satisfy regulator's requirements a bank needs a proper validation methodology. For instance, under the IRB approach, regulators are interested to know whether the calculated capital charges will be high enough to absorb the potential losses from borrower defaults. The design of a validation methodology depends on the type of rating system (BIS, 2005, p. 7).

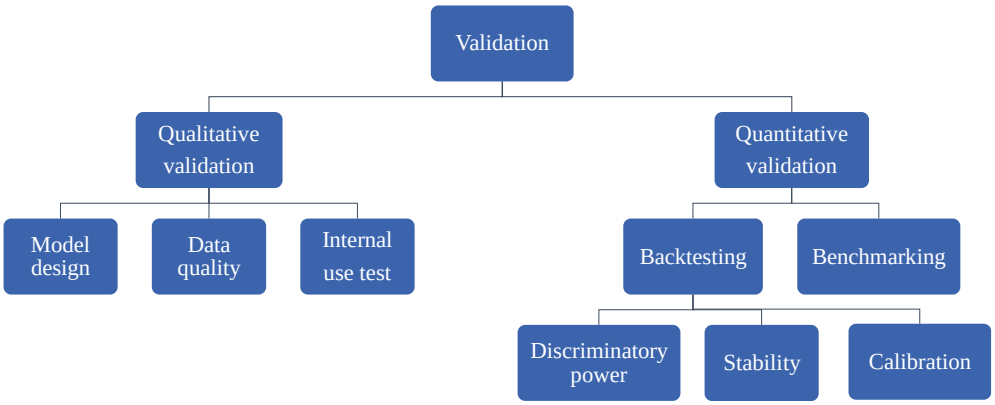
The Subgroup on Validation of the Basel Committee on Banking Supervision (hereinafter: BCBS) developed six important principles for validation methodology (Blochwitz & Hohl in Engelmann & Rauhmeier, 2011, p. 253):

- First they recognize the importance of good discriminative power of a model. A model and its rating grade structure must discriminate well between risky and good borrowers. The estimates of probabilities of default must be accurate. Very important is also the implementation of a rating system into all relevant credit processes (see also article 185 of the CRR directive (EC, 2013)).
- The bank has primary responsibility for validation.
- Validation and monitoring should be ongoing processes.
- There is no single validation method. Blochwitz and Hohl (in Engelmann & Rauhmeier, 2011, p. 253) note that there is an unanimous agreement that there is still no statistical test available, which could be used as the main validation tool.
- Validation should encompass both quantitative and qualitative elements. Very important aspects are also data quality, documentation, internal use and the use of the model in all credit processes (see articles 174–176 of the CRR directive (EC, 2013)).
- The CRR Directive (EC, 2013) in articles 144, 190 and 190 requires also that validation processes and outcomes are subject to an independent review by an independent credit risk control unit and by internal audit.

In Figure 11 it is shown what the main building blocks of validation process are. First one needs to distinguish between quantitative and qualitative validation. Under quantitative validation it is important to test the model’s discriminative power, calibration and stability. In the next chapters we will present each of these blocks in detail.

The regulators usually mark the whole model as unsuitable only when the qualitative part of validation is rejected. Failing quantitative discriminatory power testing usually leads to obligatory modifications or rebuilding of the model. In case of unsuccessful calibration, one usually needs to recalibrate the model. When failing the stability review, the bank needs to review causes and possibly lengthen the time series in the model (ONB, 2004, p. 95).

Figure 11. Types of validation



Source: ONB, *Rating models and validation*, 2004, p. 97.

8.1 Qualitative validation

Although we will not go deeply into qualitative validation, it is important to provide some insight, since it is a very important part of validation from a supervisor's perspective. EBA (2015, p. 31) stresses the importance of corporate governance that includes in particular: credit process, credit risk control unit, validation function, internal audit and the function of senior management. ONB (2004, p. 96) distinguishes three main areas of qualitative validation:

- model design (scope, transparency and comprehensiveness of documentation),
- data quality,
- internal use test (actual integration of the rating system into the bank's credit approval, risk management, internal capital allocations, pricing, provisioning, reporting systems, etc.)[61]

For regulators it is also very important that the bank is staffed with personnel that has sufficient expertise and is supported with adequate resources. Balthazar (2006, p. 116) and article 189 of the CRR directive (EC, 2103) add that it is also important that all material aspects of the rating process are understood and endorsed by the senior management. The German central bank DB (2003, p. 66) emphasises the acceptability and understanding of the rating systems for all users of a rating system. The institution must implement a regular cycle of model validation and monitoring. Borrowers should be reviewed by an independent credit unit at least annually (article 172 of the CRR directive). Higher risk borrowers need even more frequent updating (BoE, 2013, p. 9; BIS, 2001, p. 44; ONB, 2004, p. 94).

Balthazar (2006, p. 116) stresses the importance of independent validation. Ideally the bank would have an independent unit responsible for development, implementation and monitoring of the rating system. The other possibility is external validation from vendors. EBA (2015, p.

[61] For the Bank of England (BoE, 2013, p. 9) and for Deutsche Bank (DB, 2003, p. 66) it is very important that a bank demonstrates its confidence in the model by using the rating system in all aspects of business.

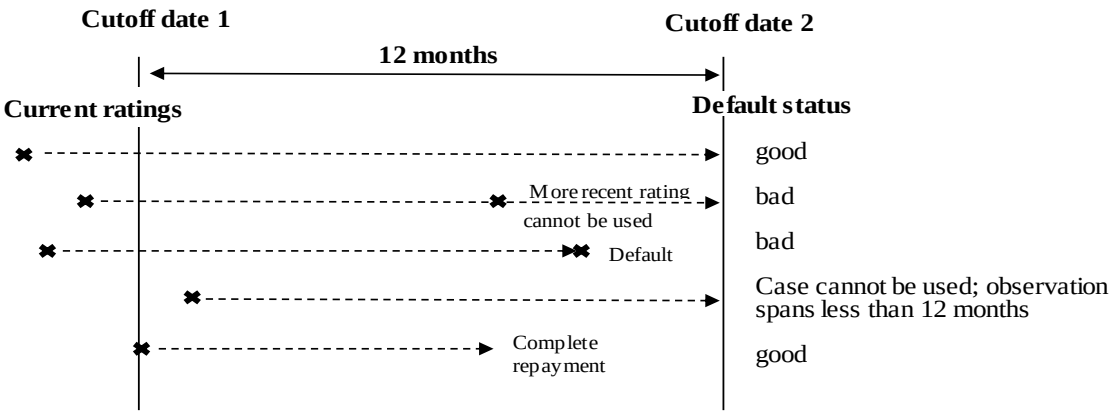
32) allows smaller institutions to operate without an independent unit, but must prove that the validators are independent. EBF (2012, p. 42) recognizes that the validation process differs from one supervisor to another. Sometimes the supervisors replicate and review the models in detail, while other focus only on the results, qualitative issues and main methodology aspects.

8.2 Quantitative validation

Löffler and Posch (2007, p. 147) provide a nice description of the two main concepts of quantitative validation. These are discrimination (How well does a rating system rank borrowers according to their true PD?*)* and calibration (How well do the estimated PDs match true PDs?). A rating system has good discriminatory power, if actual defaults occur mainly among borrowers that were assigned a bad rating. The best rating system would have all defaulters classified by the lowest rating category. DB (2003, p. 62) defines that a rating system is well calibrated, if the estimated PDs deviate only slightly from the actual default rates. We cite Falkenstein et al. (2000, p. 30) for a nice distinction between two concepts: "a model may be powerful, but not calibrated, and vice versa. For example, a model that predicts that all firms have the same default probability as the population average is calibrated in that its expected default rate for the portfolio equals its predicted default rate. A powerful model, on the other hand, can distinguish goods and bads, but if it predicts an aggregate default rate of 10 % as opposed to a true population default rate of 1 %, it would not be calibrated."

An additional issue is the stability of the model. A stable model would show stable predictions over time. In addition one must understand the difference between backtesting and benchmarking. BIS (2005, p. 8) describes backtesting as comparing the PD estimates to realized outcomes. It is based on the bank's internal data. In contrast, benchmarking uses external benchmarking datasets. It refers to a comparison of internal estimates vs estimates of other banks, external ratings, vendor models or models developed by supervisors. The two approaches are based on similar validation procedures but are interpreted differently.

Figure 12. Creating a rating validation sample for a forecasting horizon of 12 months



Source: ONB, *Rating models and validation*, 2004, p. 100.

Before presenting discrimination, calibration and stability in detail, we must define how to prepare data in a way that it is useful for validation. ONB (2004, p. 98) notes that a sufficient dataset for quantitative validation is available only when all loans have been rated with the new model for the first time and observed over the forecasting horizon of the rating model. As proposed by ONB (2004, p. 99) we first need to define two cutoff dates (usually end-of-year) with an interval of 12 months. In Figure 12 is shown how to define a sample for validation.

8.2.1 Discrimination tests

The discriminatory power denotes the ability of a model to discriminate ex ante between defaulting and non-defaulting borrowers. The tests should be applied not only in the estimation sample but also in the validation sample (out-of-sample validation), to reduce the risk of overfitting, which would otherwise result to a low stability of the model. In the study related to IRB validation techniques, BIS (2005, p. 30) discusses the most commonly used statistical methodologies for testing of the discriminatory power:

- Cumulative Accuracy Profile (hereinafter: CAP)[62] and its summary index, the Accuracy Ratio (hereinafter: AR),
- Receiver Operating Characteristic (hereinafter: ROC) and its summary index, the area under the ROC curve (hereinafter: AUC),
- Pietra index and Kolmogorov-Smirnov test (hereinafter: KS test),
- Bayesian error rate[63],
- Entropy measures such as: Conditional entropy, Kullback-Leibler distance (hereinafter: KL measure), Conditional Information Entropy Ratio (hereinafter: CIER) and Information Value,
- Correlation measures such as: Kendall's tau and Somers' D,
- Brier score (it can also be seen as a calibration test).

The CAP and ROC are visual tools but can be calculated as summary indices AR and AUC respectively. Another summary index for ROC is the KS statistic. BIS (2005, p. 32) recognizes AR and AUC as the most useful measures because of their statistical properties. For both, it is possible to calculate confidence intervals. Entropy is a concept from information theory. It is related to the extent of uncertainty that is eliminated by an experiment. Measures related to this are Conditional Entropy, Kullback-Leibler distance, CIER, and the Information value. If one wants to compare the degree of concordance of two rating systems, BIS (2005, p. 32) proposes two rank-order statistics such as Kendall's tau and Somers' D. BIS (2005, p. 32) describes the Brier score as a sample estimator of the mean squared difference of the default indicator variable and the predicted PDs. A very common measure is also Percentage correctly classified (hereinafter: PCC). Here one calculates a confusion matrix of actual versus predicted class labels. To produce discrete class predictions we have to define a cut-off value. One way to

[62] Also known as the Gini curve, Power curve or Lorenz curve.
[63] Also known as the Classification error or Minimum error.

calculate it is to use KS-statistic, as we will discuss later. Another possibility often used in articles (for instance Lessmann et al. (2015, p. 29)) is to use the prior probability in the estimation set (or in the validation set).

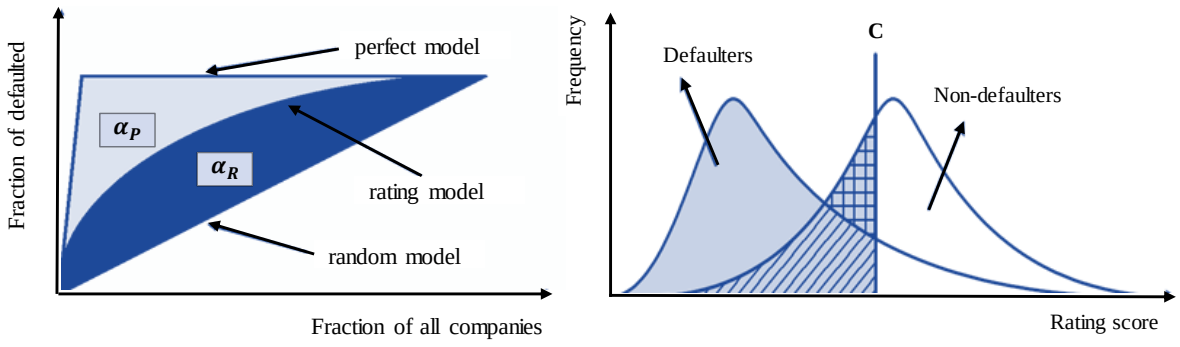
It is interesting that different measures sometimes return a different ranking of the models. This was discussed in Lessmann et al. (2015, pp. 35–39). They calculate Kendall's rank correlations and note that AUC, H-measure and PPC are highly correlated. On the other hand the Brier score gives quite different rankings, which are not correlated to the rankings from the other three measures.

Now we discuss the theoretical properties of discriminatory power tests. We discuss in more detail AUC, AR, KS, KL, Kendall tau and Brier score. Practical calculations of these tests on hypothetical data are done in the Appendixes C and D.

8.2.1.1 CAP and AR

The first and most common measure for testing discriminatory power is the accuracy ratio (hereinafter: AR). To plot it, we need data on borrower ratings in the year *t* and the default behaviour in the subsequent year *t+1*. Borrowers that are already in default at year *t* need to be excluded (Löffler & Posch, 2007, p. 148). The construction of the CAP is as follows. We first plot the fraction of all defaults that occurred among borrowers rated *r* or worse (y-axis) against the fraction of all borrowers that are rated *r* or worse (x-axis). The calculation with the corresponding graph is shown in a figure in Appendix C (BIS, 2005, p. 36; Löffler & Posch, 2007, p. 148).

Figure 13. CAP curve and distribution of rating scores for defaulters and non-defaulters



Source: BIS, *Studies on the Validation of Internal Rating Systems*, 2005, pp. 36–37.

The measure that condenses the whole information given by the CAP into a single number is the accuracy ratio (AR). It is calculated as the ratio between the area between the diagonal (α_R)[64]

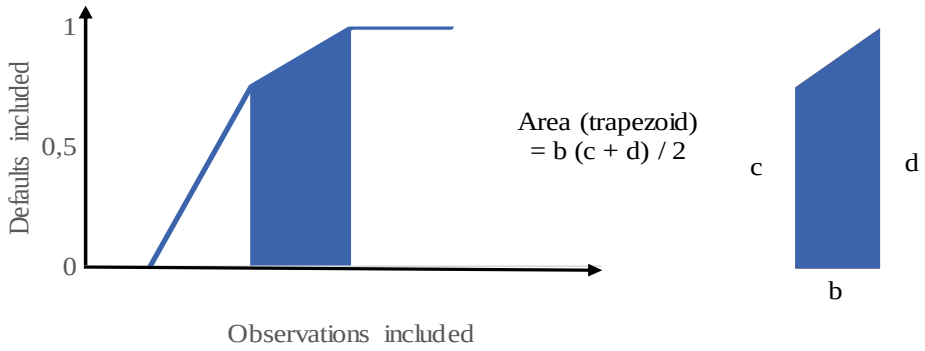
[64] The diagonal on the CAP curve corresponds to the rating system with no discrimination power. The AR for such a rating system would be 0. The AR can lie in the range between [-1,1].

and the CAP curve, versus the area for the "perfect" rating system (α_P) [65] as shown in Figure 13: $AR=\alpha_R/\alpha_P$.

Ideally we would like a scoring system that would successfully separate the distribution of defaulters from the distribution of non-defaulters. For real scoring systems perfect discrimination is not possible as shown by overlapping distributions in Figure 13. For a good rating system we obtain a concave CAP curve. On the other hand, non-optimal rating systems are non-concave.

ONB (2004, p. 109) notes that good rating systems based on logistic regression or machine learning algorithms in practice reach an AR between 70 % to 80 %. Löffler and Posch (2007, p. 150) show that the area under the CAP between two points can be treated as a trapezoid as shown in Figure 14. By using that fact they compute the AR ratio.

Figure 14. Trapezoid rule for calculating AR



Source: G. Löffler & G. Posch, *Credit Risk Modeling Using Excel and VBA*, 2007, p. 150.

An AR on the individual score (or on the rating grade) can also be calculated analytically based on comparing distributions as in Joseph (2005, p. 30):

$$AR=1-2\sum_i P_D(S_i)\left[\frac{F_{ND}(S_{i-1})+F_{ND}(S_i)}{2}\right] \quad (33)$$

where $F_{ND}(S_0)=0$. S_i (could also be S_r on the rating grade level calculation) is the score of the borrower or rating grade i, $P_D(S_i)$ is the probability of the bads (defaults) for score i, $F_{ND}(S_{i-1})$ is the cumulative probability of the goods (non-defaults) for score $(i-1)$ and $F_{ND}(S_i)$ is the cumulative probability of the goods (non-defaults) for score i.

8.2.1.2 ROC and AUC

A closely related concept is the Receiver Operating Characteristic (hereinafter: ROC). It is obtained by plotting the fraction of all defaults that occurred among borrowers rated r or worse

[65] A perfect rating model assigns the highest PDs to the actual defaulters.

(y-axis) against the fraction of all non-defaulters that are rated *r* or worse (x-axis). The two graphs differ in the definition of the x-axis. A summary measure for ROC is the area under the ROC curve (hereinafter: AUC or AUROC). AR and AUC have an exact linear relationship:

$$AR=2\cdot AUC-1 \quad \text{and} \quad AUC=\frac{AR+1}{2} \tag{34}$$

None of the two measures depends on the proportion of defaulters in the sample that is used for their estimation (BIS, 2005, p. 39). A model with greater discriminatory power has a larger AUC. Balthazar (2006, p. 142) defines that a good rating system would have an AUC measure of above 80 % and an acceptable system would have above 70 %.

One way of calculating an AUC is based on cut-off values. The value *C* in Figure 13 is the cut-off value that divides the borrowers into the ones that are predicted to default and those that are predicted to survive. This leads to four decision results that are shown in Figure 13. If the rating score is below *C* and the borrower actually defaults in the next period, then the decision was correct. Otherwise we have wrongly classified a non-defaulter as a defaulter. If the rating score is above *C* and the borrower did not default, then the classification was correct. Otherwise, a defaulter was incorrectly classified to the non-defaulters group (BIS, 2005, p. 37). To calculate an AUC based on this concept, BIS (2005, p. 37) defines the fraction of correctly classified defaulters as the hit rate *HR(C)*:

$$HR(C)=\frac{H(C)}{N_D} \tag{35}$$

where *H(C)* is the number of defaulters predicted correctly (with the cut-off value *C*) and N_D is the total number of defaulters in the sample. On the contrary, the false alarm rate *FAR(C)* is defined as:

$$FAR(C)=\frac{F(C)}{N_{ND}} \tag{36}$$

where *F(C)* is the number of false alarms (non-defaulters classified incorrectly as defaulters), N_{ND} is the number of actual non-defaulters in the sample. In Figure 13 *HR(C)* is the area to the left of the cut-off value *C* (coloured plus hatched area) under the score distribution of defaulters, while *FAR(C)* is the area to the left of *C* under the score distributions of the non-defaulters (chequered plus hatched area) (BIS, 2005, p. 37).

The ROC curve is a plot of *HR(C)* versus *FAR(C)* (we compute these two values for all cut-off values that are contained in the range of the rating scores). BIS (2005, p. 38) notes that in theory we could calculate an infinite number of pairs *HR(C)* versus *FAR(C)* (for all cut-off values). In practice we have a finite sample of points on the ROC curve, hence the other points are calculated by interpolating the set of points. Related to this, Wu (2008, p. 21) explains the errors that might happen. See Table 2.

Table 2. Types of errors

	Defaults	**Non-defaults**
Scores below C (high PD)	True positive prediction (Sensitivity)	False positive prediction (1-Specifity) (Type I error)
Scores above C (low PD)	False negative prediction (1-Sensitivity) (Type II error)	True negative prediction (Specificity)

Source: modified from X. Wu, *Credit Scoring Model Validation*, 2008, p. 21.

The proportion of correctly predicted defaults is called sensitivity (it corresponds to $HR(C)$) and the proportion of correctly predicted non-defaulters is called specificity. For a given cut-off value a good rating system would have a high sensitivity and specificity. A false positive prediction (corresponds to $FAR(C)$ or 1-specificity) is also called a type I error. This is an error of rejecting a null hypothesis that should have been accepted. On the other hand a false negative prediction is a type II error, which means accepting a null hypothesis that should be rejected (Wu, 2008, p. 22). The area below ROC (AUC) can then be calculated as:

$$AUC = \int_0^1 HR(FAR)d(FAR) \tag{37}$$

Another approach to calculate an AUC (similarly as for an AR)[66] on the individual score (or on the rating grade) can be done analytically as in Joseph (2005, p. 30):

$$AUC = 1 - \sum_i P_D(S_i)\left[\frac{F_{ND}(S_{i-1}) + F_{ND}(S_i)}{2}\right] \tag{38}$$

where $F_{ND}(S_0) = 0$. AUC can also be calculated numerically with trapezoidal sections summation as discussed above for the AR.

Löffler and Posch (2007, p. 155) remind that we must be careful when interpreting ARs and AUCs because they depend on the portfolio structure (it changes in time). There are in general no problems, if one compares AUCs of different rating systems within the same dataset.

One problem that might occur is presented in ONB (2004, p. 106). If one wants to compare two rating systems where the corresponding ROC curves for those systems intersect, it can happen that the AUC is the same for both. In such a case, the procedure which shows a steeper ROC curve in the lower range of scores on the left is to be preferred. Secondly Engelmann (in Engelmann & Rauhmeier, 2011, p. 279) provides a modified formula for an AUC for cases when the score discrimination is bimodal (if defaulters have either very high or very low score values - remember variable *growth in sales*). Hand (2009 in Lessmann et al., 2015, pp. 29–30) notices severe limitations of the AUC measure, because it does not treat all classifiers alike. He

[66] For derivation and interpretation of different statistical properties of AUC see Satschell and Xia (in Christodoulakis & Satschell, 2008, p. 118).

proposes an H-measure, which is gaining much interest in the academia in the last years. However, Lessmann et al. (2015, pp. 35, 39) found out that despite the theoretical advantages of the new measure, the overall rankings of borrowers are almost the same for both measures (they observed very high rank correlation while testing many different models).

Satschell and Xia (in Christodoulakis & Satschell, 2008, p. 120) derive analytical formulae for the AUC by assuming that the rating scores (produced by rating models) follow distributions other than normal distribution (i.e. for Weibull, logistic and mixed distribution). They test formulas with simulated data and obtain more accurate results compared to normal approximation. Once we compare two models in terms of AUC measures it is also possible to test, if the difference between the two is statistically significant. The test was developed by DeLong et al. (1988, in Engelmann & Rauhmeier, 2011, p. 286).

We need to note that ARs and AUCs are only estimates of a rating system's discriminatory power and as such are dependent on the underlying data. To calculate their standard errors and associated confidence intervals there exists an analytical equation based on normal approximation. For the derivation see Engelmann (in Engelmann & Rauhmeier, 2011, p. 282), Wu (2008, p. 24) and BIS (2005, p. 39). How wide the confidence interval is, depends on the value of α and on the values of N_D and N_{ND}. The lower the number of borrowers N or the higher the α (i.e. increase from 95 % to 99 %) the wider is the confidence interval[67]. Another way (although computationally inefficient) to calculate confidence intervals of a test statistic is the bootstrap method presented in Löffler and Posch (2007, p. 153).

8.2.1.3 The Kolmogorov-Smirnov (KS) test

The Pietra Index is another one-dimensional measure of discriminatory power. It is closely connected to the ROC curve concept. We will be more interested in the Kolmogorov-Smirnov nonparametric[68] statistic (KS), which is a special case of the Pietra Index. The one-sample version of the KS test can be used to compare the desired empirical cumulative distribution with the reference probability distribution. We are more interested in its two-sample version that allows us to compare two empirical distribution functions (of defaults and non-defaults). SAS (2015) notes it can also be used to determine the best cut-off in scoring. The best cut-off maximizes the KS, which becomes the best differentiator between the two populations.

ONB (2004, p. 106) provides several interpretations of the KS test. It can be derived from the ROC curve and interpreted as twice the area of the largest triangle that can be drawn between the diagonal and the ROC curve. If the ROC curve is concave, then the area of this triangle can also be calculated as the product of the diagonal's length and the largest distance between the ROC curve and the diagonal. It can also be interpreted as the maximum difference between the

[67] Satschell and Xia (in Christodoulakis & Satschell, 2008, p. 130) show that a small proportion of defaults in the sample also leads to wide confidence intervals.
[68] Note that to calculate the KS statistic we do not specify what are the underlying distributions. However, to run tests on the KS value, we need to use critical values that correspond to different distributions.

cumulative frequency distributions of good and bad cases. BIS (2005, p. 42) gives also another interpretation. They define it as the maximum difference between hit rate *HR(C*) and false alarm rate *FAR(C*) for a cut-off value.

First we show the calculations based on the difference of cumulative distributions of defaults and non-defaults. By F_{ND} and F_D we denote the cumulative distribution functions of non-defaulters and defaulters, respectively. The analytical formula for the Kolmogorov-Smirnov test is (Joseph, 2005, p. 29)[69] or as in BIS (2005, p. 42)[70]:

$$KS = \max_i \left(\widehat{F}_D(S_i) - \widehat{F}_{ND}(S_i)\right) = \max_c \left(\widehat{F}_D(c) - \widehat{F}_{ND}(c)\right) \qquad (39)$$
$$KS = \max_c |HR(c) - FAR(c)|$$

A KS test can take values from 0 to 1, where 1 means ideal discrimination. The formulas for cumulative distributions are as follows (Maarse, 2012, p. 26):

$$\widehat{F}_D(x) = \widehat{F}_X(x) = \frac{1}{N}\sum_{i=1}^{N} I_{\{X_i \le x\}} \qquad (40)$$
$$\widehat{F}_{ND}(x) = \widehat{F}_Y(x) = \frac{1}{M}\sum_{i=1}^{M} I_{\{Y_i \le x\}}$$

where X_i is the observation from sample X with $i=1,\ldots,N$ and Y_i is the observation from sample Y with $i=1,\ldots,M$. I is an indicator function.

The three assumptions for this test are given in Maarse (2012, p. 26). Firstly, the two samples must be independent and random. Their cumulative distribution functions must be continuous. Similarly as the AUC, the KS does not depend on the total portfolio probability of default. Therefore it may be estimated on samples with non-representative default/non-default proportions (BIS, 2005, p. 31).

It is convenient that one can calculate significance tests for the KS statistic. The significance of a rejection of the null hypothesis (the rating system has no more power than the random rating) with the Kolmogorov-Smirnov tests at (at i.e. 5 % level) could serve as a minimum requirement for rating systems (BIS, 2005, p. 31). ONB (2004, p. 107) describes the null hypothesis for the KS test for two independent samples as: "The score distributions of good and bad cases are identical" (i.e. the samples of both distributions are drawn from the same distribution). This hypothesis is rejected at level $1-\alpha$ if the KS test value equals or exceeds the following value (ONB, 2004, p. 107):

[69] For derivation see Clavero-Rasero (2006, pp. 8–9).

[70] The formula for the Pietra Index is given in BIS (2005, p. 42) as: $Pietra\ Index = \sqrt{2}/4 \cdot \max_c |HR(c) - FAR(c)|$. The Pietra Index can take values between 0 and 0,353. As a rating model's performance improves, the value is closer to 0,353 (KS is between 0 and 1).

$$D=\frac{D_{1-\alpha}}{\sqrt{N\cdot\bar{PD}(1-\bar{PD})}}=D_{1-\alpha}\sqrt{\frac{N_{ND}+N_D}{N_{ND}N_D}} \quad (41)$$

Where $D_{1-\alpha}=D_q$ is the value for significance levels q in the statistical table for Kolmogorov-Smirnov distribution. The table is in general valid when the sample size goes to infinity (must be at least of size 40). Then the KS test is Kolmogorov-Smirnov distributed (Maarse, 2012, p. 26). If the KS $\geq D$ we conclude that significant differences exist between the score values of good and bad cases.

In our case we determine the form of empirical cumulative distribution from the data, hence we must be careful with the determination of the critical values because critical values are calculated based on a distributional assumption[71]. There exist tables for some distributions (i.e. for normal distribution, exponential, Gumbel etc.). Anderson (2007, p. 196) notes that the KS critical values that are obtained for exponential and Weibull distribution are lower compared to normal distribution. We will not go into the details of deriving the Kolmogorov distribution that is the underlying concept for calculating critical values. However, it is important to note that there exist several tables with critical values that differ in the definition of the underlying distribution. Critical values for different distributions can be calculated by Monte Carlo techniques.

GraphPad (n.d.) interprets the p-values as the answer to the following question: "If the two samples were randomly sampled from identical populations, what is the probability that the two cumulative frequency distributions would be as far apart as observed?" If the p-value is small we can conclude that the two groups were sampled from populations with different distributions.

ONB (2004, 107) warns that this test gives only approximate results because it includes all kinds of differences between two distributions (not only differences between average rating scores for good and bad cases, but also differences in variance and other higher moments of the distribution). However, Mays (2004, in Anderson, 2007, p. 196) notes that the KS test is the most widely used statistic within the USA. In other countries AR or AUC measures seem to be preferred. According to the author very bad values for the KS are below 20 % and extremely good values are above 70 %.

8.2.1.4 Entropy measures (KL)

Entropy is a concept derived from information theory (a branch of computer science). It is related to the extent of uncertainty that is eliminated by an experiment (observing a borrower over time in order to decide his solvency status). The uncertainty of the solvency status is highest if the rating system has no discriminatory power (for example if all the rating grades have the

[71] It is possible to test the underlying normality assumption for instance with the R package "fitdistrplus", with Q-Q plots, histograms, CDF's and with Shapiro-Wilks test. There exist non-parametric approaches that do not require distributional assumptions. Techniques based on specific distributional assumptions (parametric) are usually more powerful.

same predicted PD). In this case the entropy measure has high value since the gain in information by finally observing the actual borrower's status will be large (BIS, 2005, p. 31).

The most common entropy measures are[72]: Conditional Entropy, Kullback-Leibler divergence (KL), CIER and Information Value (IV). In this thesis we present the KL divergence measure and use the approaches of Joseph (2005) and Quantiki (2015) to calculate it[73]. For implementing mutual entropy[74], IV or CIER see Joseph (2005), Anderson (2007), BIS (2005) or ONB (2004).

The calculation of the KL measure is based on segmenting default scores into groups. Groups are usually rating grades, however in Falkenstein et al. (2000, p. 82) they group borrowers into 50 quantile buckets and examine the default rate in the next year. Keenan and Sobehart (1999, pp. 10–11) use a bin counting approach.

In probability and information theory, the Kullback-Leibler divergence (KL) measures the difference between two probability distributions (defaults and non-defaults). It has similar properties as the Information value[75]. For two probability distributions of a discrete random variable (p and q) it is defined as in Quantiki (2015):

$$KL(p,q)=\sum_x p(x)\, log_2 \frac{p(x)}{q(x)}=E_{p(x)}\left[log_2 \frac{p(x)}{q(x)}\right] \tag{42}$$

If the logarithms are taken to the base 2, then the KL measure can be interpreted in units of bits, but the other properties of it hold irrespective of the logarithm base. The formula can be rewritten as (Quantiki, 2015):

$$KL(p,q)=-\underbrace{\sum_x p(x) ln\, q(x)}_{Cross\ entropy}+\underbrace{\sum_x p(x) ln\, p(x)}_{Entropy}=-H(p,q)+H(p) \tag{43}$$

By $H(p,q)$ we denote the cross entropy of p and q, and by $H(p)$ the entropy of p (this is also called the Shannon entropy[76]). The cross entropy is always greater than or equal to the entropy, hence the KL is always non-negative. If $p=q$ then $KL(p,q)=0$. It is a non-symmetric measure of difference, hence it holds that $KL(p,q)\neq KL(q,p)$. To calculate the KL divergence we follow the calculations of Joseph (2005, p. 32):

$$KL=\sum_i P_D(S_i)\cdot ln\left[\frac{P_D(S_i)}{P_{ND}(S_i)}\right] \quad \forall i \tag{44}$$

[72] Note that entropy measures are also used as a node impurity measure in constructing classification trees. However, there we use an alternative method, the Gini impurity measure.
[73] We test it with R package "Entropy" and obtain the same results.
[74] It is an interesting measure because it allows to quantify the dependence between any two models (it is a measure of how much information can be predicted about model B given the output of model A). See Keenan and Sobehart (1999, p. 11).
[75] It is actually a weighted sum of the difference between conditional default and conditional non-default rates.
[76] It is one of the most important metrics in information theory. It measures the uncertainty associated with a random variable or in other words the expected value of the information in the message which is measured in bits.

where $P_{ND}(S_i)>0$, $P_{ND}(S_i)$ is the probability of the goods (non-defaults) for bucket/rating grade *i* and *log* is the natural logarithm function. The rest was defined above under KS test.

8.2.1.5 Kendall-tau

Kendal's τ and Somer's *D* are rank order statistics and as such are nonparametric measures of association. They measure the degree of comonotonic dependence[77] of two random variables. We say that any pair of random variables with correlation 1 (i.e. any linearly dependent pair) is comonotonically dependent. Also if one of the variables can be expressed as increasing transformation of the other, then the two variables are comonotonic. This sort of dependence is considered as the strongest form of dependence of random variables (BIS, 2005, p. 45).

Kendal's τ is based on the relative ordering of ranks. Ranks will be higher, lower, or the same, if and only if the values are larger, smaller, or equal, respectively. It is similar to Spearman's correlation, but a bit harder to calculate. Kendall's τ values range between -1 and +1, with a positive correlation indicating that the ranks of both variables increase together. A negative correlation means that as the rank of one variable increases, the rank of the other variable decreases (Press, Teukolsky, Vetterling, & Flannery, 1992, p. 642).

If (x,y) are a pair of random variables, then BIS (2005, p. 45) defines the statistic as:

$$\tau_{xy}=P(x_1<x_2,y_1<y_2)+P(x_1>x_2,y_1>y_2)-P(x_1<x_2,y_1>y_2)-P(x_1>x_2,y_1<y_2) \tag{45}$$

To define τ Press, Teukolsky, Vetterling, & Flannery (1992, p. 642) start by defining the *n* data points (x_i,y_i). In our case these are the values for default events (binary variable) and the values for predicted PDs. In combinatorics there are $\binom{n}{2}=\frac{n(n-1)}{2}$ pairs of data points. This number consists of data points that cannot be paired with themselves and those where the points in either order count as one pair. Press et al. (1992, p. 643) define a pair concordant (*P*) if the relative ordering of the ranks of the two *x*'s is the same as the relative ordering of the ranks of two *y*'s. The pair is discordant (*Q*), if the relative ordering of the two *x*'s is opposite from the ordering of the two *y*'s. If there is a tie in either of the ranks of the two *x*'s or two *y*'s, then we call it a tie (and not *P* or *Q*). We count the ties separately for *x*'s and for *y*'s. Similarly PennState (n.d.) define that the observations (x_i,y_i) and (x_j,y_j) are concordant if they are in the same order with respect to each variable: $x_i<x_j$ and $y_i<y_j$ or if $x_i>x_j$ and $y_i>y_j$ for $i<j$. They are discordant, if they are in the reverse ordering as: $x_i<x_j$ and $y_i>y_j$ or if $x_i>x_j$ and $y_i<y_j$ for $i<j$. The two observations are tied if $x_i=x_j$ and/or $y_i=y_j$ for $i<j$.

[77] Linear dependence is expressed via (linear) correlation.

There are more versions of Kendal's τ. Because our data consists of ties, it is important to note that our implementation is actually the Kendal's τ-b. The formula for Kendal's τ-b is (R-tutor, n.d.; SAS, n.d.):

$$\tau_b = \frac{P-Q}{\sqrt{N_1}\sqrt{N_2}} = \frac{P-Q}{\sqrt{(T_0-T_1)(T_0-T_2)}} = \frac{\sum_{i<j}\left(sign(x_i-x_j)sign(y_i-y_j)\right)}{\sqrt{(T_0-T_1)(T_0-T_2)}} \tag{46}$$

where N_1 is the number of data pairs that are not tied in the first target feature (x), N_2 is the number of data pairs that are not tied in the second target feature (y), P is the number of concordant pairs, Q is the number of discordant pairs. We need to define also: $T_0 = \frac{n(n-1)}{2}$, $T_1 = \sum_k \frac{t_k(t_k-1)}{2}$ and $T_2 = \sum_l \frac{u_l(u_l-1)}{2}$ where t_k is the number of tied x values in the kth group of tied x values, u_l is the number of tied y values in the lth group of tied y values, n is number of observations and *sign(z)* is defined as (SAS, n.d.):

$$sign(z) = \begin{cases} 1 & if\ z>0 \\ 0 & if\ z=0 \\ -1 & if\ z<0 \end{cases} \tag{47}$$

We compute the statistic from a number of interchanges of the pairs and correct it for tied pairs (pairs of observations with equal values of x or equal values of y). In the empirical part we calculate the number of concordant, the number of discordant pairs and the number of ties with the help of the VBA code done by Joseph (2005) and the notes in Press et al. (1992, p. 643). The formula we implement is the one from R-tutor (n.d.) as stated above. In R we use the package "Kendall" that computes Kendal's τ-b that returns as a result also the p-values under the null hypothesis of no association (built-in several algorithms regarding the number of ties present in the data).

8.2.1.6 Brier score

The Brier score (Mean Square Error (hereinafter: MSE) test) actually measures both discrimination and calibration. The formula is as follows (Löffler & Posch, 2007, p. 156):

$$Brier\ score\ (MSE) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{PD_i})^2 \tag{48}$$

where i indexes the N borrowers. $\widehat{PD_i}$ is the predicted probability of default of the borrower i at the beginning of a period and y_i is an indicator variable at the end of a period that takes the value *1* if the borrower i defaulted in the period and *0* otherwise. To compute the Brier score, we need individual PDs, which we do not necessarily need for calculating the AUC. The Brier calculates squared prediction errors based on probability predictions and it lies in the range of [0,1]. The better rating system has lower score values (Löffler & Posch, 2007, p. 156).

The test measures the difference between default/non-default $y_i=(0,1)$ indicators and the corresponding predicted $\widehat{PD}_i$, that are averaged across all borrowers. The MSE gets small, if the predicted PDs assigned to defaulted firms are high and the ones assigned to non-defaults are low (Rauhmeier, in Engelmann & Rauhmeier, 2011, p. 322). The best that one can do is to predict each default probability perfectly: i.e. $PD_i=\widehat{PD}_i$ for each borrower *i*. In this case the expected value of the MSE is $E(MSE_{PD_i=\widehat{PD}_i})=\frac{1}{N}\sum_{i=1}^{N} PD_i(1-PD_i)$ (Rauhmeier & Scheule, 2005, p. 4).

Murphy and Winkler (1992, in Rauhmeier & Scheule, 2005, p. 5) propose a method for decomposing the MSE statistic into an over-all-calibration and variance-of-the-forecast-error components of the forecast error as follows:

$$MSE= \overbrace{(\bar{y}-\bar{\widehat{PD}})^2}^{\text{Over-all-calibration}} + \underbrace{\overbrace{s_y^2}^{\text{Uncertainty}} + \overbrace{s_{\widehat{PD}}^2}^{\text{Refinement}} -2s_y s_{\widehat{PD}} \overbrace{r_{y\widehat{PD}}}^{\text{Association}}}_{\text{Variance of forecast error}} \tag{49}$$

The over-all-calibration measures the fit between the default rate $\bar{y}=\frac{1}{N}\sum_{i=1}^{N} y_i$ and the mean of the predicted PDs: $\overline{\widehat{PD}}=\frac{1}{N}\sum_{i=1}^{N}\widehat{PD}_i$. The expression $\bar{y}-\overline{\widehat{PD}}$ is also known as Mean Error. The last three terms form the variance of the forecast error. It is comprised of the variance of the default events s_y^2 (Uncertainty), the variance of predicted PDs $s_{\widehat{PD}}^2$ (Refinement) and the covariance between the two variables $r_{y\widehat{PD}}$[78]. Good rating models with low MSE have a small variance over-all-calibration and variance-of-the-forecast-error. The Uncertainty cannot be influenced by the modeller (Rauhmeier & Scheule, 2005, p. 5). Rauhmeier and Scheule (2005, p. 5) show also another decomposition for the MSE. Now it reveals the properties of Refinement and Discrimination.

$$MSE= \overbrace{s_{\widehat{PD}}^2}^{\text{Refinement}} + \overbrace{\sum_y \frac{N_y}{N}(\bar{\widehat{PD}}_y-\bar{y})^2}^{\text{Discrimination I}} - \overbrace{\sum_y \frac{N_y}{N}(\bar{\widehat{PD}}_y-\bar{\widehat{PD}})^2}^{\text{Discrimination II}} \tag{50}$$

To minimize the MSE we need that the average predicted PDs for all defaults $\overline{\widehat{PD}_{y=1}}$ are close to one and for all non-defaults $\overline{\widehat{PD}_{y=0}}$ close to 0 (Discrimination I). The predicted PDs for defaulters and non-defaulters should also be far away from the average predicted PD for all borrowers (Discrimination II). In general, Discrimination is important in relation to bank's exposure to idiosyncratic credit risk (individual borrowers), over-all-calibration is especially important for homogenous, infinitely granular (e.g. large retail) portfolios (Rauhmeier & Scheule, 2005, p. 14).

[78] This is the Bravais-Pearson-Correlation Coefficient. The formula can be found in Joseph (2005, p. 17).

8.2.2 Calibration tests

The regulators require banks to compare the estimated PDs with the observed default rates. DB (2003, p. 63) notes that the validators need to determine whether the deviations are purely random or they occur systematically and the predicted PDs underestimate the true risks. Calibration is usually performed on pooled PDs. Given the rating structure, the calibration tests (those on rating grade level) tell us how the average expected PD per rating grade fits the observed default experience for this rating grade. Rauhmeier (in Engelmann & Rauhmeier, 2011, p. 320) describes that the calibration assures that each forecast probability of default is correct $PD_{rt} = \widehat{PD_{rt}}$. However, in practice the realisation $y_i=1$ or $y_i=0$ is the result of a random process. Even, if the parameters of the model β are correctly specified, some unpredictable randomness ε_i remains. The one-period calibration tests are based on individual or rating grade predicted PDs ($\widehat{PD_i}$) and realized individual default events (y_i) in the next year. Calibration tests are mostly based on the concept of p-values. This is the probability of making an error when rejecting the null hypothesis, which in our case is that the model is adequately calibrated. If the p-value is less than 5% then we reject the null hypothesis (we reject the model calibration). The p-value is the probability of obtaining an effect at least as extreme as the one in your sample data, assuming the truth of the null hypothesis.

Hence, Balthazar (2006, p. 182) is of opinion that validation of calibration is a tough exercise, as default rates are usually very low[79] and highly volatile. There might be even issues with the small size of the sample[80]. Even BIS (2005, p. 3) notes that methods for validating calibration are at a much earlier stage compared to discriminatory power tests. A major issue is how to incorporate the correlations between defaults. BIS warns that the default correlation causes that the observed default rates can systematically exceed the critical PD values, if these are determined under the assumption of independence of the default events, which is usually the case in calibration tests. Tests based on the independence assumption are sometimes too conservative. Hence, even well-calibrated rating systems may perform poorly in these tests. BIS (2005, p. 3) is of the opinion that statistical tests alone are insufficient to adequately validate an internal rating system. A simple graphical tool to test calibration is the calibration plot (we have discussed it already above), where one plots predicted PDs versus observed DRs on the x- and y-axis respectively. The most common statistical calibration tests are:

- Binomial test (standard binomial, normal approximation, normal with incorporated correlations – i.e. one-factor model),
- Chi-square test (Hosmer-Lemeshow test, hereinafter: HSLS),
- Normal test,
- Traffic lights approach (basic traffic lights, granularity adjustment, extended traffic lights approach - ETLA),

[79] This is even of more concern in the low default segments. Such models must be validated with a different procedure as discussed in Blochwitz et al. (in Engelmann & Rauhmeier, 2011, p. 306).
[80] Balthazar (2006, p. 184) discusses the use of Monte Carlo if we have a too small sample.

- Spiegelhalter test (hereinafter: SPGH).

The distinctions between statistical tests presented in this section go into three directions. Blochwitz et al. (in Engelmann & Rauhmeier, 2011, p. 296) distinguish between those that can be classified either by validation period (single- versus multi-period tests) or by number of rating grades (single- versus multi-grade tests). Another important distinction is whether they need the default independence assumption or not. ONB (2004, p. 119) reports that empirical studies show that default events are generally not uncorrelated. Typically default correlations range between 0,5 % and 3 %. Their effect is on strengthening of fluctuations in default probabilities, hence tests with the assumption of independence reject model calibrations more often (are more conservative).

The example of a single-grade single-time-period[81] is the binomial test and its normal approximation. It is also the most commonly used test in practice. Under the assumption of independence of default events it tests the calibration of a one period predicted PDs. With only one data point it is difficult to develop robust validation technique. The next step is to use tests for several rating grades. The Hosmer-Lemeshow test (hereinafter: HSLS) is an example of a single-period, multi-grade test and tests the adequacy of PDs for several rating grades simultaneously.

Also multi-period tests exits. Examples are the normal test and traffic lights tests. Blochwitz et al. (in Engelmann & Rauhmeier, 2011, p. 308) warn that they require at least 4 years of data. Unfortunately, in 4 years time a rating system might have gone through some revisions or has been replaced. The normal test is applied under the assumption of a stable default rate over time and that default events in different years are independent (inter-temporal independence). The normal test is based on a normal approximation of the distribution of time-averaged default rates. It allows for cross-sectional dependence (BIS, 2005, p. 34). Another multi-period, single-grade test is the traffic lights approach. BIS (2005, p. 34) notes that it is well accepted among regulators. EBA (2013b, p. 80) reports that most of large EU banks use this approach. In contrast to the normal test it is independent of the assumption of constant PDs over time. It is based on the normal approximation and assumes cross-sectional and inter-temporal independence of default events (BIS, 2005, p. 34).

Comparing the differences between predicted PDs and realised default rates is usually done simultaneously for all rating grades, or separately for each rating grade. However, if one does not want to test bucket PDs it is also possible to test the calibration of individual PDs with the multi-grade SPGH test.

[81] The usual starting point for single-period tests is to obtain data about one and a half year after the first implementation of a rating system.

Table 3. Calibration tests

Test	Multi-grade	Multi-period	Buckets	One-factor	Assumptions
Binomial test	No	No	Yes	No	binomial distribution; independence of default events
Normal approximation	No	No	Yes	No	normal approximation; independence of default events
One-factor binomial test	No	No	Yes	Yes	normal approximation; one factor dependence structure
Chi-square HSLS	Yes	No	Yes	No	normal approximation; independence of default events
Normal test	No	Yes	Yes	-->	normal approximation; assume independence of defaults across different years (inter-temporal independence of default events) and stable mean default rates
Granularity (traffic lights)	No	Yes	Yes	Yes	normal approximation; one factor dependence structure
ETLA (traffic lights)	No	Yes	Yes	-->	normal approximation; assumption of cross-sectional and inter-temporal independence of default events; deals with correlations through volatility
SPGH	Yes	No	No	No	normal approximation; independence of default events

There are also possible extensions for SPGH and HSLS tests, if one wants to incorporate correlation effects. Rauhmeier (in Engelmann & Rauhmeier, 2011, p. 327) proposes the Monte Carlo simulation method to implement circumstances of correlated defaults[82] and to substitute the theoretical test statistic (e.g. normal distribution in the case of the SPGH test), by one obtained by the simulation approach. Test decisions (i.e. p-values) are then based on the "new" simulated distribution of the test statistic.

Now we discuss the theoretical properties of calibration tests. In the empirical part we implement binomial tests, one-factor test, HSLS and SPGH test. Practical calculations of these tests on hypothetical data are done in Appendix E.

8.2.2.1 Binomial tests (normal approximation and one-factor binomial test)

a) Binomial test

With the binomial test we can test the hypothesis whether the PDs predicted by a model are consistent with observed defaults. The test can be applied only to one single rating grade over a single time period (usually 1 year). It is based on the assumption of independent default events between all borrowers (N_{rt}) within the chosen rating grade $r \in \{1,....,R\}$ in year t. The number of defaults (D_{rt}) in rating grade r in year t can be modelled as a binomially distributed ($D_{rt} \sim Bin(N_{rt}PD_{rt} ; N_{rt}PD_{rt}(1-PD_{rt}))$) random variable X with size parameter N_{rt} and success probability PD_{rt} (this is the proxy for a true default rate that is based on the historical default rate $DR_{rt} \approx \overline{PD_{rt}} = \frac{D_{rt}}{N_{rt}}$), which is assumed to be the same for all borrowers within one rating grade.

[82] It can also be used for small sample corrections.

Blochwitz et al. (in Engelmann & Rauhmeier, 2011, p. 297) define the null hypothesis for rating grade *r* in one period as (one sided test):

$\boldsymbol{H_0}$: $PD_{rt} \leq \widehat{PD_{rt}}$ (the predicted $\widehat{PD_{rt}}$ of the rating category *r* is conservative enough (i.e. the actual default rate PD_{rt} is (less than or) equal to the predicted default rate $\widehat{PD_{rt}}$ by the PD model))
$\boldsymbol{H_1}$: $PD_{rt} > \widehat{PD_{rt}}$ (the predicted PD underestimates the true probability of default)

The null hypothesis is rejected at a confidence level α (95 %), if the number of historically observed defaults D_{rt} in this rating grade is greater than or equal to the critical value c_α. Rejection means that our model underestimates the true probability of default. To define the critical value (c_α) we need to know the distribution of observed defaults D_{rt} in rating grade *r* in year *t*. Rauhmeier (in Engelmann & Rauhmeier, 2011, p. 320) defines it by starting with the definition of the binomial distribution under the independence assumption [83]:

$$P(D_{rt}|N_{rt},PD_{rt}) = \binom{N_{rt}}{D_{rt}} PD_{rt}^{D_{rt}}(1-PD_{rt}^{D_{rt}})^{N_{rt}-D_{rt}} \tag{51}$$

Where PD_{rt} is the default probability estimated at the start of year *t* (proxy of true default probability), D_{rt} is the actual number of defaults in year *t* and grade *r*. N_{rt} is the number of borrowers at the date of setting the model into operation. Knowing the distribution of D_{rt} under H_0 we can calculate critical values with the cumulative binomial distribution as defined in Blochwitz et al. (in Engelmann & Rauhmeier, 2011, p. 297):

$$c_\alpha = min\left\{D_{rt} \ : \ \sum_{i=D_{rt}}^{N_{rt}} \binom{N_{rt}}{i} PD_{rt}^{i}(1-PD_{rt}^{i})^{N_{rt}-i} \leq \alpha\right\} \tag{52}$$

We want to find the minimum value of D_{rt} for which the summation term from above is less than α.

b) Normal approximation

In practice the normal approximation of the exact binomial test is often used. This is due to the convergence of the discrete binomial distribution to the normal distribution for increasing sample sizes (can be shown by applying the central limit theorem). As a rule of thumb this is possible when two inequalities hold: $(N_{rt} \cdot PD_{rt} \geq 10)$ and $(N_{rt} \cdot PD_{rt} \cdot (1-PD_{rt}) \geq 10)$. Then the number of defaults within one rating grade is normally distributed $D_{rt} \sim N(N_{rt}PD_{rt}\ ;\ N_{rt}PD_{rt}(1-PD_{rt}))$ and the *z* test statistic (with standard normal distribution) can be constructed (Rauhmeier in Engelmann & Rauhmeier, 2011, p. 322):

$$z = \frac{D_{rt}-N_{rt}PD_{rt}}{\sqrt{N_{rt}PD_{rt}(1-PD_{rt})}} \sim N(0,1) \tag{53}$$

[83] $\binom{N_{rt}}{D_{rt}} = \frac{N_{rt}!}{D_{rt}!(N_{rt}-D_{rt})!}$, where $N_{rt}! = 1 \times 2 \times \ldots \times N_{rt}$.

The critical value would be defined similarly as for the binomial test:

$$c_\alpha = min\{D_{rt} \ : \Phi(z) \leq \alpha\} \tag{54}$$

c) One-factor binomial test

In practice the independence assumption is violated, since the default correlations are normally within the range of 0–3 %, as stated by BIS (2005, p. 47). A popular modification of the binomial model to incorporate default correlations is the one-factor model, which is used also as a basis for the calculations of the minimum regulatory capital charge in the IRB approach. ONB (0024, p. 122) and Blochwitz et al. (in Engelmann & Rauhmeier, 2011, p. 298) note that from a conservative risk assessment point of view, overestimating significance is not critical, so we can easily operate under the assumption of uncorrelated defaults. However, there might be negative effects for the stability of the model, because one would have to recalibrate the model very frequently.

There are in general two reasons why the observed default rate in year *t* is larger than the underlying default probability. So far we have discussed only the possibility that borrowers default due to individual "bad luck". However, it may be the case that year *t* was generally a "bad year" for all borrowers (i.e. take the correlations into account). The idea behind the one-factor model is to model the equation for a common/systematic risk factor *F* that corresponds to the market condition in a certain year. We then compare the predicted PDs to this statistic. The default correlations δ can be modelled through correlations in asset values ρ (A_i is asset value of borrower *i)*. The formula is: $A_i = \sqrt{1-\rho}\varepsilon_i + \sqrt{\rho}F$. The systematic factor F is common to all borrowers and as such it can capture the "bad year" definition, however the factor ε_i is specific to the borrower and captures "individual bad luck". We say that our predicted PD underestimates the true default probability, if it is too unlikely to assume that the year under consideration was extremely bad. Löffler and Posch (2007, p. 158) define a one-factor model[84] by setting the average default probability to our prediction PD_{rt}, and the default probability $\overline{PD_{rt}}$ equal to the default rate in year *t*.

$$\overline{PD_{rt}} = \frac{D_{rt}}{N_{rt}} = \Phi\left[\frac{\Phi^{-1}(PD_{rt}) - \sqrt{\rho}F_t}{\sqrt{1-\rho}}\right] \tag{55}$$

For detailed derivation see the Appendix F. If we solve for the factor F_t we get:

$$F_t = \frac{\Phi^{-1}(PD_{rt}) - \sqrt{1-\rho}\,\Phi^{-1}\left(\frac{D_{rt}}{N_{rt}}\right)}{\sqrt{\rho}} \tag{56}$$

[84] If the factor F is very negative (bad year) then asset correlations ρ would increase the adverse effect and increase the default rate $\overline{PD_{rt}}$ in year t.

From the equation we see what kind of year we need, if we want to bring the PD and the default rate in line. A negative F_t (the more negative it is, the more of a "bad year" we have) will push the default rate above the predicted PD. The probability of observing a year as bad as year t or worse is $\Phi(F_t)$ (we assume that F_t is the standard normally distributed). Once we know the value of F, the default event of borrower i provides no information on how likely another borrower is to default. This is because the correlations are now modelled in the systematic factor and the randomness is now entirely due to the term ε_i (we have assumed in the model that ε_i and ε_j are independent for $i \neq j$). Conditional on the factor realization, defaults are now independent. At significance level α, we thus reject the H_0 if (Löffler & Posch, 2007, p. 159):

$$\Phi[F_t] \leq 1-\alpha \tag{57}$$

If equation $\Phi(F_t) \leq 1-\alpha$ holds, then we conclude that the predicted PD was too low (the scenario F_t is too extreme at our significance level). Similarly as before we can define the critical value c_α as:

$$c_\alpha = min\{D_{rt} \ : \Phi(F_t) > 1-\alpha\} \tag{58}$$

Intuitively, we expect higher default rates in the one-factor level when compared to the binomial model in recessionary times. The critical values of correlation-modified PD tests tend to be much greater than the critical values of the (independence-based) binomial test, hence it is harder to reject the null hypothesis in the one-factor model. The independence based binomial test is very conservative (we reject the null hypothesis too early). We will not discuss here how to properly estimate default correlations, because we possess a too short time series of defaults[85]. We note that a valid estimation is not an easy task. For estimating default correlations with the Asset value approach see Löffler and Posch (2007, pp. 103–118).

8.2.2.2 Chi-square (Hosmer Lemeshow, HSLS) test

The drawback of the binomial test is that it is a single-period validation test for one rating grade at a time. The two-sided Hosmer-Lemeshow Chi square test is an alternative approach that allows us to simultaneously compare all rating grades. Intuitively the test compares the expected number of defaults per rating grade with the number of realised defaults in this rating grade. The HSLS test then summarizes this information for all rating grades.

The test statistic is based on the assumption that the predicted default probabilities $\widehat{PD}_r$ and the observed default rates $\overline{PD}_r$ are identically distributed. In addition, we assume that all default events within each of the different rating grades as well as between all rating grades are independent (Blochwitz et al. in Engelmann & Rauhmeier, 2011, p. 300).

[85] For instance, Löffler and Posch operate with 25 years of data.

We apply it to the rating grade structure. Rauhmeier (in Engelmann & Rauhmeier, 2011, p. 324) and Balthazar (2006, p. 138) comment that it is often applied to buckets, where the buckets (*b*)[86] are constructed by arranging individual predicted PDs into e.g. ten centiles (ten buckets). The test groups data into several intervals and compares the observed default with those predicted by the model (average predicted PD per bucket). The null hypothesis for our test statistic is: *H0:* $PD_{rt}= \widehat{PD_{rt}}$. The Hosmer-Lemeshow Chi square test statistic for a given year *t* is defined as in Rauhmeier (in Engelmann & Rauhmeier, 2011, p. 324):

$$\chi_t^2 = \sum_{r=1}^{R} \frac{(N_r\overline{PD_r}-N_r\widehat{PD_r})^2}{N_r\widehat{PD_r}(1-\widehat{PD_r})} = \sum_{r=1}^{R} N_r \frac{(\overline{PD_r}-\widehat{PD_r})^2}{\widehat{PD_r}(1-\widehat{PD_r})} \tag{59}$$

Or as in BIS (2005, p. 52):

$$\chi_t^2 = \sum_{r=1}^{R} \frac{(D_r-N_r\widehat{PD_r})^2}{N_r\widehat{PD_r}(1-\widehat{PD_r})} \tag{60}$$

Where $D_r=N_r\overline{PD_r}$ is the number of actual defaults in rating grade *r*, N_r is the number of borrowers in risk grade *r*, $\widehat{PD_r}$ is the predicted average probability of default and $\overline{PD_r}$ is the actual default rate in the corresponding bucket. By the central limit theorem, when $N_r \to \infty$ simultaneously for all borrowers, then the distribution of χ_{HL}^2 converges to a chi-square χ_R^2 distribution with *R* degrees of freedom. χ_{HL}^2 consists in fact of *R* independent squared standard normal distributed random variables. It was proposed (based on Hosmer-Lemeshow simulations) to use *R-2* degrees of freedom (Rauhmeier in Engelmann & Rauhmeier, 2011, p. 325).

To assess the goodness of fit of the calibrated model we can calculate the Chi-probability of our χ_R^2 test. The closer the Chi-probability is to zero (large chi-square values), the worse is our prediction/calibration. We would normally categorize a prediction as satisfactory, if the Chi-probability is greater than 5 %. We need to repeat the test for every year *t*.

The test has a few drawbacks. First the buckets (rating grades) must be reasonably large to get good approximations[87]. Secondly, the independence between borrowers can be assumed only, if we test a PIT model (it is not useful for TTC models) (Blochwitz et al. in Engelmann & Rauhmeier, 2011, p. 300; Maarse, 2012, p. 43). Additionally Allison (2013) emphasises that the HSLS statistic is very sensitive to the choice of number of buckets[88]. Rauhmeier (in Engelmann

[86] For instance, Hayden (in Engelmann & Rauhmeier, 2011, p. 23) uses the approach with buckets. PDs are grouped based on quantiles (10 % intervals were used).

[87] The number of observations is large enough if $N_r\widehat{PD_r} \geq 10$ and $N_r\widehat{PD_r}(1-\widehat{PD_r}) \geq 10$. This corresponds to approximating a binomial distribution with the normal distribution. BIS (2005, p. 52) further emphasises that the convergence is low if the PDs are very small.

[88] Using less buckets gives less opportunity to detect mis-specification. On the other hand, using too many buckets causes a low number of observations in each bucket, hence it is difficult to determine whether differences between observed and expected PDs are random or they indicate a model's mis-specification. Some sensitivity analysis is recommended (Bartlett, 2014).

& Rauhmeier, 2011, p. 325) notes that it can be shown that, when there is just one rating grade, the HSLS test, the (squared) SPGH test and the (squared) approximated binomial test statistic are identical.

8.2.2.3 Spiegelhalter (SPGH) test

The other test that allows us to simultaneously test several rating grades is the SPGH test (it is a multi-grade test). However it does not need bucket (grade) segmentation and can be applied on individual PDs. The SPGH test is based on the Brier Score (Mean Square Error, hereinafter: MSE):

$$MSE=Brier\ score=\frac{1}{N}\sum_{i=1}^{N}(y_i-\widehat{PD_i})^2 \tag{61}$$

The SPGH test tests whether an observed MSE is significantly different from its expected (predicted) value or not. Under the null hypothesis (H_0: $PD_i=\widehat{PD_i}$), the MSE has an expected value of (Rauhmeier in Engelmann & Rauhmeier, 2011, p. 323):

$$E\left(MSE_{PD_i=\widehat{PD_i}}\right)=\frac{1}{N}\sum_{i=1}^{N}PD_i(1-PD_i) \tag{62}$$

and a variance:

$$Var\left(MSE_{PD_i=\widehat{PD_i}}\right)=\frac{1}{N^2}\sum_{i=1}^{N}(1-2PD_i)^2PD_i(1-PD_i) \tag{63}$$

The absolute value of the MSE does not have a meaningful interpretation because its value depends not only on the quality of the rating system, but also on true but unknown default probabilities (Rauhmeier in Engelmann & Rauhmeier, 2011, p. 323). Using the central limit theorem, it can be shown that under the null hypothesis the test statistic is defined as (Rauhmeier in Engelmann & Rauhmeier, 2011, p. 323):

$$Z_{SPGH}=\frac{MSE-E\left(MSE_{PD_i=\widehat{PD_i}}\right)}{\sqrt{Var\left(MSE_{PD_i=\widehat{PD_i}}\right)}}=\frac{\frac{1}{N}\sum_{i=1}^{N}(y_i-PD_i)^2-\frac{1}{N}\sum_{i=1}^{N}PD_i(1-PD_i)}{\sqrt{\frac{1}{N^2}\sum_{i=1}^{N}(1-2PD_i)^2PD_i(1-PD_i)}} \tag{64}$$

Asymptotically it follows a standard normal distribution (by the central limit theorem). As a special case of the SPGH statistic, if there is just one single PD in the entire portfolio (all borrowers are in the same rating grade), then the SPGH statistic is exactly equal to the statistic of the approximated binomial test (equation (55) for normal binomial)[89].

The main advantage of the SPGH test over the binomial test is that it can test all grades simultaneously. Secondly, many other tests require averaging the predicted PDs (which are individually calculated) across rating grades or across years. Wu (2008, p. 28) notes that this

[89] The derivation is in Rauhmeier (2011, in Engelmann & Rauhmeier, 2011, p. 343).

might introduce some bias. The advantage of the SPGH test is that it can be performed on individual PDs that are not averaged. Maarse (2012, p. 44) emphasises that the test has also two drawbacks. First it assumes independence between defaults (valid only for PIT models). Second, the test cancels out borrowers with a too high predicted PD against buckets with a too low predicted PD.

Additionally, Redelmeier et al. (1991, in Rauhmeier & Scheule, 2005, p. 13) proposed a test whether the MSEs of two models are significantly different or not (can be used for models that have already passed the SPGH test).

8.2.3 Stability

If the model's results in the validation sample (and the sample of new data after the model was implemented) are much worse compared to the estimation sample, then such models might not be appropriate, because they fail when applied to new data. ONB (2004, p. 81) distinguishes between the stability of the discriminatory power when applied to a validation sample and the stability when applied to longer forecasting horizons (not only 1 year). In addition ONB (2004, p. 127) notes that there are two aspects of stability:

- stability of the discriminatory power of a rating model for different prediction time horizons (and stability when loans become older[90]),
- stability in the general conditions underlying the use of the model (effects on individual model parameters).

Since rating models are calculated for a period of 1 year, their discriminatory power decreases for longer time horizons. The purpose of stability testing is to ensure that their discriminatory power deteriorates only slowly. ONB (2004, p. 127) requires that models demonstrate sufficient discriminatory power over forecasting horizons of three or more years. In stability testing one reviews also the developments in the external (economic, political, or legal)[91] and internal (business strategies, changes in market segments, organizational structure) environment of a bank and their influence on the rating model. These changes can cause systematic changes in parameter estimates or in the default rates. The bank may also decide to develop a new rating model, if it is expected that a potential change in general conditions would impact the current model considerably (ONB, 2004, p. 127). The stability is connected to the concept of on-going validation, which is our next point of interest.

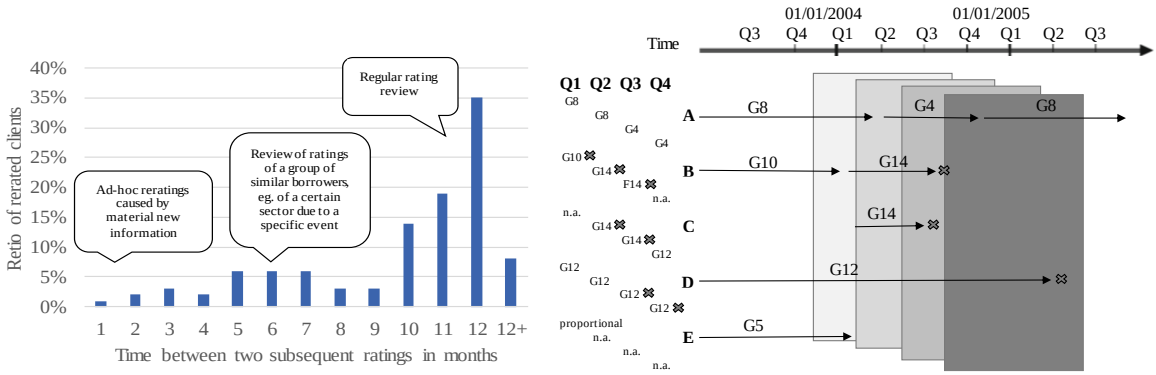
8.3 On-going-validation

First we describe what it means that models are continuously monitored. Then we will present a method that allows us to perform an on-going model validation. Blochwitz and Hohl (in

[90] To remedy this issue, the practioners use behavior scoring models which evaluate more recent information from the development of the credit transaction.
[91] For instance: changes in accounting standards, in bankruptcy procedures, country ratings etc.

Engelmann & Rauhmeier, 2011, p. 259) note that rating systems are ideally monitored continuously. Ratings must be updated at least annually or when new important information comes in. The expected pattern is shown in Figure 15. Most borrowers are revaluated regularly once every 12 months. Sometimes it might happen that ratings change due to some new information that has some systematic affect on part of the portfolio.

Figure 15. Data patterns and preparation for monthly on-going monitoring



Source: S. Blochwitz & S. Hohl (in B. Engelmann & R. Rauhmeier, *The Basel II Risk Parameters*, 2011, p. 259); R. Rauhmeier (in B. Engelmann & R. Rauhmeier, *The Basel II Risk Parameters*, 2011, p. 337).

Rauhmeier (in Engelmann & Rauhmeier, 2011, p. 336) presents the approach that is commonly used for on-going validation of models. He calls it "cutting slices" or "rolling 12-months-window" approach. It is shown on the right side of Figure 15. We start with the slice Q1-2004, which begins in January 2004. For validation we can use the whole period up to the end of Q4 in the year 2004. In this sample we would have borrowers A (good, grade 8), B (bad, because he defaulted in the period, enters with grade 10) and D (good, grade 12). The borrower E repaid in the period after this slice, hence he must be weighted as a non-default with the number of months that were observed in this slice[92] (good, grade 5). We do not include borrower C because he was not yet present in January 2004.

An interesting concept to address stability issues is a population stability index presented in Anderson (2008, pp. 194, 485). It is based on the KL measure and measures the difference between the development sample and a more recent distribution.

9 EMPIRICAL RESULTS

In the empirical part we apply all steps from the flowchart in Appendix A. We concentrate on building a PIT rating model for corporate exposures. From point 5 onwards (from our flowchart) we continue only with our logistic regression model that is developed up to point 5. In addition, we also compare its out-of-sample results to the results of neural networks and random forest methods.

[92] (number of months it has been observed)/12

The data is prepared with the structured query programming language SQL from the bank's datawarehouse. The validation is conducted mostly in Excel and its programming language VBA. The main part, which is statistical modelling, is performed using the statistical programming language R.

9.1 Data (Ch. 2)

First we created a catalogue[93] of 133 variables (Appendix G). Our data of corporate borrowers (private companies and limited liability companies) was retrieved from a small local bank for the period from 1. 1. 2010 to 30. 4. 2015 and is in line with its definition of default (filed bankruptcy, compulsory settlements and 90 days overdue). We excluded small companies that have less than 1.000 EUR assets or less than 1.000 EUR total sales. In addition, we excluded sectors with specific financial statements, such as the financial and the public sector. We have programmed an SQL procedure for automatic data extraction.

The data for non-defaulted companies at the start of each year consisted of factor variables, financial ratios and the information whether the default event occurred in the following year for these companies. Because we used growth variables, we excluded the companies that did not report financial statements in the last two years[94].

Due to the bank's privacy policy we are not disclosing the actual data. We report only the results of statistical tests. With regards to data preparation, we performed all required steps:

- design of financial ratios (deal with zero or small denominator, fraction issues etc.),
- checked descriptive statistics (default rate for missing values and non-missing, min, max, median for defaulters and non-defaulters, etc.),
- treatment of outliers with the truncation/winsorisation procedure (truncation at 2 %),
- removed the variable, if it had more than 15 % of missing values (we removed one variable),
- replaced missing values with the medians of defaulters and non-defaulters respectively[95],
- transformation of the data to assure the theoretical relationship between dependent and independent variables – we used two main transformations: HP and MINMAX transformation.

For the HP transformation we categorized the variables in 8 quantiles[96] and replaced the original values with the empirical odds of the respective quantile[97]. Then we applied the Hodrick-

[93] We used similar ratios as Hayden (2002), Altman and Sabato (2007), Cipollini et al. (2009), Jovan (2005) and also Ortl and Gorenjec (2015).
[94] We have noticed that few companies do not report financial statements in the year prior to distress. This might lead to some bias. In addititon, note that we dealt only with data for companies that were granted a loan from the bank. The borrowers that were rejected are not part of the database, which is another bias.
[95] In the portfolio to which we applied the PD estimates we replaced the missing values with the overall medians.
[96] Having around 8 groups turned out to return the best results. We have also run some tests with more groups, but the database is too small for many groups and the results are therefore worse.
[97] This transformation is used and proposed by many authors: Jovan (2005), ONB (2004, p. 79), Lőffler and Posch (2007) and Hayden (2002).

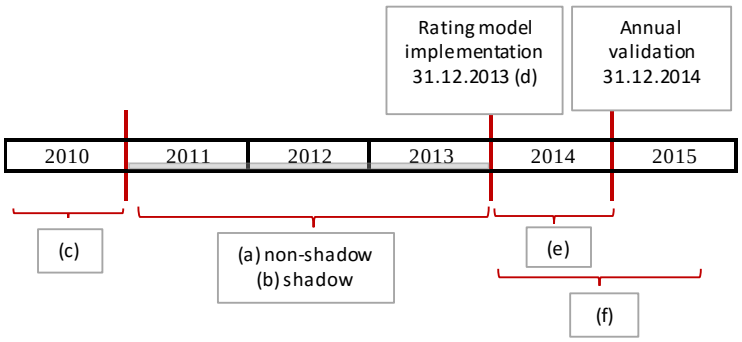
Prescott filter (HP) with the parameter $\lambda=100$ and smoothed the data. For the MINMAX transformation we first winsorized the data (cap with 2 %) and then scaled the data to the [0,1] interval with the MINMAX transformation.

Note that among the listed 132 variables there are 112 variables, which are appropriate to use for all data transformations. The transformations where we added the most negative value of the accounting variable in the denominator to the denominator itself (we did this to avoid the problem of having negative values in the denominator), can be used only with the HP transformation and not with raw data or with the MINMAX transformation.

Table 4. Sample formation

Stage	Data sample definitions	Bucket or individual level	Measures used for validation
Model development	(a) In-sample (estimation) [Ch. 5]	individual (bucket for KL test)	Discrimination: AR, AUC, KS, KL, Kendall-tau, Brier
	(b) Out-of-sample (validation) [Ch. 5]	individual (bucket for KL test)	Discrimination: AR, AUC, KS, KL, Kendall-tau, Brier
	(c) Out-of-time (validation) [Ch. 5]	individual (bucket for KL test)	Discrimination: AR, AUC, KS, KL, Kendall-tau, Brier
Implementation	(d) Assign PDs to non-defaulters at implementation [Ch. 8]	rating grade (individual for Brier & Tau)	Discrimination: AR, AUC, KS, KL, Kendall-tau, Brier
Validation and monitoring	(e) Validation-sample (validation after 1 year) [Ch. 9]	rating grade (individual for Brier & Tau) rating grade (individual for SPGH)	Discrimination: AR, AUC, KS, KL, Kendall-tau, Brier Calibration: Binomial (also one-factor), HSLS, SPGH
	(f) On-going-validation (monthly after 1 year) [Ch. 9]*	rating grade (individual for Brier & Tau)	Discrimination: AR, AUC, KS, KL, Kendall-tau, Brier

Note. * We have data up to April 2015.



The next step was to divide the data sample into several subsamples (see Table 4). Up to the implementation date (31. 12. 2013) we had three samples. The estimation sample covered the years 2011–2013 (these are the years of observing defaults, financial ratios were taken on the last date of the previous year). We formed samples by the following procedure:

- first we divided defaulted companies randomly into two samples (70 % to the estimation/in-sample, 30 % to the out-of-sample),

- then we added non-defaulters randomly to both samples in such a way that the default rate in the estimation sample was 20 %[98] and the default rate in the out-of-sample sample was the default rate from the original sample (actual default rate)[99],
- in addition we divided the in-sample into five folds, because we later performed the 5-fold cross-validation for optimal tuning of the RF and ANN algorithms.

After the model was implemented (PDs were assigned to non-defaulted corporates on 31. 12. 2013), we waited one year (this was our scenario). In practice there would usually be a modular system in place. In the meantime experts would classify companies not only by the results from the statistical model, but would also add their expert opinion, which would be based on the model's shortcomings and new information. After one year (on 31. 12. 2014) we performed a validation of the deviations of the model's probability of default from the observed default rates in the year 2014. If necessary, we would recalibrate the results or improve (or reestimate) the model.

9.2 Model (Ch. 3&4)

In this chapter we first present the main logistic model, which will be used in the next chapters. In addition we also present the results of alternative models (one can look at them as benchmark models for validation purposes).

Our main model is the logistic model developed on the experience gained in the article of Ortl and Gorenjec (2015). In the article we have shown the usefulness of HP data transformation and the benefits of choosing candidate variables for the final logistic regression with the random forest algorithm (variable importance feature).

After calibrating the RF model and obtaining the best RF VI variables (among initially 132 variables) we chose the best 30 variables in terms of RF VI (mean decrease in the Gini index) and decorrelated them. We used the Spearman pairwise correlation coefficient of 50 % and designed an algorithm for the elimination of correlated variables. We eliminated the variables that were correlated to the variables that have better ranking in RF VI, one by one.

Finally we were left with 9 variables to which we added 2 factor variables for size and legal status of the firm (we have tried combinations with other factor variables and some other financial ratios also, but performance was worse). The results showed that the coefficient for one ratio was negative wrongly (all coefficients after HP transformations should be positive), hence we eliminated this variable and ended up with 8+2 variables in the final model (model 1 in Table 5). The final variables were: X120, X113, X92, X109, X30, X96, X50, X8, and factors

[98] This was proposed for instance by Jovan (2007) and Kraus (2014). Also in Ortl and Gorenjec (2015) we obtained better results for a dataset created with such proportions, when compared to the proportions in the original dataset.
[99] Additionally it might be a good idea to use the data for one borrower only once (for a random year).

for size and legal status of the firm (F3, F4). Our variables covered more than five different credit risk groups and we were happy not to have more than 10 variables in the final model.

The final model resulted in an out-of sample AUC of 89,1 %[100], which we believe is a very good result for such a small dataset and is in line with the results of Ortl and Gorenjec (2015) or Jovan (2005) in Slovenia. Additionally, also the results of the out-of-time sample show high discriminatory power of the model with an AUC of 86,1 %, which allows us to conclude that the model is quite stable in time.

The results from the logistic regression were compared to the machine learning algorithms (RF, and ANN). First we calibrated these algorithms on the actual data (note that we performed separate calibrations for each dataset and for each data transformation, however all results are not reported here). The calibration was done by the following procedure:

- ANN: first we fixed the maximum number of iterations (500 mio), absolute tolerance error (0,0001) and decay parameter (0,1) to reasonable numbers and calibrated the optimal number of nodes in the hidden layer with the objective of maximum AUC in 5-fold validation. Then we fixed this value of nodes and started calibrating the decay parameter. After that we fixed these parameters and started calibrating the range of initial weights, which was set to 0,7 in the beginning. In the end we changed also the absolute tolerance error. After lots of testing we managed to tune the algorithm. For the ANN model we used an entropy cost function with one binary output unit.
- RF: similar to Kraus (2014) we first set the number of trees (T=1000) and the nodesize (nodesize=1) and chose the optimal number of features (m) that were randomly chosen in each tree. In the next step we calibrated the nodesize by fixing the number of trees and the number of features. In the last step we calibrated also the number of trees. We use the RF for classification problems.

The parameters were chosen in a 5-fold cross-validation procedure (folds were created on the estimation set) with the objective to get the maximum AUC (averaged across folds). The optimal parameters were then used to train a model on the estimation set.

For the ANN algorithm we do not have to calibrate the learning rate, as discussed in chapter 1.6. This is because we use a more sophisticated implementation of the ANN algorithm in the R package "nnet". This package does not use the standard gradient descent method, but uses the more sophisticated BFGS methods. It does also not rely on the standard backpropagation training algorithm, but uses some modifications. However, this implementation is restricted to the use of only one hidden layer. An alternative implementation that is based on the standard

[100] Note that we have also tried some manual adding or eliminating of variables but the model was not improved. Additionally, we tested a few other combinations of correlation threshold and number of RF VI variables to use as a starting point before running decorrelation.

backpropagation with gradient descent, as discussed in chapter 1.6, is the “neuralnet” package in R.

Note that for the ANN algorithm it is very important that the data is transformed into a more narrow interval, which makes training faster and reduces the chance of getting stuck in local optima. It is often preferred that the data is scaled to a [0,1] interval. Hence we applied the HP and MINMAX transformation (once for the raw data and once for the HP transformed data).

We tested the performance of the RF and ANN models also on raw winsorised data, but the results were worse (for ANN the performance dropped substantially). Best results were again obtained by the HP transformation.

In Table 5 we show the results for all developed models. Models 3–7 are directly comparable, as they were estimated on a set of 112 variables that were appropriate for use with any data transformation. All results (except for the last two columns were made out-of-sample). These models (3–7) have no added factor variables. The logistic regression model 3 was used in combination with the RF variable importance feature to rank the importance of the variables. Then we chose the best 30 variables and decorrelated them with a coefficient of correlation of 50 %. In the RF model the best results were obtained, when all 112 variables were put into the RF algorithm. In the ANN model it was beneficial to use 2/3 of RF VI variables that were also decorrelated (eliminated variables that were more than 75 % correlated with a better variable in terms of RF VI).

It turned out that logistic regression and RF models (models 3 and 4) perform equally well. The best model was the ANN model on HP transformed data, which resulted in an AUC of 89,2 % which is similar to model 1, which included more variables (also factor variables and was based on 132 variables) and thus not directly comparable to models 3–7. It was shown that the MINMAX transformation and the scaling of HP data performed worse (models 5 and 6) compared to the HP transformation (model 7).

We can surprisingly conclude that the ANN model performed best in our testing on a small dataset. Its disadvantage is the cumbersome tuning procedure and difficult interpretation. Also the RF algorithm performed very well and was very easy to calibrate and run. However, it was good to see that also the traditional logistic regression, which is very much used in practice, provides reasonable results.

It was comforting to see that similar model rankings were also suggested from other measures for the discriminatory power of the model (KS, Kendall, KL, Brier score). For model 1 we also tested the equation (41) and confirmed that $KS \geq D$ as $0{,}642 \geq 0{,}183$. This allowed us to conclude that significant differences exist between the score values of good and bad cases.

Table 5. Final models

Model	Algorithm	Data transf.	# of vars	AUC	AUC (95 % CI)	AR	KS statistic	Kendall tau	Brier score	KL measure	AUC in-sample	AUC out-of-time
1	Logistic regression*	**HP**	**8+2**	**89,1**	**(84,6 - 93,7)**	**78,3**	**64,2**	**24,1**	**5,17**	**1,62**	**90,9**	**86,1**
2		HP	8	87,4	(82,6 - 92,2)	74,8	61,3	23,1	6,11	1,39	89,0	80,1
3	Logistic regression	HP	7	87,4	(82,6 - 92,2)	74,8	65,6	23,1	6,29	1,50	88,5	80,4
4	Random forest (RF)	HP	112	87,3	(82,6 - 91,9)	74,5	62,6	23,0	6,11	1,38	100,0	83,6
5	Neural networks (ANN)	HP	29	89,2	(85,1 - 93,3)	78,5	65,2	24,0	6,12	1,57	91,8	81,6
6		HPminmax	29	88,0	(83,5 - 92,6)	76,1	63,3	23,5	6,26	1,47	90,7	88,0
7		minmax	28	84,8	(80,0 - 89,5)	69,5	57,8	21,4	7,92	1,06	95,5	80,0

Note. * Note that models 3–7 are directly comparable because they do not include factor variables and are based on the same number of starting variables (112). Models 1 & 2 are for instance based on 132 variables, the other models exclude variables as in Appendix G.

Comments to the transformations and variable choice procedure:

models 1&2: Models 1 & 2 are the only models based on all 132 variables. Out of them we chose 30 variables that had best performance in terms of the Gini decrease (RF VI). Then we removed the variables which were more than 50 % correlated (Spearman) with the variables that have higher ranking in the RF VI. Additionaly we removed one variable with a wrong (negative) parameter sign in the logistic regression. Note that with HP transformed variables (categorized) parameters for all variables should be positive. In model 2 we have 8 variables, in model 1 we add two additional factor variables for size and legal form of the firm.

model 3: Model 3 differs from models 1 & 2 in the starting variables. It is based on 112 variables because we excluded variables which we noted as appropriate for untransformed data as noted in Appendix G. We could use them with HP transformed data but we excluded them to achieve comparability with models 4–7 which also exclude these variables.

model 4: In the RF model we used all 112 variables (other are excluded as in model 3).

models 5&6&7: In models 5 & 6 & 7 we started with 112 variables (others were excluded as in model 3). Out of these variables we chose 2/3 of the best variables in terms of RF VI ranking (we were left with 75 variables in total). After decorrelation (here we excluded the variables which were more than 75 % correlated) we had 29 variables. Note that for model 6 we used double transformation. HP transformed data were further scaled to [0,1] interval with the MINMAX transformation. In model 7 we also used double transformation. Original data were first winsorised with 2 % caps and then the MINMAX transformation was applied on them.

Comments to the calibration of RF and ANN models.

RF: Note that we performed a RF calibration for each different dataset/transformation because RF VI variables were used also in other models. For instance, the calibration parameters for model 4 are: T=2000, m=8, nodesize=1. The objective in the calibration procedure was to get the maximum AUC within 5-fold validation.

ANN: Similar to the RF model, the objective in the calibration procedure for the ANN was the maximum AUC within 5-fold validation. The calibration parameters for model 5 were for instance: abstol = 0,0001, decay = 0,9, number of nodes = 10 and the initial weights were generated within the +/- 0,7 range.

As expected RF and ANN had the best results in-sample, because they had used more variables. However, it is important to note that the models differ in out-of-time testing, where the ANN model 6 performs best and both ANN models and RF outperform logistic regression. The drop in the out-of-time sample in some models is quite huge, although the AUCs are still quite good. We believe such behaviour is also due to some structural changes in Slovenia, for instance we

observe in the raw data a huge fluctuation of leverage ratios, because companies have been deleveraging in the last years.

This tells us that it is important to validate the model and monitor its performance frequently, because economy and firms change over time and the models need to be reestimated to capture these changes. However, out-of-time performance in the past is not as important as the performance in the future, because our model tries to forecast the future and not the past. This is subject of the next chapters.

9.3 Calibration, rating structure and implementation (Ch. 5&6&7)

Now that we had a working model (further on we will build only on the logistic regression model 1), we wanted to calibrate its predicted probabilities of default to the default rates that are expected to be realized in our corporate portfolio segment in the next year (2014). We followed the approach presented in chapter 5. We first calculated the average sample default rate resulting from logistic regression (using a sample that is representative of the non-defaulted portfolio) and the expected default rate for our bank portfolio.

We approximated the latter with the default rate in the year prior to the model implementation (we assume that the late history is our best guess for the future default rates). Note that more appropriate approaches would be: a) to include macroeconomic variables into the model[101] or b) to model the expected DR with a separate model and then scale our estimated PDs to these forecasted DRs[102].

We obtained calibrated/scaled PDs for our portfolio and now was the time to assign rating grades. Note that setting the rating structure is an important step, as bounds should in principle not change for several years and the economic cycle will be mirrored in the rating migration and not in the average rating grade PDs or DRs (in the PIT rating philosophy). This will be tested with calibration tests later in this chapter. If the tests show violations, recalibrations to higher PDs are needed.

First we needed to define the upper bounds to separate our 8 rating grades (A1, A2, B1, B2, B3, C1, C2, C3). To obtain an optimal rating structure, as defined in chapter 6, we set the bound optimally by maximizing the rating grade AUC. For forming an optimal rating structure we used the differential evolution optimisation algorithm. The main building blocks of the implementation were the definitions of the objective function and constraints and tuning of the DE parameters. The objective function we used is as follows:

$$f(b_i) = -AUC(b_i) + \alpha \cdot penalty \tag{65}$$

[101] These variables might provide better PIT PD estimations for the following year.
[102] There is a vast literature that describes satellite models for forecasting models (i.e. exogenous macroeconomic models or macroeconomic with endogenous economic variables – vector autoregression).

Because DE is the minimization algorithm, we added a minus sign in front of the AUC. To the objective function we added the penalty function to which we enforced the following restrictions to hold:

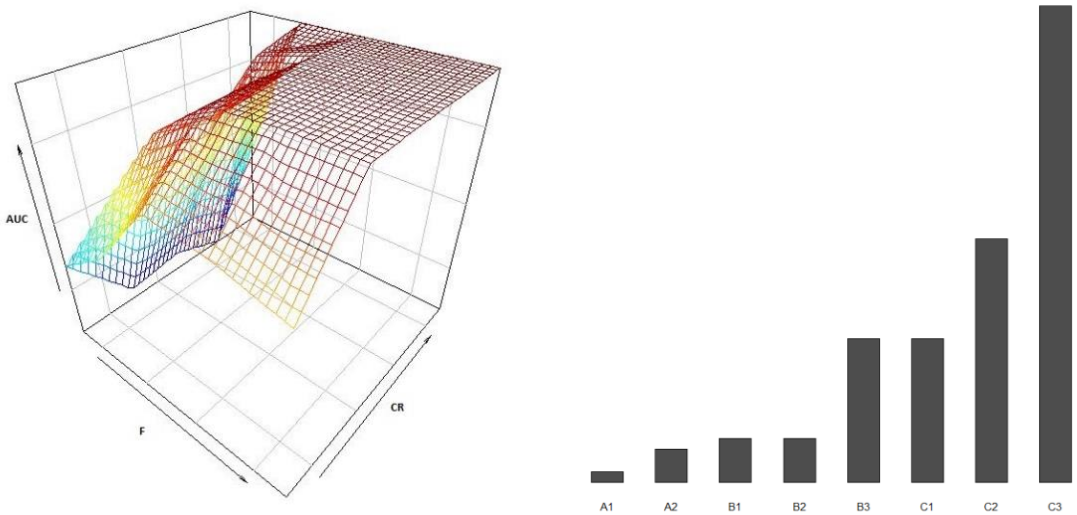
$b_i > b_{i+1}$ for all $i=1:6$ because the bounds need to be increasing,
$b_i \leq 1$ because the bound cannot be higher than 100% and,
$b_i \geq 0$ because our bound cannot be negative.

α is the penalty parameter (we set it to *10E6* for every restriction of constraint). The penalty function is proportional to the size of the violation of given constraint. b_i are the 7 resulting upper bounds which define 8 rating grades (the last grade is defined by the upper bound 100 %). $AUC(b_i)$ is our main objective function. We implemented the approach for calculating AUC based on Löffler and Posch (2007, pp. 214–222)[103]. For the DE algorithm we used the R implementation of Gilli, Grosse, & Schumann (2010).

However, note that the algorithm is prone to changes in tuning parameters. Hence we calibrated the parameters to obtain the optimal solution. Our problem turned out not to be very hard to solve. The only parameters with a huge effect on the results (AUCs) were the number of generations and the size of the population matrix. When we had set a big enough number of iterations and the size of the population to n_P=200, n_G=600, then the resulting AUCs did not change by more than 1 % when we changed other parameters.

As shown in Figure 16 (our calibration on our implementation portfolio with respect to changing combinations of the F and CR parameters was the range 0,1 to 0,9 for both parameters) the choice of F=0,9 and CR=0,9 provided stable results.

Figure 16. DE calibration and default rate for companies in 2013 by rating grade



[103] The input data for calculating an AUC of the rating structure are the PDs, to which the algorithm assigns rating grades in each iteration and evaluates the AUC for the rating structure.

We had also tested our algorithm with the code of Löffler and Posch (2007, pp. 218–222), which does the same thing as our DE algorithm, but uses Monte Carlo simulations. Their approach is very dependent on the choice of starting values and their randomisation. After a few hours of testing we couldn't find as good a solution as with the DE algorithm (our best try of running 10.000 simulations several times with different starting values still provided around 0,5 % lower AUC than the DE solution which needed no guessing of starting values).

Our non-defaulted borrowers on 31. 12. 2013 were then assigned to 8 rating grades for non-defaulters (all defaulted companies were given 100 % PD and combined to one additional rating class for defaulters). The AUC for this optimised rating structure was 86,4 %. Figure 16 shows how the companies that defaulted in the year 2013 would be classified with our rating structure, if we would have applied the model on 31. 12. 2012 (the y-axis is hidden, not to disclose any bank information).

Such a model would then be implemented into the bank's systems. However, one should think of some possible shortcomings of ratios used in the model. For instance one financial ratio includes cash. We know that having less cash can mean liquidity problems on one hand or good financial management (or parent support) on the other hand.

For this and for a variety of other reasons (see chapters 1.3 and 7) the bank should implement a system of overrides or some other way of including soft information and up-to-date data. The other possibility would be to eliminate such a variable.

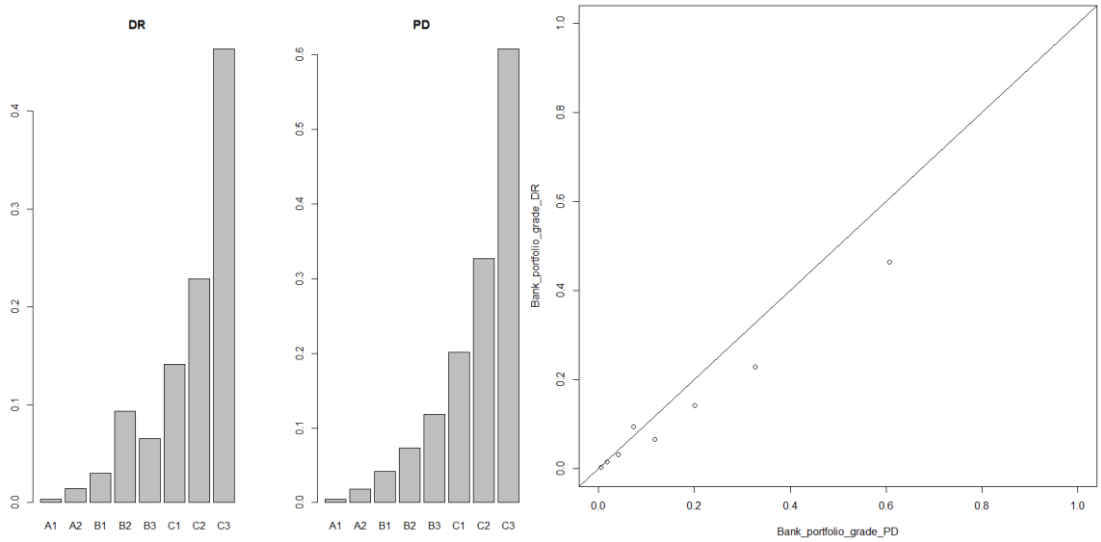
9.4 Quantitative validation (Ch. 8)

One year after the implementation we can test the discriminatory power of the model and whether the realized default rates per rating grades are in line with the PDs that were predicted on 31. 12. 2013 (here we assume no overrides have been made in the meantime). The new data needs to be transformed in the same way as the data that was used for estimating the model. Keep in mind that our validation procedure is more applicable to PIT rating systems, because TTC rating systems are based on much longer time horizons and long-term calibration.

In Figure 17 we show that the realized DRs per rating grade in the year 2014 are mostly in line with the predicted PD per rating grade on 31. 12. 2013 for the year 2014. The calibration plot on the right of Figure 17 shows that the predicted and realized default rates are close to the 45 % line, which means perfect calibration. It is also comforting to see that the realized default rates per rating grade are increasing for the worst rating grades (with the exception of the rating grade B3).

Also the tests for the discriminatory power showed good performance. As can be seen in Table 6 the AUC is 85,2 % and is stable in a one month moving-window monitoring up to April 2015.

Figure 17. Comparison of realized vs predicted default rates and calibration plot



Note that the DRs in Figure 17 are scaled by some multiplicative number with an objective not to disclose the real bank percentages.

Table 6. Validation of discriminatory power

Date of validation	AUC measure	AUC (95% CI)	AR measure	KS statistic	Kendall tau	Brier score	KL measure
31.12.2014	**85,2**	**(81,1 - 89,4)**	**70,5**	**56,4**	**25,0**	**5,59**	**1,08**
31.01.2015	85,3	(81,4 - 89,3)	70,7	56,7	25,2	5,64	1,10
28.02.2015	85,1	(81,0 - 89,1)	70,2	56,4	24,9	5,63	1,09
31.03.2015	86,1	(82,5 - 89,8)	72,3	57,1	25,2	5,51	1,14
30.04.2015	84,8	(80,6 - 89,0)	69,6	55,0	22,5	5,06	1,09

In Table 7 we show the calibration test results only for 31. 12. 2014 (the analysis could be easily extended to the monitoring samples). The null hypothesis that the actual defaults and the expected defaults are statistically equivalent is not rejected.

Table 7. Validation of calibration

Rating grade	Binomial (p-value in %)	Normal binomial (p-value in %)	One-factor (p-value in %)	HSLS (Chi-prob %)	SPGH (p-value in %)
A1	43,2	65,1	55,1		
A2	49,8	63,6	50,3		
B1	68,6	76,6	61,4		
B2	16,3	21,5	25,2		
B3	92,5	93,9	84,6		
C1	89,6	91,6	77,4		
C2	95,1	96,0	84,5		
C3	95,7	97,1	91,4		
ALL				91,06	94,6

In the figure above we were mostly worried about the rating grade B2, which showed the worst results in the binomial tests, but still was not rejected. That means we are safe to say that the predicted PDs are not underestimating the realized DRs with 95 % confidence. Note that we did not estimate the default correlations for the one-factor model, as our time series were not long enough. We assumed the asset correlation was 7 % as in the DB (2003). This means that our validation was positive and model reestimation or recalibration were not needed.

CONCLUSION

In the thesis we have discussed and implemented the PD rating system. We covered the whole model development cycle, from data preparation, modelling, calibration of PDs and model validation. In addition, we have compared the results of estimating PDs with different models. The results show that the ANN algorithm on HP transformed data outperformed our logistic regression model, which is quite surprising on a rather small dataset. We have shown that machine learning algorithms are not only useful as an alternative to estimating probabilities of default, but also as a complementary method at various stages of rating model development. We have successfully used random forests to select variables that were later input variables for logistic regression and also for the ANN algorithm. For defining the optimal rating structure we have applied the differential evolution algorithm.

Based on our hypothetical implementation of the main logistic regression model with a good out-of-sample and out-of-time performance, we have obtained good results also in the validation step one year after the implementation of the model. Our discriminatory and calibration tests show satisfactory stability and the calibration of the model. However, a possible extension of this thesis would be to implement a more sophisticated procedure for predicting future default rates by adding macroeconomic variables or by developing separate satellite models and using their results in our calibration procedure.

However, as noted in EBF (2012, pp. 16, 40) there are different approaches from different central banks when applying these methodologies. We think that our methodology and comments (with referencing alternative approaches) provides a nice introduction to modelling PDs for different use (i.e. for impairments under the IFRS 9 approach). For the IRB approach one should do some more research on how to validate models based on the TTC rating philosophy and how to calculate TTC PDs from PIT PDs. We believe this thesis can provide a nice introduction to different issues related to the construction of an operational rating system.

Throughout the thesis we reported different identified regulatory shortcomings and a too high degree of flexibility of the IRB approach, which may lead to risks regarding the robustness of models (mostly noted by EBA, 2013c). To regain public trust in IRB models, the models need to be comparable. Recently EBA (2015) set out an outline for further improvements which will affect primarily the harmonisation of default definitions, harmonised formula for the computation of default rates (and multiple defaults), PD model calibration (calibration approach, definition on economic cycle, how to cope with the absence in the available stressed period time

series), low default modelling approaches (and calibration) and allowance for the partial use of a standardised approach (also important for low default portfolios).

All these topics are planned to be addressed separately by the end of the year 2017 in different guidelines or technical standards. The timeline can be seen in EBA (2015, p. 15). This is very important, because these topics contribute most to the differences in capital requirements calculations, as stated by EBA (2013c, pp. 16–20). Changes in all these topics will lead to a significant operational burden, both for institutions and for supervisors.

However, almost at each step in the flowchart there are several possible approaches (i.e. we could set the optimal rating structure by maximizing AUC with DE or alternatively by using cluster analysis). We have decided for approaches that are commonly used in banks, commonly cited or we thought they would provide good results. However, each approach can be tested by applying alternatives (i.e. compare different transformations on variables, different models for estimation, different approaches to define rating structure or variable selection etc.).

REFERENCE LIST

1. Adams, T. (2009). *Demystifying the pro-cyclicality of Basel II.* Retrieved January 22, 2016, from https://www.nedbank.co.za/content/dam/nedbank/site-assets/AboutUs/Information%20Hub/Corporate%20Presentations/2009/Deutsche_Securities_Credit_Teach-in_-_Demystifying_the_Procyclicality_of_Basel_II_-_May_2009.pdf
2. Aguais, C., Forest, L, Wong, E., & Ledezma, D. (2004). Point-in-Time versus Through-the-Cycle Ratings. *Barclays Capital.* Retrieved January 20, 2016, from https://www.z-riskengine.com/media/1029/point-in-time-versus-through-the-cycle-ratings.pdf
3. Allison, P. (2013). *Why I don't trust the Hosmer-Lemeshow test for logistic regression.* Retrieved January 3, 2016, from http://statisticalhorizons.com/hosmer-lemeshow
4. Altman, E., & Sabato, G. (2007). Modeling Credit Risk for SMEs: Evidence from the US Market. *Abacus, A Journal of Accounting, Finance and Business Studies, 43*(3), 332–357.
5. Anderson, R. (2007). *The credit scoring toolkit.* Oxford: Oxford University Press.
6. Angelini, E., Tollo, G., & Roli, A. (2008). A neural network approach for credit risk evaluation. *Elsevier. The quarterly Review of Economics and Finance, 48,* 733–755.
7. Atiya, A. (2001). Bankruptcy Prediction for Credit Risk Using Neural Networks: A Survey and New Results. *IEE Transactions on Neural Networks, 12*(4), 929–935.
8. Balthazar, L. (2006). *From Basel 1 to Basel 3.* New York: Palgrave Macmillan.
9. Bank for International Settlements – BIS. (2001). *The New Basel Capital Accord, Consultative Document, The Internal ratings-Based Approach.* Basel: Bank for International Settlements, Basel Committee on Banking Supervision.
10. Bank for International Settlements – BIS. (2005). *Studies on the Validation of Internal Rating Systems* (Working Paper No. 14). Basel: Bank for International Settlements, Basel Committee on Banking Supervision.
11. Bank for International Settlements – BIS. (2006). *Stability of a »through-the-cycle« rating system during a financial crisis.* Basel: Bank for International Settlements, Basel Committee on Banking Supervision.
12. Bank of England – BoE. (2013). *Credit risk: internal ratings based approaches* (Consultation paper CP4/13). London: Bank of England, Prudential Regulation Authority.
13. Bartlett, J. (2014). *The Hosmer-Lemeshow goodness of fit test for logistic regression.* Retrieved January 5, 2016, from http://thestatsgeek.com/2014/02/16/the-hosmer-lemeshow-goodness-of-fit-test-for-logistic-regression/
14. Begin, B., & Thomas, S. (2012). Basel II retail modelling approaches PD Models. *Actuaries Institute.* Retrieved January 22, 2016, from http://actuaries.asn.au/Library/Events/Insights/2012/Presentation-BaselIIRetailModellingApproaches.pdf
15. Brezigar-Masten, A., & Masten, I. (2012). CART-based selection of bankruptcy predictors for the logit model. *Expert Systems with Applications, 39*(11), 10153–10159.
16. Christodoulakis, G., & Satchell, S. (2008). *The analytics of risk model validation.* Oxford: Elsevier Finance.
17. Cipollini, F., Dainelli, F., & Giunta, F. (2009*). Determinants of SME Creditworthiness after Basel II: Evidence from Italy.* Retrieved December 15, 2015, from

http://www.gcbe.us/9th_GCBE/data/Fabrizio%20Cipollini,%20Francesco%20Dainelli,%20Francesco%20Giunta.doc
18. Clavero-Rasero, B. (2006). *Statistical aspects of setting up a credit rating system* (Doctoral dissertation). Kaiserslautern: Universität Kaiserslautern.
19. Deutsche Bundesbank – DB. (2003). *Approaches to the validation of internal rating systems.* Frankfurt: Deutsche Bundesbank.
20. Engelmann, B., & Rauhmeier, R. (2011). *The Basel II Risk Parameters* (2nd ed.). Heidelberg: Springer.
21. European Banking Authority – EBA. (2013a). *Interim results update of the EBA review of the consistency of risk-weighted assets, Low default portfolio analysis.* London: European Banking Authority.
22. European Banking Authority – EBA. (2013b). *Third interim report on the consistency of risk-weighted assets, SME and residential mortgages.* London: European Banking Authority.
23. European Banking Authority – EBA. (2013c). *Summary report on the comparability and pro-cyclicality of capital requirements under the Internal Ratings Based Approach in accordance with Article 502 of the Capital Requirements Regulation*. London: European Banking Authority.
24. European Banking Authority – EBA. (2015). *Future of the IRB Approach* (Discussion Paper 2015/01). London: European Banking Authority.
25. European Banking Federation – EBF. (2012). *Study on Internal Rating Based (IRB) models in Europe. Residential mortgages.* Brussels: European Banking Federation.
26. European Commission – EC. (2013). *Regulation (EU) No 575/2013 of the European parliament and of the council on prudential requirements for credit institutions and investment firms and amending Regulation (EU)*. Retrieved April 2, 2016, from http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:321:0006:0342:EN:PDF
27. Falavigna, G. (2006). Models for Default Risk Analysis Focus on Artificial Neural Networks, Model Comparisons, Hybrid Frameworks. *WP Ceris-CNR, 10*. Moncalieri (TO): National Research Council of Italy, Ceris-Cnr.
28. Falkenstein, E., Boral, A., & Carty, L. (2000). *RiskCalc for Private Companies: Moody's default model: Rating Methodology*. New York: Moody's Investors Service.
29. Gilli, M., Grosse, S., & Schumann E. (2010). Calibrating the Nelson-Siegel-Svensson model. *Comisef Working Papers Series, WPS-031 30/03/2010.* Retrieved April 5, 2016, from http://comisef.eu/?q=Nelson%E2%80%93Siegel%E2%80%93Svensson
30. Gilli, M., Maringer, D., & Schumann E. (2011). *Numerical Methods and Optimization in Finance.* Oxford: Elsevier.
31. GraphPad (n.d.). *Interpreting results: Kolmogorov-Smirnov test.* Retrieved January 5, 2016, from http://www.graphpad.com/guides/prism/6/statistics/index.htm?interpreting_results_kolmogorov-smirnov_test.htm
32. Hamilton, D., Sun, Z., & Ding, M. (2011). Through-the-Cycle EDF Credit Measures. *Moody's Analytics.* Retrieved January 20, 2016, from http://www.moodysanalytics.com/~/media/Microsites/ERS/2011/through-cycle-

EDF/MoodysAnalytics_Through-the-Cycle%20EDF%20Measure%20Methodology%20Overview.pdf
33. Hastie, T, Tibshirani, R., & Friedman, J. (2008). *The elements of statistical learning. Data mining, inference, and prediction* (2nd ed.). New York: Springer.
34. Hastie, T, Tibshirani, R., James, G., & Witten, D. (2013). *An introduction to statistical learning with applications in R.* New York: Springer.
35. Hayden, E. (2002). *Modeling an Accounting-Based Rating System for Austrian Firms* (Doctoral dissertation). Vienna: University of Vienna, Fakultät für Wirtschaftswissenscheften und Informatik.
36. Hinton, G. (2012). *Coursera course: Neural Networks for Machine Learning.* [Video File]. University of Toronto. Retrieved March 20, 2016, from https://www.coursera.org/course/neuralnets
37. Ingolfsson, S., & Elvarsson, B. (2007). *Cyclical adjustment of Point-in-Time (PiT) PD.* Retrieved January 25, 2016, from http://www.business-school.ed.ac.uk/waf/crc_archive/2007/papers/ingolfsson-elvarsson-cyclical.pdf
38. Jagrič, V., Kračun, D., & Jagrič, T. (2011). Does Non-linearity Matter in Retail Credit Risk Modeling? *Czech Journal of Economics and Finance, 61*(4), 384–402.
39. Joseph, M. (2005). *A PD Validation Framework for Basel II Internal Ratings-Based Systems.* Retrieved February 15, from http://crededata.com/crede/publish/joseph-maurice%20(Credit%20Scoring%20and%20Credit%20Control%20IX).pdf
40. Jovan, M. (2005). *Od česa je odvisno razločevanje statističnih modelov?* Ljubljana: Bank of Slovenia.
41. Kastrin, A. (2015). *Uvrščanje in diskretizacija mnogorazsežnih mikromrežnih DNA-podatkovij* (Doctoral dissertation). Novo Mesto: Faculty of information studies.
42. Keenan, S., & Sobehart, J. (1999). Performance Measures for Credit Risk Models. *Moody's Risk Management Services.* Retrieved January 10, 2016, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.456.9608&rep=rep1&type=pdf
43. Kosič, U. (2012). *Mera podobnosti na podlagi naključnih gozdov* (Bachelor thesis). Ljubljana: University of Ljubljana, Faculty of Computer and Information Science.
44. Kraus, A. (2014). *Recent methods from statistics and machine learning for credit scoring* (Doctoral dissertation). München : Ludwig-Maximilians-Universität München, Faculty of mathematics, informatics and statistics.
45. Lantz, B. (2013). *Machine Learning with R.* Birmingham: PACKT Publishing.
46. Lessmann, S., Seow, H., Baesens, B., & Thomas, L. (2015). Benchmarking state of the art classification algorithms for credit scoring: A ten-year update. *European Journal of Operational Research, 247*(1), 124–136. Retrieved 14 April, 2016, from http://www.business-school.ed.ac.uk/waf/crc_archive/2013/42.pdf
47. Löffler, G., & Posch, P. (2007). *Credit Risk Modeling Using Excel and VBA.* West Sussex: John Wiley & Sons.
48. Maarse, B. (2012). *Backtesting framework for PD, LGD and EAD* (Master's thesis). Twente: University of Twente.
49. Mitchell, T. (1997). *Machine learning.* Columbus Ohio: McGraw-Hill Science.

50. Ng, A. (2015). *Coursera course: Machine Learning.* [Video File] Stanford University. Retrieved March 12, 2016, from https://www.coursera.org/learn/machine-learning
51. Novotny-Farkas, Z. (2015). The significance of IFRS 9 for financial stability and supervisory rules. *Directorate-general for internal policies. Policy department: Economic and scientific policy. European Parliament.* Retrieved April 3, 2016, from http://www.europarl.europa.eu/RegData/etudes/STUD/2015/563461/IPOL_STU(2015)563461_EN.pdf
52. Oesterreichische Nationalbank – ONB. (2004). *Rating models and validation. Guidelines on credit risk management.* Vienna: Oesterreichische Nationalbank.
53. Ortl, A. (2014). *Mathematical Portfolio Optimizaction with Mean-Variance, Mean-CVaR and Max-Omega Approaches with Cardinality Constraint* (Master's thesis). Maribor: University of Maribor, Faculty of economics and business.
54. Ortl, A., & Gorenjec, R. (2015). Corporate default prediction with the Random Forest algorithm. *The Journal of Money and Banking, 2015*(12), 22–28.
55. PennState, Eberly College of Science (n.d.). *Kendall Tau-b Correlation Coefficient.* Retrieved January 10, 2016, from https://onlinecourses.science.psu.edu/stat509/node/158
56. Petrov, A. (2015). *Rating models development in modern world.* Retrieved January 20, 2016, from https://www.sas.com/content/dam/SAS/ru_ru/doc/other2/Petrov_Alexander_A2015.pdf
57. Press, W., Teukolsky, S., Vetterling, W., & Flannery, B. (1992). *Numerical Recipes in C*. Cambridge: Cambridge University Press.
58. PWC (2014). IFRS 9 – Expected credit losses. *In depth: A look at current financial reporting issues.* Retrieved March 30, 2016, from https://www.pwc.com/us/en/cfodirect/assets/pdf/in-depth/us2014-06-ifrs-9-expected-credit-losses.pdf
59. Quantiki (2015). *Classical relative entropy.* Retrieved January 17, 2016, from https://quantiki.org/wiki/classical-relative-entropy
60. Rauhmeier, R., & Scheule, H. (2005). Rating Properties and their Implication on Basel II-Capital. *Risk* (18), 78–81. Retrieved January 8, 2016, from https://editorialexpress.com/cgi-bin/conference/download.cgi?db_name=QMF2005&paper_id=38
61. Refaeilzadeh, P., Tang, L., & Liu, H. (2008). *Cross-validation.* Arizona: Arizona State University.
62. R-tutor (n.d.). *Significance Test for Kendall's Tau-b.* Retrieved January 10, 2016, from http://www.r-tutor.com/gpu-computing/correlation/kendall-tau-b
63. Samuels, D. (2012). *From Second Wave Basel II to Basel III.* Retrieved December 1, 2015, from https://www.capitaliq.com/
64. SAS (n.d.). *Kendall's Tau-b Correlation Coefficient.* Retrieved January 10, 2016, from http://support.sas.com/documentation/cdl/en/procstat/63104/HTML/default/viewer.htm#procstat_corr_sect015.htm
65. SAS (2015). *Statistical Measures Used in Basel II Reports.* Retrieved January 13, 2016, from http://support.sas.com/documentation/cdl/en/mdsug/65072/HTML/default/viewer.htm#n194xndt3b3y1pn1ufc0mqbsmht4.htm

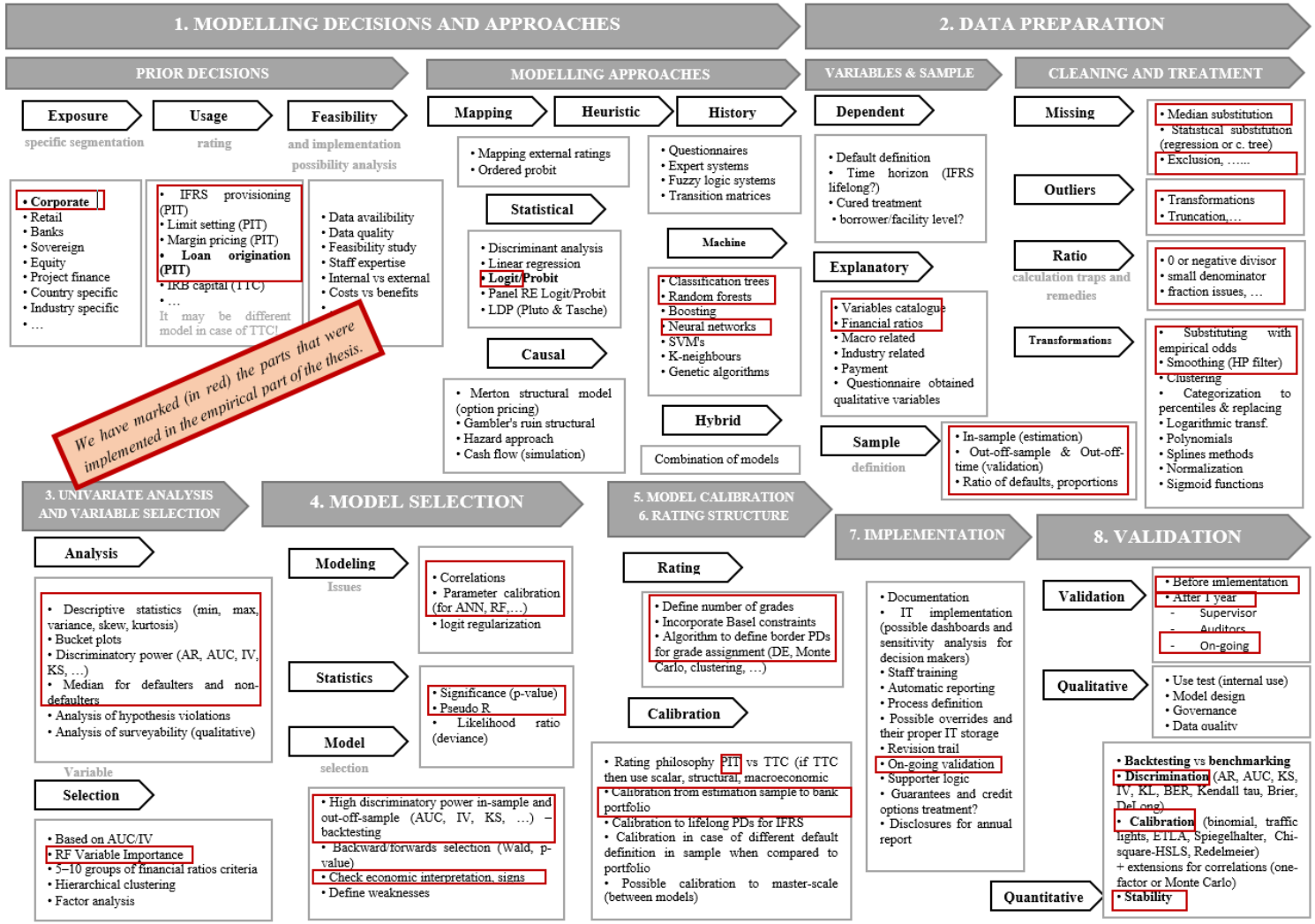
66. Saurina, J., & Trucharte, C. (2007). *An assessment of Basel II procyclicality in mortgage portfolios.* Retrieved January 25, 2016, from http://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosTrabajo/07/Fic/dt0712e.pdf
67. Tata (2015). *IFRS 9 Expected Loss Impairment Accounting Model versus Basel Framework.* Retrieved February 27, 2016 from http://www.tcs.com/SiteCollectionDocuments/White-Papers/Expected-Loss-0515-1.pdf
68. Tucker, J. (1996). *Neural networks versus logistic regression in financial modelling: A methodological comparison.* Retrieved March 16, 2016, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.54.1759&rep=rep1&type=pdf
69. Wu, X. (2008). *Credit Scoring Model Validation* (Master's thesis). Amsterdam: University of Amsterdam, Faculty of science, Institute for mathematics.
70. Zhang, G., Hu, M., Patuwo, B., & Indro, D. (1999). Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis. *European Journal of Operational Research, 1999*(116), *16–32.*

APPENDIXES

TABLE OF APPENDIXES

Appendix A: Flowchart

Figure 1. Flowchart



Appendix B: Derivations for ANN

a) derivations for the output layer

We need to prove the equation from the text (in the derivation we use concepts from Mitchell (1997, pp. 92–93, 97–103 and pp. 170–171)):

$$\frac{\partial E_1}{\partial w_{1h}^{(2)}} = \frac{\partial E_1}{\partial z_1^{(3)}} \frac{\partial z_1^{(3)}}{\partial w_{1h}^{(2)}} = \ldots = a_h^{(2)} \delta_1^{(3)} = a_h^{(2)} (y_1 - h_1) \quad (1)$$

Notice that the weight $w_{1h}^{(2)}$ influences the output through $z_1^{(3)}$. We use the cross-entropy error function from the text (equation 16), whereas the squared error derivations will be described in the footnotes[104] (note that the minus sign in the text is due to the minimizing nature of an optimisation algorithm):

$$E_1(\vec{w}) = y_1 log(h_1) + (1-(y_1)) log(1-h_1) \quad (2)$$

Then we compute the second part (and use the fact that $z_1^{(3)} = \sum_{h=0}^{H} a_h^{(2)} w_{1h}^{(2)}$):

$$\frac{\partial z_1^{(3)}}{\partial w_{1h}^{(2)}} = a_h^{(2)} \quad (3)$$

The first part is a bit more complicated[105] because we need to use chain rule again as we need to account for the sigmoid function, which is applied over:

$$\frac{\partial E_1}{\partial z_1^{(3)}} = \frac{\partial E_1}{\partial h_1} \frac{\partial h_1}{\partial z_1^{(3)}} \quad (4)$$

$$\frac{\partial E_1}{\partial h_1} = \frac{\partial}{\partial h_1} \left(y_1 log(h_1) + (1-(y_1)) log(1-h_1) \right) = \left(\frac{y_1}{h_1} - \frac{(1-y_1)}{1-h_1} \right) = \frac{y_1(1-h_1) - h_1(1-y_1)}{h_1(1-h_1)} = \frac{y_1 - h_1}{h_1(1-h_1)} \quad (5)$$

Here we have used the rule for derivative of a logarithm $log(h_1) = \frac{1}{h_1}$ (for log being the natural logarithm). We continue with the other part (derivative of the sigmoid function):

$$\frac{\partial h_1}{\partial z_1^{(3)}} = \frac{\partial \sigma(z_1^{(3)})}{\partial z_1^{(3)}} = \ldots = h_1(1-h_1) \quad (6)$$

[104] The squared error cost function would be $E_1(\vec{w}) = \frac{1}{2}(y_1 - h_1)^2$.

[105] For the squared error cost function we would obtain: $\frac{\partial E_1}{\partial h_1} = \frac{\partial}{\partial h_1} \frac{1}{2} (y_1 - h_1)^2 = \frac{1}{2} 2 (y_1 - h_1)(-1) = -(y_1 - h_1)$.

Because:

We follow Hinton (2012) when doing the derivative of the sigmoid function:

$$h_1=\sigma\left(z_1^{(3)}\right)=\frac{1}{1+e^{-z_1^{(3)}}}=(1+e^{-z_1^{(3)}})^{-1} \tag{7}$$

$$\frac{\partial h_1}{\partial z_1^{(3)}}=\frac{-1\left(-e^{-z_1^{(3)}}\right)}{\left(1+e^{-z_1^{(3)}}\right)^2}=\left(\frac{1}{1+e^{-z_1^{(3)}}}\right)\left(\frac{e^{-z_1^{(3)}}}{1+e^{-z_1^{(3)}}}\right)=h_1(1-h_1) \tag{8}$$

Because:

$$\frac{e^{-z_1^{(3)}}}{1+e^{-z_1^{(3)}}}=\frac{\left(1+e^{-z_1^{(3)}}\right)-1}{1+e^{-z_1^{(3)}}}=\frac{1+e^{-z_1^{(3)}}}{1+e^{-z_1^{(3)}}}+\frac{-1}{1+e^{-z_1^{(3)}}}=1-h_1 \tag{9}$$

It follows that[106]:

$$\frac{\partial E_1}{\partial z_1^{(3)}}=\frac{\partial E_1}{\partial h_1}\frac{\partial h_1}{\partial z_1^{(3)}}=\frac{y_1-h_1}{h_1(1-h_1)}h_1(1-h_1)=y_1-h_1 \tag{10}$$

If we combine all terms we get the final result:[107]

$$\frac{\partial E_1}{\partial w_{1h}^{(2)}}=\frac{\partial E_1}{\partial z_1^{(3)}}\frac{\partial z_1^{(3)}}{\partial w_{1h}^{(2)}}=a_h^{(2)}(y-h_1) \tag{11}$$

b) derivations for the hidden layer

We need to prove the equation from the text (in the derivation we use concepts from Mitchell (1997, p. 103) and derive with respect to the weights in the hidden layer):

$$\frac{\partial E_1}{\partial w_{hk}^{(1)}}=\frac{\partial E_1}{\partial z_h^{(2)}}\frac{\partial z_h^{(2)}}{\partial w_{hk}^{(1)}}=\ldots=x_k^{(1)}\delta_h^{(2)}=x_k^{(1)}w_{hk}^{(2)}\left(y_1-h_1\right)\left(a_h^{(2)}\left(1-a_h^{(2)}\right)\right) \tag{12}$$

This is again more complicated because to get the first part we need to use chain rule again:

[106] For squared error we would obtain: $\frac{\partial E_1}{\partial z_1^{(3)}}=\frac{\partial E_1}{\partial h_1}\frac{\partial h_1}{\partial z_1^{(3)}}=-(y_1-h_1)h_1(1-h_1)$.

[107] For squared error we would obtain: $\frac{\partial E_1}{\partial w_{1h}^{(2)}}=\frac{\partial E_1}{\partial z_1^{(3)}}\frac{\partial z_1^{(3)}}{\partial w_{1h}^{(2)}}=-(y_1-h_1)h_1(1-h_1)a_h^{(2)}$.

$$\frac{\partial E_1}{\partial z_h^{(2)}} = \frac{\partial E_1}{\partial z_1^{(3)}} \frac{\partial z_1^{(3)}}{\partial z_h^{(2)}} \tag{13}$$

From above we propagate the error backwards:

$$\frac{\partial E_1}{\partial z_1^{(3)}} = (y_1 - h_1) \tag{14}$$

Then we need to use another chain rule:

$$\frac{\partial z_1^{(3)}}{\partial z_h^{(2)}} = \frac{\partial z_1^{(3)}}{\partial a_h^{(2)}} \frac{\partial a_h^{(2)}}{\partial z_h^{(2)}} \tag{15}$$

And because $z_1^{(3)} = \sum_{h=0}^{H} a_h^{(2)} w_{1h}^{(2)}$ we calculate:

$$\frac{\partial z_1^{(3)}}{\partial a_h^{(2)}} = w_{1h}^{(2)} \tag{16}$$

Similar as above we have the derivative of a sigmoid function:

$$\frac{\partial a_h^{(2)}}{\partial z_h^{(2)}} = \ldots = a_h^{(2)}(1 - a_h^{(2)}) \tag{17}$$

The term $\frac{\partial z_h^{(2)}}{\partial w_{hk}^{(1)}}$ is actually $x_k^{(1)}$ because we use the equation $z_h^{(2)} = \sum_{k=0}^{K} x_k^{(1)} w_{hk}^{(1)}$. Now we combine all the terms and the final result is:

$$\frac{\partial E_1}{\partial w_{hk}^{(1)}} = \frac{\partial E_1}{\partial z_h^{(2)}} \frac{\partial z_h^{(2)}}{\partial w_{hk}^{(1)}} = \ldots = x_k^{(1)} \delta^{(2)} = x_k^{(1)} w_{1h}^{(2)} (y_1 - h_1) a_h^{(2)} (1 - a_h^{(2)}) \tag{18}$$

Appendix C: Examples of discriminatory power validation measures (bucket level)

First we need to prepare the hypothetical data that will be used in the tests below (the whole appendix refers to this data). The formulas are the same as described in the main text.

Figure 2. Data preparation for discriminatory power validation (bucket level)

	Individual borrower i	Rating grade r	Prob. of default PD	Default status D
5	1	C	8,0%	1
6	2	C	8,0%	1
7	3	C	8,0%	1
8	4	C	8,0%	0
9	5	B	2,0%	1
10	6	B	2,0%	0
11	7	B	2,0%	0
12	8	A	0,1%	0
13	9	A	0,1%	0
14	10	A	0,1%	0

	Cut-off value C	HC $H(C)$	FC $F(C)$	HR $HR(C)$	FAR $FAR(C)$
20	**8,0%**	0	0	0%	0%
21	**2,0%**	3	1	75%	17%
22	**0,1%**	4	3	100%	50%
23	**0,0%**	4	6	100%	100%

25	**# of defaults:**	4	N_D
26	**# of non-defaults:**	6	N_{ND}

B20: =COUNTIFS($C$5:$C$14;">8%";$D$5:$D$14;"1")
C20: =COUNTIFS($C$5:$C$14;">8%";$D$5:$D$14;"0")
D20: =B20/$C$25
E20: =C20/$C$26

	Rating / score $(r = i)$ S_i	Counting (goods) ND_i	Counting (bads) D_i	Probability (share) (goods) P_{ND}	Probability (share) (bads) P_D	Cumulative (goods)	Cumulative (bads)	Cumulative prob. (goods) F_{ND}	Cumulative prob. (bads) F_D
33	**C**	1	3	17%	75%	1	3	17%	75%
34	**B**	2	1	33%	25%	3	4	50%	100%
35	**A**	3	0	50%	0%	6	4	100%	100%
36	**Totals**	**6**	**4**	**100%**	**100%**				

Fraction of non-defaults

Fraction of defaults

	Rating	Proportion per rating class	Fraction of borrowers	Default rate per rating
52	**C**	40%	0%	75%
53	**B**	30%	40%	33%
54	**A**	30%	70%	0%
55	**Totals**	**100%**	100%	**40%**

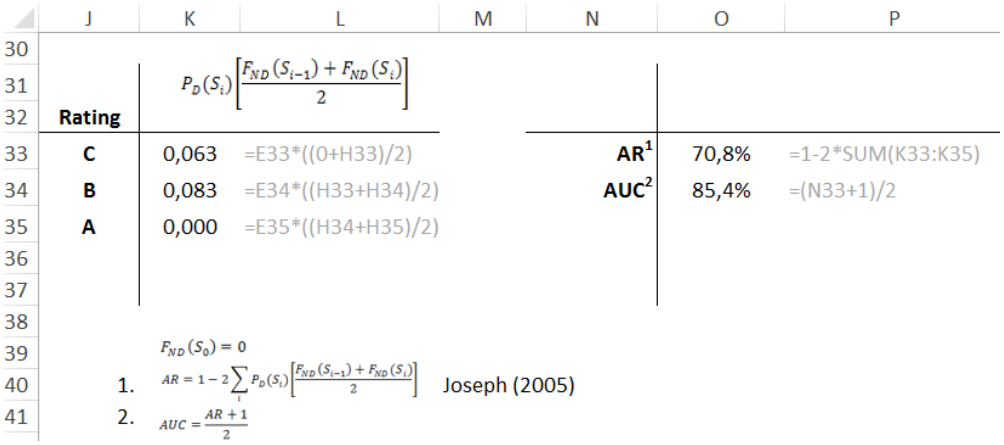
B52: =COUNTIF($B$5:$B$14;A52)/10
C53: =C52+B52
D52: =SUMIF($B$5:$B$14;A52;$D$5:$D$14)/COUNTIF($B$5:$B$14;A52)

B33: =COUNTIF($B$5:$B$14;A33)-C33
C33: =SUMIF($B$5:$B$14;A33;$D$5:$D$14)
D33: =B33/$B$36
E33: =C33/$C$36
F34: =F33+B34
H34: =H33+D34

Accuracy ratio (AR) and Area under ROC curve (AUC)

Calculation of the AR is based on Joseph (2005) and shown in the figure below:

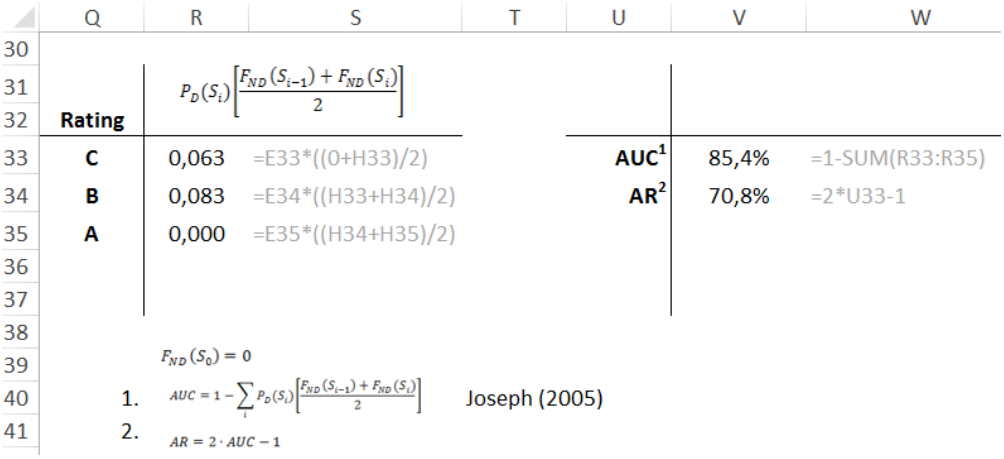
Figure 3. Calculation of AR



Source: modified from M. Joseph, *A PD Validation Framework for Basel II Internal Ratings-Based Systems*, 2005, p. 10.

Calculation of the AUC is based on Joseph (2005) and shown in the figure below:

Figure 4. Calculation of AUC

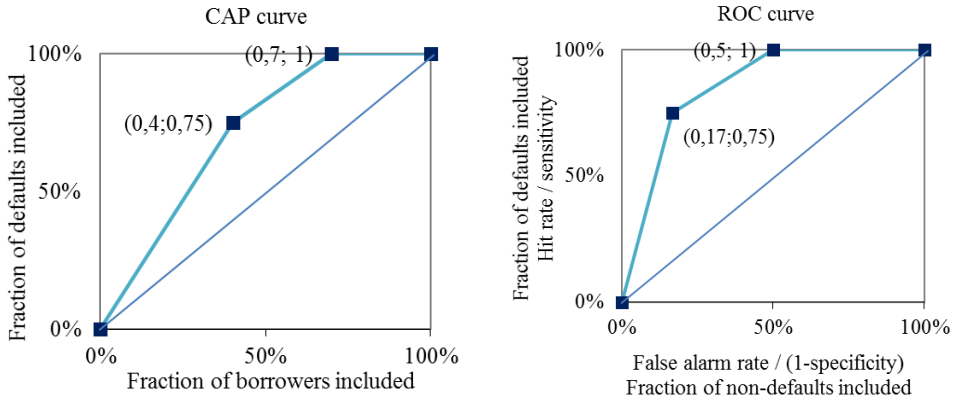


Source: modified from M. Joseph, *A PD Validation Framework for Basel II Internal Ratings-Based Systems*, 2005, p. 10.

Based on the calculated data one can construct the CAP and ROC curves. In case of CAP curve the point [0,4 ; 0,75] means there are 75 % of all defaults classified into the first rating class (C) (see cells C53:C55 and I33:I36). From all borrowers, 40 % are given class C (cell C53). In classes B and C there are 70 % of all borrowers and 100 % of all defaulters (see cells C53 and I34). In the class A there are the remaining 30 % of borrowers.

In the figures below there are the CAP (left) and the ROC (right) curve. For construction of the CAP curve we use cells C52:C54 for the x-axis and I33-I36 for the y-axis. Similar for the construction of the ROC curve we use cells E20:E23 (or H33:H35) for the x-axis and D20:D23 (or I33:I36) for the y-axis.

Figure 5. CAP and ROC curves



Source: modified from G. Löffler & P. Posch, *Credit risk modelling using Excel and VBA*, 2007, p. 148.

Kolmogorov-Smirnov test (KS)

We compare two methods for calculating the KS (see references in the figure below).

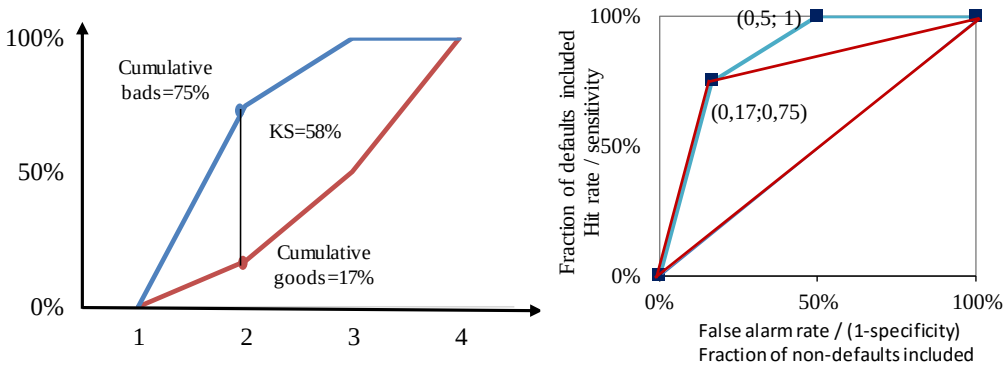
Figure 6. Calculation of KS test

	X	Y	Z	AA	AB	AC
29						
30			**KS[1]**	**KS[2]**		
31			$F_D - F_{ND}$	$HR - FAR$		
32		**Rating**				
33			0,00%	0,00%		
34		**C**	**58,33%**	**58,33%**	**<-- Max**	
35		**B**	50,00%	50,00%		
36		**A**	0,00%	0,00%		
37						
38						
39		1. $KS = \max_c(\bar{F}_D(c) - \bar{F}_{ND}(c))$		Joseph (2005)		Z34: =I33-H33
40		2. $KS = \max_c \lvert HR(c) - FAR(c) \rvert$		BIS (2005)		AA34: =D21-E21

Source: modified from M. Joseph, *A PD Validation Framework for Basel II Internal Ratings-Based Systems*, 2005, p. 28.

The idea of KS test is presented also on the two graphs that are related to the actual calculations. On the left figure KS is the maximum difference between the cumulative frequency distributions of good and bad cases. On the right figure the KS is twice the area of the largest triangle that can be drawn between the diagonal and the ROC curve (the ROC curve is constructed as above).

Figure 7. Interpretation of KS test



Kullback-Leibler statistic (KL)

We compare two methods of calculating the KL (see references in the figure below).

Figure 8. Calculation of KL statistic

	AT	AU	AV	AW	AX	AY	AZ	BA
31	**Rating**	**KL[1]**	**Cross Entropy**	**Entropy**				
32						Cross entropy	162%	=SUM(AV33:AV35)
33	**C**	113%	134%	-22%		Entropy	-56%	=SUM(AW33:AW35)
34	**B**	-7%	27%	-35%		**KL[1]**	**106%**	=SUM(AU33:AU35)
35	**A**	0%	0%	0%		**KL[2]**	**106%**	=AZ32+AZ33
36								
37								
38	1.	$KL = \sum_i P_D(S_i) \cdot ln\left[\frac{P_D(S_i)}{P_{ND}(S_i)}\right]$				Joseph (2004)		
39								
40	2.	$KL(p,q) = \underbrace{-\sum_x p(x) \ln q(x)}_{Cross\ entropy} + \underbrace{\sum_x p(x) \ln p(x)}_{Entropy}$				Quantiki (2015)		
41								
42								
43		AU33: =E33*LN(E33/D33)						
44		AV33: =-E33*LN(D33)						
45		AW33: =E33*LN(E33)						

Source: modified from M. Joseph, *A PD Validation Framework for Basel II Internal Ratings-Based Systems*, 2005, p. 32.

Appendix D: Examples of discriminatory power validation measures (individual level)

First we need to prepare the hypothetical data that will be used in the tests below (the whole appendix refers to this data). The formulas are the same as described in the main text.

Figure 9. Data preparation for discriminatory power validation (individual level)

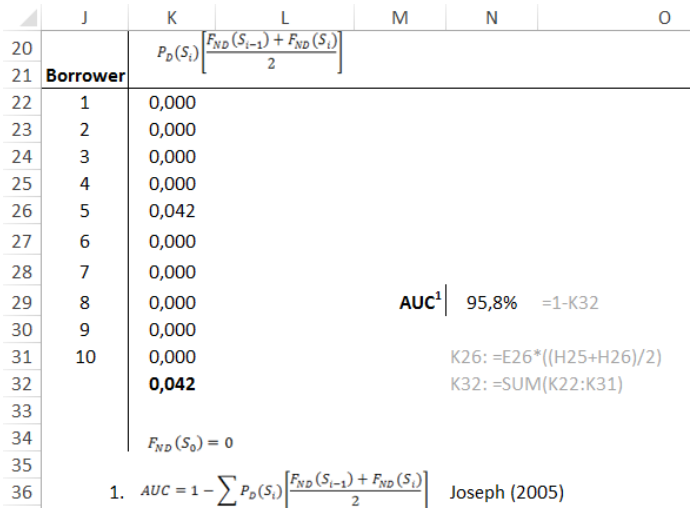
	A	B	C	D	E	F	G	H	I
1									
3	**Individual borrower**	**Prob. of default**	**Default status**						
4	i	PD	D						
5	1	8,2%	1						
6	2	8,1%	1						
7	3	7,9%	1						
8	4	7,8%	0						
9	5	2,1%	1						
10	6	2,0%	0						
11	7	1,9%	0						
12	8	0,2%	0						
13	9	0,1%	0						
14	10	0,0%	0						
18									
19	**Borrower**	**Counting**		**Probability (share)**		**Cumulative**		**Cumulative prob.**	
20	i	(goods) ND_i	(bads) D_i	(goods) P_{ND}	(bads) P_D	(goods)	(bads)	(goods) F_{ND}	(bads) F_D
22	1	0	1	0%	25%	0	1	0%	25%
23	2	0	1	0%	25%	0	2	0%	50%
24	3	0	1	0%	25%	0	3	0%	75%
25	4	1	0	17%	0%	1	3	17%	75%
26	5	0	1	0%	25%	1	4	17%	100%
27	6	1	0	17%	0%	2	4	33%	100%
28	7	1	0	17%	0%	3	4	50%	100%
29	8	1	0	17%	0%	4	4	67%	100%
30	9	1	0	17%	0%	5	4	83%	100%
31	10	1	0	17%	0%	6	4	100%	100%
32	**Totals**	**6**	**4**	**100%**	**100%**				
33								Fraction of non-defaults	Fraction of defaults
34									
35	Calculations are similar than on a bucket level.								

Source: modified from M. Joseph, *A PD Validation Framework for Basel II Internal Ratings-Based Systems*, 2005, p. 10.

Area under ROC curve (AUC)

We calculate this similarly as before on grade level, but now we calculate it on individual borrower level.

Figure 10. Calculation of AUC (individual level)

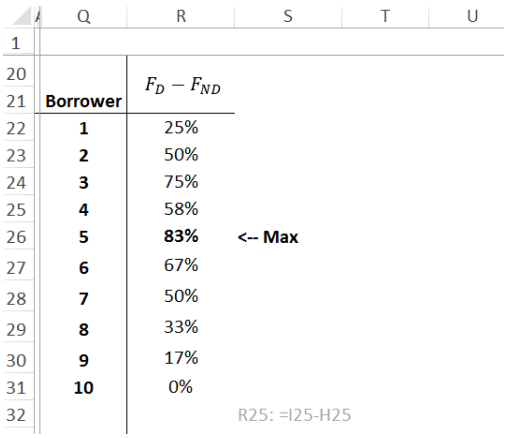


Source: modified from M. Joseph, *A PD Validation Framework for Basel II Internal Ratings-Based Systems*, 2005, p. 10.

Kolmogorov-Smirnov test (KS)

We calculate this similarly as before on grade level, but now we calculate it on individual borrower level.

Figure 11. Calculation of KS test statistic (individual level)

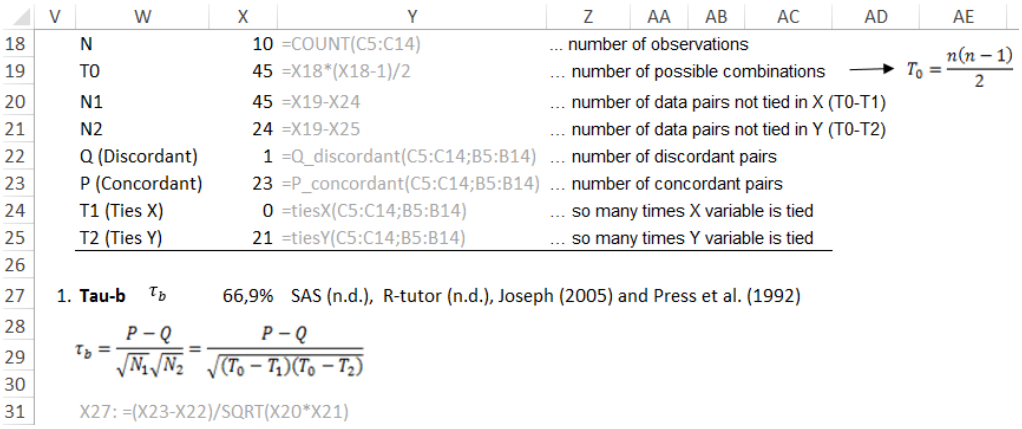


Source: modified from M. Joseph, *A PD Validation Framework for Basel II Internal Ratings-Based Systems*, 2005, p. 28.

Kendall tau-b correlation

The VBA code for calculating discordant/concordant pairs and for calculating ties is based on Joseph (2005) and modified by the other references cited below.

Figure 12. Calculation of Kendall tau-b correlation measure

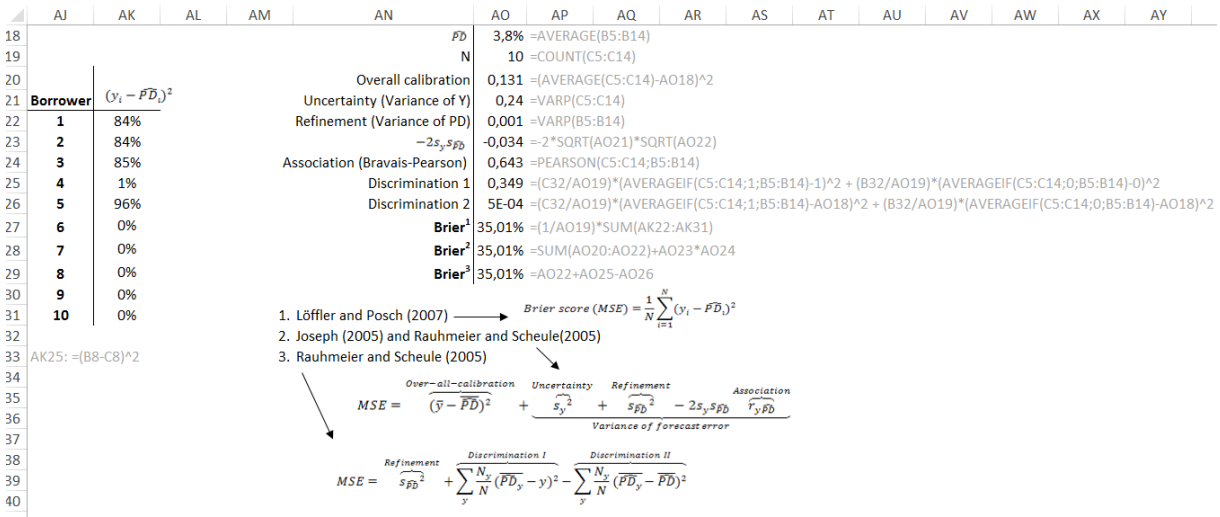


Source: modified from M. Joseph, *A PD Validation Framework for Basel II Internal Ratings-Based Systems*, 2005, p. 19.

Brier score

We compare three methods/decompositions of calculating Brier score (see references in the figure below).

Figure 13. Calculation of Brier score



Source: modified from M. Joseph, *A PD Validation Framework for Basel II Internal Ratings-Based Systems*, 2005, p. 17; G. Löffler & P. Posch, *Credit risk modelling using Excel and VBA*, 2007, p. 157.

Appendix E: Examples of calibration tests

Binomial tests

The implementation of binomial tests is based on Löffler and Posch (2007). Red colour (p-value < 1 %) means that predicted PDs in this grade underestimate the true default rates. Yellow colour corresponds to the p-value of < 5 %.

Figure 14. Basic tests for calibration

r	PD_{rt}	DR_{rt}	N_{rt}	D_{rt}	$\widehat{D_{rt}}$	p-values			
Rating grade in year 2014	Predicted PDs at 31.12.2013 (calibrated by historical average DR)	Observed default rates in year 2014	Number of borrowers at 31.12.2013	Number of observed defaults in year 2014	Number of predicted defaults in year 2014	Binomial	Normal	One-factor	
A1	0,22%	0,22%	465	1	1	26,4%	50,0%	35,0%	asset corr.= 7,0%
A2	0,16%	0,56%	1253	7	2	0,108%	0,0%	2,9%	alpha= 95%
A3	0,61%	0,89%	2463	22	15	3,2%	3,5%	20,0%	
B1	2,31%	1,97%	2643	52	61	86,6%	87,8%	49,0%	
B2	4,59%	4,20%	1546	65	71	74,5%	76,7%	47,0%	
C1	7,15%	6,94%	1384	96	99	59,7%	62,3%	44,5%	
C2	26,24%	24,35%	423	103	111	79,6%	81,2%	55,2%	

I5: =1-BINOMDIST(F5;E5;C5;1)
J5: =1-NORMSDIST((F5-C5*E5)/((C5*(1-C5))*E5)^0,5)
K5: =NORMSDIST((NORMSINV(C5)-NORMSINV(D5)*(1-M$4)^0,5)/M$4^0,5)

Source: modified from G. Löffler & P. Posch, *Credit risk modelling using Excel and VBA*, 2007, p. 160.

In our hypothetical example, our rating grades A2 and A3 are not well calibrated at 99 % and 95 % confidence levels respectively (because the p-values of one-sided tests are below 1 % or 5 %)). This can be seen in the table from columns F and G, where the actual number of defaults was much bigger compared to predicted number of defaults in rating grades A2 and A3. When assuming asset correlation in the one-factor model, the test is less conservative.

In the example below we illustrate the case where $N_{rt}=1253$ in rating grade $r=A2$. The two-sided null hypothesis would be rejected if we would predict more than 11 defaults or less then 4 defaults as can be seen in the figure below. We say that the predicted PDs overestimate or underestimate the true default probability.

In the cells E31:E52 (in the example below) we calculate the value of c_α as the minimum value of D_{rt} for which it holds that the summation term is less than $\alpha = 95$ %. The p-value for this value of $D_{rt}= 4$ is above 5 %, which means that for predicted 4 defaults we would not underestimate the true number of defaults, which is 7. If the p-value is less than 5 % we reject the null hypothesis at 95 % confidence level.

Figure 15. Calculation of Binomial test for calibration

1. $P(D_{rt}|N_{rt},PD_{rt}) = \binom{N_{rt}}{D_{rt}} PD_{rt}^{D_{rt}}(1-PD_{rt}^{D_{rt}})^{N_{rt}-D_{rt}}$

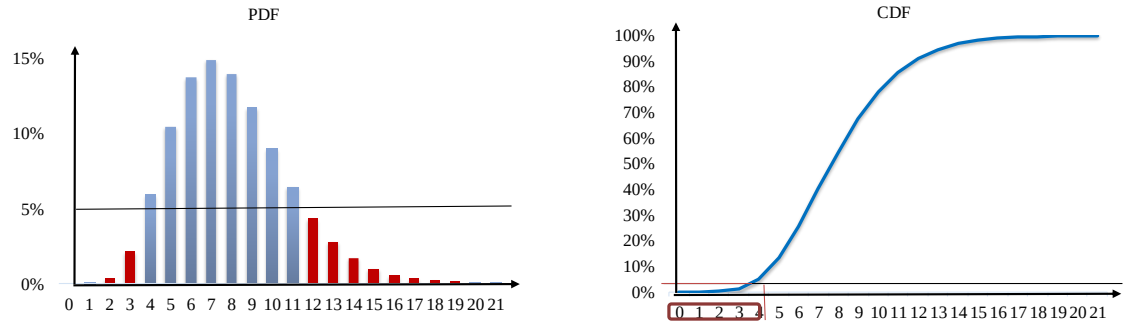
2. $c_\alpha = min\left\{D_{rt} : \sum_{i=D_{rt}}^{N_{rt}} \binom{N_{rt}}{i} PD_{rt}^{i}(1-PD_{rt}^{i})^{N_{rt}-i} \leq \alpha\right\}$

C31: =B31/$E$5
D31: = COMBIN($E$5;$F$5) * C31^$F$5 * (1-C31)^($E$5-$F$5)
or D31: =BINOMDIST($F$5;$E$5;C31;FALSE)
E31: =BINOMDIST($F$5;$E$5;C31;TRUE)
F31: =1-E31

predited D_{rt}	PD_{rt}	*eq. 1*		p-value	*eq. 2*
pred D	**PD**	**PDF**	**CDF**	**1-CDF**	c_α
0	0,00%	0,0%	100,0%	0,000%	underestimation
1	0,08%	0,0%	100,0%	0,001%	underestimation
2	0,16%	0,3%	99,9%	0,108%	underestimation
3	0,24%	2,2%	98,8%	1,180%	underestimation
4	**0,32%**	**5,9%**	**94,9%**	**5,085%**	**<-- min is D=4 as CDF < 95%** *(eq. 2)*
5	0,40%	10,5%	86,7%	13,295%	
6	0,48%	13,8%	74,4%	25,569%	
7	0,56%	14,9%	59,9%	40,129%	
8	0,64%	14,0%	45,3%	54,749%	
9	0,72%	11,7%	32,3%	67,695%	
10	0,80%	9,0%	21,9%	78,086%	
11	0,88%	6,4%	14,2%	85,795%	
12	0,96%	4,3%	8,8%	91,154%	
13	1,04%	2,8%	5,3%	94,685%	
14	1,12%	1,7%	3,1%	96,906%	overestimation
15	1,20%	1,0%	1,8%	98,249%	overestimation
16	1,28%	0,6%	1,0%	99,034%	overestimation
17	1,36%	0,3%	0,5%	99,479%	overestimation
18	1,44%	0,2%	0,3%	99,725%	overestimation
19	1,52%	0,1%	0,1%	99,858%	overestimation
20	1,60%	0,0%	0,1%	99,927%	overestimation
21	1,68%	0,0%	0,0%	99,964%	overestimation

In the figures below we show the probability density function (PDF) and the cumulative distribution function (CDF) of a binomial distribution (see cells D31:D52 and F31:F52).

Figure 16. PDF and CDF of a binomial distribution



The calculations for normal approximation to the binomial test are similar (the formulas correspond to the ones in the main text).

Figure 17. Calculation of Normal approximation to binomial test

1. $z = \frac{D_{rt} - N_{rt} PD_{rt}}{\sqrt{N_{rt} PD_{rt}(1 - PD_{rt})}}$

2. $\Phi = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

3. $c_\alpha = min\{D_{rt} : \Phi(z) \leq \alpha\}$

Y31: =X31/$E$5
Z32: =($F$5-X32)/(SQRT($E$5*Y32*(1-Y32)))
AA32: =1/(SQRT(2*PI())) * EXP(1)^(-(Z32^2)/2)
or AA32: =NORM.S.DIST(Z32;FALSE)
AB32: =NORM.S.DIST(Z32;TRUE)
AC32: =1-AB32

predited D_{rt}	PD_{rt}	*eq. 1*		$\Phi(z)$		*eq. 3*
pred D	**PD**	**z**	**PDF**	**CDF**	**1-CDF**	c_α
0	0,00%	NA	NA	NA	NA	underestimation
1	0,08%	6,00	0,0%	100,0%	0,0%	underestimation
2	0,16%	3,54	0,1%	100,0%	0,0%	underestimation
3	0,24%	2,31	2,8%	99,0%	1,0%	underestimation
4	**0,32%**	**1,50**	**12,9%**	**93,4%**	**6,6%**	**<-- min is D=4 as CDF < 95 %** *(eq. 3)*
5	0,40%	0,90	26,7%	81,5%	18,5%	
6	0,48%	0,41	36,7%	65,9%	34,1%	
7	0,56%	0,00	39,9%	50,0%	50,0%	
8	0,64%	-0,35	37,5%	36,1%	63,9%	
9	0,72%	-0,67	31,9%	25,2%	74,8%	
10	0,80%	-0,95	25,3%	17,0%	83,0%	
11	0,88%	-1,21	19,2%	11,3%	88,7%	
12	0,96%	-1,45	13,9%	7,3%	92,7%	
13	1,04%	-1,67	9,8%	4,7%	95,3%	overestimation
14	1,12%	-1,88	6,8%	3,0%	97,0%	overestimation
15	1,20%	-2,08	4,6%	1,9%	98,1%	overestimation

For one-factor binomial model it is similar (the formulas correspond to the main text):

Figure 18. Calculation of One-factor binomial test

1. $$F_t = \frac{\Phi^{-1}(PD_{rt}) - \sqrt{1-\rho}\,\Phi^{-1}(\frac{D_{rt}}{N_{rt}})}{\sqrt{\rho}}$$

AV32: =AU32/$E$5
AW32: =(NORMSINV(AV32)-SQRT(1-$M$4)...
...*NORMSINV($F$5/$E$5))/SQRT($M$4)
AX32: =NORM.S.DIST(AW32;TRUE)

2. $$c_\alpha = min\{D_{rt} \ : \ \Phi(F_t) > 1 - \alpha\}$$

predited D_{rt}	PD_{rt}		$\Phi[F_t]$	*eq. 2*	
pred D	**PD**	**F**	**CDF**	c_α	
0	0,00%	NA	NA	underestimation	
1	0,08%	-2,68	0,4%	underestimation	
2	0,16%	-1,90	2,9%	underestimation	
3	**0,24%**	**-1,41**	**7,9%**	**<-- min is D=3 as CDF > 5 %** *(eq. 2)*	
4	0,32%	-1,06	14,5%		
5	0,40%	-0,78	21,8%		
6	0,48%	-0,54	29,3%		
7	0,56%	-0,34	36,6%		
8	0,64%	-0,16	43,5%		
9	0,72%	0,00	49,8%		
10	0,80%	0,14	55,6%		
11	0,88%	0,27	60,7%		
12	0,96%	0,39	65,3%		
13	1,04%	0,51	69,4%		
14	1,12%	0,61	73,0%		
15	1,20%	0,71	76,2%		

Chi-square Hosmer-Lemeshow test (HSLS)

In the figure below we can see that our model is not well calibrated. The rating grade A2 increases the Chi-value the most (as expected). Higher Chi-value leads to the lower Chi-probability. Since the Chi-probability measures the chance of the null hypothesis being true, we can see that our model is not performing well (the chi-probability is 0,5 %). The implementation is partially based on Joseph (2005).

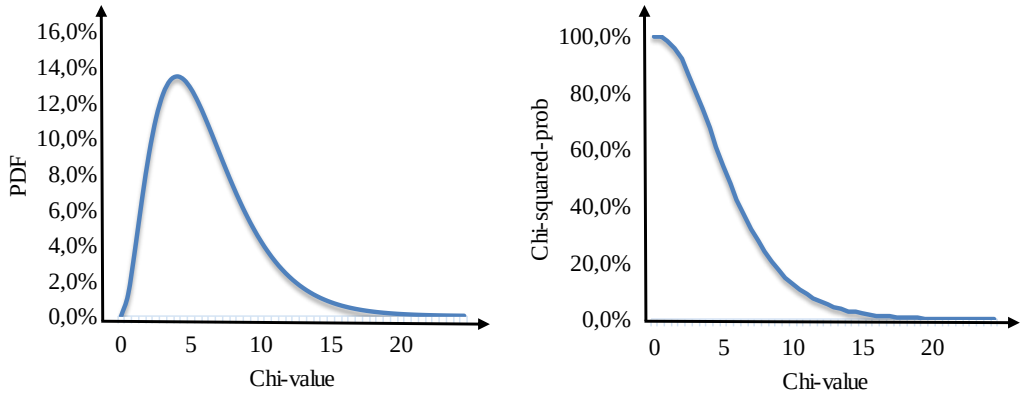
Figure 19. Calculation of HSLS test

	B	C	D	E	F	G	H	I	J	K	L	M
1												
2	r	PD_{rt}	DR_{rt}	N_{rt}	D_{rt}	$\widehat{D_{rt}}$						
3	Rating grade in year 2014	Predicted PDs at 31.12.2013 (calibrated by historical average DR)	Observed default rates in year 2014	Number of borrowers at 31.12.2013	Number of observed defaults in year 2014	Number of predicted defaults in year 2014	$\frac{(D_r - N_r\widehat{PD_r})^2}{N_r\widehat{PD_r}(1-\widehat{PD_r})}$					
4	A1	0,22%	0,22%	465	1	1	0,00			DF= 6		
5	A2	0,16%	0,56%	1253	7	2	12,52			Chi-squared Value = 18,58		
6	A3	0,61%	0,89%	2463	22	15	3,29			Chi-squared Probability = 0,49%		
7	B1	2,31%	1,97%	2643	52	61	1,36					
8	B2	4,59%	4,20%	1546	65	71	0,53					
9	C1	7,15%	6,94%	1384	96	99	0,10					
10	C2	26,24%	24,35%	423	103	111	0,78			H5: =((F5-E5*C5)^2) / (E5*C5*(1-C5))		
11										L4: =(COUNT(C4:C10)-1)*(COUNT(F4:G4)-1)		
12										L5: =SUM(H4:H10)		
13	1.	$\chi_t^2 = \sum_{r=1}^{R} \frac{(D_r - N_r\widehat{PD_r})^2}{N_r\widehat{PD_r}(1-\widehat{PD_r})}$		BIS (2005)						L6: =CHIDIST(L5;L4)		
14												
15												

Source: modified from BIS, *Studies on the Validation of Internal Rating Systems*, 2005, p. 52.

Below we plotted the relation between different Chi-values (with 6 degrees of freedom) and Chi-probabilities. We also graph the probability density function (PDF).

Figure 20. Chi-values and PDF



Note that the Chi-probability gets lower for both, for the underestimation and for the overestimation effect.

Spiegelhalter test (SPGH)

For the SPGH test (assume one period *t*) we show an example of well calibrated model, since the p-value of one-sided test is above 5 %.

Figure 21. Calculation of SPGH test

Individual borrower i	Prob. of default PD_i	Default status y_i	Brier score (MSE)	E(MSE)	VaR(MSE)	z-value z_{SPGH}	p-value $\Phi(z)$
1	82,0%	1	3%	0,15	0,06	**0,417**	**33,83%**
2	81,0%	1	4%	0,15	0,06		model is OK
3	79,0%	1	4%	0,17	0,06		
4	78,0%	0	61%	0,17	0,05	**N=** 10	
5	21,0%	1	62%	0,17	0,06	**alpha =** 95%	
6	20,0%	0	4%	0,16	0,06	**1-alpha=** 5%	
7	19,0%	0	4%	0,15	0,06	**predicted D=** 3,83	
8	2,0%	0	0%	0,02	0,02	**actual D=** 4	
9	1,0%	0	0%	0,01	0,01		
10	0,0%	0	0%	0,00	0,00		
			0,14	**0,115**	**0,004**		

1. $MSE = Brier\ score = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{PD}_i)^2$
2. $E\left(MSE_{PD_i=\widehat{PD}_i}\right) = \frac{1}{N}\sum_{i=1}^{N} PD_i(1-PD_i)$
3. $Var\left(MSE_{PD_i=\widehat{PD}_i}\right) = \frac{1}{N^2}\sum_{i=1}^{N}(1-2PD_i)^2 PD_i(1-PD_i)$
4. $Z_{SPGH} = \frac{MSE - E\left(MSE_{PD_i=\widehat{PD}_i}\right)}{\sqrt{Var\left(MSE_{PD_i=\widehat{PD}_i}\right)}}$

D4: =(B4-C4)^2
D14: =1/H7 * SUM(D4:D13)
E4: =B4*(1-B4)
E14: =(1/H7)*SUM(E4:E13)
F4: =E4*(1-2*B4)^2
F14: =(1/(H7^2))*SUM(F4:F13)
G4: =(D14-E14)/SQRT(F14)
H4: =1-NORM.S.DIST(G4;TRUE)
H7: =COUNT(A4:A13)
H10: =SUM(B4:B13)
H11: =SUM(C4:C13)

Source: modified from Rauhmeier (in Engelmann & Rauhmeier, *The Basel II Risk Parameters*, 2011, p. 323).

If we would insert the value of 20 % (assumed one rating grade average PD) in all cells B4:B13, then we would obtain the same p-value as in the binomial test.

Appendix F: One-factor model derivation

To derive the formula we need to develop some statistical background first. Rauhmeier (in Engelmann & Rauhmeier, 2011, p. 317) starts with defining the probability of default (and non-default) with the logistic regression model (note that the notation is similar as in chapter for logistic regression):

$$PD_i(x_i)=P(y_i=1|x_i)=F(x'_i\beta)$$

$$1\text{-}PD_i(x_i)=P(y_i=0|x_i)=1\text{-}F(x'_i\beta) \tag{19}$$

$$score=x'_i\beta=x_1\beta_1+x_2\beta_2+\ldots+x_K\beta_K$$

On scores we would apply logistic distribution to obtain probabilities of default as explained in the thesis. Now suppose that behind the observable depended variable y_i there is a non observable (latent), continuous variable $\tilde{y}_i$. Its value depends on the value of the risk drivers x_i. Latent variables determine an observed, discrete outcome. If this latent variable falls below the (also latent) threshold θ_i then we predict $y_i=1$ (Rauhmeier in Engelmann & Rauhmeier, 2011, p. 318):

$$\begin{aligned} y_i=1 &\Leftrightarrow \tilde{y}_i=x'_i\beta+\tilde{\varepsilon}_i\le\theta_i \\ y_i=0 &\Leftrightarrow \tilde{y}_i=x'_i\beta+\tilde{\varepsilon}_i>\theta_i \end{aligned} \tag{20}$$

The error term $\tilde{\varepsilon}_i$ allows for randomness and accounts for idiosyncratic risk factors that are not covered in x_i. Then we can write:

$$\begin{aligned} PD_i(x_i)=P(y_i=1|x_i)=P(\tilde{y}_i\le\theta_i)= \\ =P(\tilde{\varepsilon}_i\le\theta_i-x'_i\beta)=F(\theta_i-x'_i\beta)=F(\tilde{\theta}_i) \end{aligned} \tag{21}$$

It follows that: $\tilde{\theta}_i=\Phi^{-1}(PD_{rt})$. The variable that we are now interested in is the **default correlation** δ. It can be derived as in Löffler and Posch (2007, p. 104) and in BIS (2005, p. 48):

$$Corr(y_i,y_j)=\delta_{ij}=\frac{Cov(y_i,y_j)}{\sigma(y_i)\sigma(y_j)}=\frac{PD_{ij}-PD_iPD_j}{\sqrt{PD_i(1-PD_i)PD_j(1-PD_j)}} \tag{22}$$

because:

$$E(y_i)=PD_i=\Phi^{-1}(\theta) \tag{23}$$

$$\begin{aligned} Var(y_i)=\sigma_i^2=Prob(y_i=1)(1-E(y_i))^2+Prob(y_i=0)(0-E(y_i))^2= \\ =PD_i(1-PD_i)^2+(1-PD_i)(0-PD_i)^2= \end{aligned} \tag{24}$$

$$=PD_i(1-PD_i)^2+PD_i^2(1-PD_i)=$$
$$=PD_i(1-PD_i)$$

$$\sigma(y_i)\sigma(y_j)=\sqrt{PD_i(1-PD_i)PD_j(1-PD_j)} \quad (25)$$

$$Cov(y_i,y_j)=E(y_iy_j)-E(y_i)E(y_j)=PD_{ij}-PD_iPD_j \quad (26)$$

We will later rewrite it as:

$$Corr(y_i,y_j)=\frac{\Phi_2(\theta,\theta,\rho)-PD_i^2}{PD_i(1-PD_i)}=\delta \quad (27)$$

where PD_i is the default probability $Prob(y_i=1)$ for borrower i, and PD_{ij} is the joint default probability $Prob(y_i=1,y_j=1)=Prob(A_i\leq\theta_i,A_j\leq\theta_J)$. The expected value of Bernoulli variable with success probability PD_i is $E(y_i)=PD_i$, the variance is $PD_i(1-PD_i)$. We also use the standard formulas for covariance $E(X_1X_2)=E(X_1)E(X_2)+cov(X_1,X_2)$ and for variance $Var(X)=E[(X-E(X))^2]=E[X^2]-E[X]^2$. However, it is not practical to calculate pairwise default correlations[108]. Hence we use the simplification based on asset correlations ρ instead (they can be transformed to default correlations). We represent defaults as a function of continuous variables and then impose structure on these variables (instead of on all the default correlations). Let us define these variables as A_i, $i=1,\dots N$. Building on our statistical framework we can follow Rauhmeier (in Engelmann & Rauhmeier, 2011, p. 319), Löffler and Posch (2007, p. 104) and BIS (2005, p. 48) and extend the model by integrating dependencies in our model. The natural choice for that is to use the one-factor model for the value of assets A_i. At the end of the period, the value A_i of the assets of borrower i depends on a systematic risk factor F common to all borrowers and an idiosyncratic (individual) risk factor ε_i that is specific to the borrower. It is assumed that the two risks are independent between each other and that idiosyncratic risk is independent for the two different borrowers $Cov(\varepsilon_i,\varepsilon_j)=0$. Now we are building the one-factor model. The model describes the creditworthiness of a borrower which is expressed by change in the asset value. We further assume that: $A_i,F,\varepsilon_i\sim N(0,1)$, $Cov(F,\varepsilon_i)=0$ and $Cov(\varepsilon_i,\varepsilon_j)=0,\ i\neq j$. Under these assumptions A_i can be modelled as:

$$A_i=\sqrt{1-\rho}\varepsilon_i+\sqrt{\rho}F \quad (28)$$

Now we can split the error term $\tilde{\varepsilon}_i$ in the components ε_i and F. The formulation now becomes (Rauhmeier in Engelmann & Rauhmeier, 2011, p. 319):

$$y_i=1\Leftrightarrow\tilde{y}_i=x_i'\beta+\sqrt{\rho}F+\sqrt{1-\rho}\varepsilon_i\leq\theta_i \quad (29)$$

[108] If we would have 1000 borrowers we would need to calculate $\frac{1000^2-1000}{2}=499.500$ correlations.

$$y_i=0 \Leftrightarrow \tilde{y}_i=x_i'\beta+\sqrt{\rho}F+\sqrt{1-\rho}\varepsilon_i>\theta_i$$

where F, A_i and ε_i are the normally distributed random variables with mean zero and standard deviation one. Similar as before we define that a borrower defaults if the value of his assets is below some threshold θ_i at the end of the period: $x_i'\beta+A_i \leq \theta_i$. We then define the default indicator y_i as:

$$y_i=\begin{cases}1, & x_i'\beta+A_i \leq \theta_i \\ 0, & else\end{cases} \tag{30}$$

Now we can define the **asset correlation** ρ as in Rauhmeier (in Engelmann & Rauhmeier, 2011, p. 319) and Löffler and Posch (2007, p. 105):

$$\begin{aligned} \sigma_i^2=Var(A_i)=Var\left(\sqrt{\rho}F+\sqrt{1-\rho}\varepsilon_i\right)= \\ =Cov\left(\sqrt{\rho}F+\sqrt{1-\rho}\varepsilon_i,\sqrt{\rho}F+\sqrt{1-\rho}\varepsilon_i\right)=Cov\left(\sqrt{\rho}F,\sqrt{\rho}F\right)+Cov\left(\sqrt{1-\rho}\varepsilon_i,\sqrt{1-\rho}\varepsilon_i\right)= \\ =\sqrt{\rho}\sqrt{\rho}\cdot Var(F)+\sqrt{1-\rho}\sqrt{1-\rho}Var(\varepsilon_i)= \\ =(\sqrt{\rho})^2+(\sqrt{1-\rho})^2=1 \end{aligned} \tag{31}$$

$$\begin{aligned} \sigma_{ij}=Cov\left(A_i,A_j\right)=Cov\left(\sqrt{\rho}F+\sqrt{1-\rho}\varepsilon_i,\sqrt{\rho}F+\sqrt{1-\rho}\varepsilon_j\right)= \\ =Cov\left(\sqrt{\rho}F,\sqrt{\rho}F\right)=\sqrt{\rho}\sqrt{\rho}\cdot Var(F)=(\sqrt{\rho})^2=\rho \end{aligned} \tag{32}$$

$$\rho_{ij}=Corr\left(A_i,A_j\right)=\frac{Cov\left(A_i,A_j\right)}{Var(A_i)Var\left(A_j\right)}=\frac{\rho}{1}=\rho \tag{33}$$

because $F,\varepsilon_i \sim N(0,1)$, $Cov(F,\varepsilon_i)=0$ and $Cov(\varepsilon_i,\varepsilon_j)=0$, $i \neq j$. We use the formulas for variance and covariance as above.

Next, we present the formula to calculate the default correlation. We make two assumptions and continue from the results for default correlations from above. First we assume that all borrowers have the same probability of default (are in the same rating grade) and that the same threshold $\theta:=\theta_i$ applies to all borrowers. Secondly, we assume that the asset correlation is the same for all pairs of borrowers. Now we can obtain the following properties for the default indicator y_i (BIS 2005, p. 48):

$$E\left(y_i\right)=PD_i=\Phi^{-1}(\theta) \tag{34}$$

$$Var\left(y_i\right)=PD_i(1-PD_i) \tag{35}$$

$$Corr\left(y_i,y_j\right)=\frac{PD_{ij}-PD_iPD_j}{\sqrt{PD_i(1-PD_i)PD_j(1-PD_j)}}=\frac{\Phi_2(\theta,\theta,\rho)-PD^2}{PD(1-PD)}=\delta \tag{36}$$

where $\Phi_2(.,.,\rho)$ denotes the bivariate standard normal distribution function with asset correlation ρ. Here we also use the formulas for the default probability $Prob(A_i \leq \theta_i) = PD_i = \Phi(\theta_i)$ and for the joint default probability $Prob(A_i \leq \theta_i, A_j \leq \theta_j) = PD_{ij} = \Phi_2(\theta_i, \theta_j, \rho_{ij})$.

Now we derive **the conditional probability of default** (conditional on the realization of the common factor F) (this is the approximation to the distribution of D_{rt} by using the ideas of the Vasicek distribution) (BIS, 2005, p. 49; Rauhmeier in Engelmann & Rauhmeier, 2011, p. 319):

$$\overline{PD_{rt}} \xrightarrow{n\to\infty} PD_i(x_i,F) = P(y_i = 1 \mid x_i, F) = P(\tilde{y}_i \leq \theta_i) = \tag{37}$$

$$= P\left(x_i'\beta + \sqrt{\rho}F + \sqrt{1-\rho}\varepsilon_i \leq \theta_i\right) =$$

$$= F\left(\varepsilon_i \leq \frac{\theta_i - x_i'\beta - \sqrt{\rho}F}{\sqrt{1-\rho}}\right) =$$

$$= \Phi\left[\frac{\tilde{\theta}_i - \sqrt{\rho}F_t}{\sqrt{1-\rho}}\right] = \Phi\left[\frac{\Phi^{-1}(PD_{rt}) - \sqrt{\rho}F_t}{\sqrt{1-\rho}}\right]$$

where $\tilde{\theta}_i = \Phi^{-1}(PD_{rt})$ and $\tilde{\theta}_i = \theta_i - x_i'\beta$ from above. This is the formula that we use in the main text.

Appendix G: Variable catalogue

Table 1. Variable catalogue

ID	Variable [exluded if HP]
F1	region
F2	sector
F3	legal form
F4	size (regulatory)
DEF	default
X1	total sales
X2	total_revenue
X3	total_revenue2
X4	Liabilities / Assets
X5	(LT_liabilities+current_liabilities+provisions)/assets
X6	Equity / Assets
X7	Equity / Assets*
X8	(Equity + provisions and long-term accruals + long-term liabilities) / (assets + inventories + 1)
X9	Tangible Assets / (Liabilities+1)
X10	Long term Liabilities / Assets
X11	Bank Debt / Assets
X12	Bank Debt / Assets*
X13	(Assets - Bank Debt)/(Bank Debt+1)
X14	(Assets - Bank Debt)*/(Bank Debt+1)
X15	Bank debt / (Liabilities+1)
X16	(equity+provisions)/(bank_debt+1)
X17	Short Term Debt/(Equity+min(Equity)) [Y]
X18	Equity/(Short Term Debt+1)
X19	Equity/(Total Liabilities+1)
X20	(bank_debt-cash)/net_sales
X21	(bank_debt-cash)/assets
X22	ST_bank_debt/assets
X23	(Equity) / (intangible assets and long-term accruals + tangible assets + 1)
X24	(Long-term liabilities + short-term liabilities + ST accrauls) / (assets)
X25	(equity+provisions)/(intangible_assets+tangible_assets+1)
X26	(liabilities*last_assets)/((last_liabilities+1)*assets)
X27	(bank_debt*last_assets)/((last_bank_debt+1)*assets)
X28	(gross profit+material_costs+financial_income)/(gross profit_last+material_costs_last+financial_income_last+1)
X29	(cash-cash_last)/(cash_last+1)
X30	Net Sales / (Last Net Sales+1)
X31	assets/last_assets
X32	(assets-last_assets)/last_assets
X33	(retained_earnings)/(retained_earnings_last+min(retained_earnings_last)) [Y]
X34	(employees-employees_last)/employees_last
X35	(net_sales+ulpl+other_operational_income+financial_income)/ (net_sales_last+ulpl_last+other_operational_income_last+financial_income_last+1)
X36	(net_income+amortization)/(net_income_last+amortization_last+min(net_income_last+amortization_last)) [Y]
X37	equity/(last_equity+min(last_equity)) [Y]
X38	Current Assets / (Current Liabilities+1)
X39	Current Assets / (Liabilities+1)
X40	Working Capital / Assets
X41	Current Liabilities / Assets
X42	Current Assets / Assets
X43	(LT assets) / (assets)
X44	Cash / Assets
X45	Working Capital / Net Sales
X46	Cash / Net Sales
X47	Current Assets / Net Sales
X48	Quick Assets / Net Sales
X49	Short Term Bank Debt / (Bank Debt+1)
X50	Cash / (Current Liabilities+1)
X51	(cash+ST_fin)/(current_liabilities+1)
X52	Working Capital / (Current Liabilities+1)
X53	Quick Ratio
X54	Intangible/Total Assets
X55	(Current assets - inventories + short-term active accruals) / (current liabilities + ST accruals+ 1)

ID	Variable [exluded if HP]
X56	(Current assets + ST accruals) / (current liabilities + ST accruals + 1)
X57	(current_assets-inventory)/(current_liabilities+1)
X58	(current_assets+kacr)/(current_liabilities+kpcr+1)
X59	(inventory+ST_receivables+ST_fin+cash+kacr)/(current_liabilities+kpcr+1)
X60	Inventory / (EBIT+min(EBIT)) [Y]
X61	Inventory / Net Sales
X62	Inventory / (Material Costs+1)
X63	Accounts Receivable / Net Sales
X64	ST_receivables/net_sales
X65	EBIT/(Accounts Receivable+1)
X66	Accounts Receivable / (Liabilities+1)
X67	Accounts Receivable / (Liabilities*+1)
X68	Accounts Receivable / (Material Costs+1)
X69	Accounts Payable / (Material Costs+1)
X70	Accounts Payable / Net Sales
X71	Inventory / (Accounts Receivable+1)
X72	total_revenue/total_expenses
X73	(net_income+odpisi)/total_expenses2
X74	(net_income+odpisi)/total_expenses
X75	total_revenue2/total_expenses2
X76	Personnel Costs / Net Sales
X77	Personnel Costs / (EBIT+min(EBIT)) [Y]
X78	EBIT/(Personnel Costs+1)
X79	Personnel Costs / ((Net Sales-Material Costs)+min(Net Sales-Material Costs)) [Y]
X80	EBIT/Material Costs
X81	EBIT / Assets
X82	EBIT / Assets*
X83	EBITDA / Assets*
X84	EBIT / Net Sales
X85	(EBIT+Interest Income)/(EBIT+min(EBIT)) [Y]
X86	(EBIT + Interest Income) / Assets
X87	Ordinary Business Income/Assets
X88	Ordinary Business Income/Assets*
X89	Net Income / Assets
X90	Net Income / Assets*
X91	Net Income / Net Sales
X92	Retained Earnings / Assets
X93	Ebitda/Total Assets
X94	(total profit + financial income - financial expenses from impairments and investment write-offs) / (Equity + long-term financial liabilities + short-term financial liabilities + min(Equity + long-term financial liabilities + short-term financial liabilities)) [Y]
X95	(EBIT - operating loss + financial income - financial expenses from impairments and investment write-offs) / (assets)
X96	EBIT/(Gross operating income+min(Gross operating income)) [Y]
X97	(EBIT+write-offs)/(Gross operating income+min(Gross operating income)) [Y]
X98	(Gross operating income)/(Gross operating income +min(Gross operating income)) [Y]
X99	(Gross operating income) / (Operating expenses + 1)
X100	(Gross operating income + finančni prihodki) / (poslovni odhodki + finančni odhodki + 1)
X101	(Gross operating income + finančni prihodki + drugi prihodki) / (poslovni odhodki + finančni odhodki + drugi odhodki)
X102	(net_income+amortization)/(LT_liabilities+current_liabilities+1)
X103	Gross operating income/(financial_income+Gross operating income+min(financial_income+Gross operating income)) [Y]
X104	Gross operating income/(financial_income+Gross operating income+other income+min(financial_income+Gross operating income+other income)) [Y]
X105	Gross operating income/(inventory+ST_receivables-ST_operational_liabilities+1)
X106	(Net sales revenue + capitalized own products and services + subsidies , grants, allowances , compensations and other revenue associated with products and services + other operating income + financial income) / (assets)
X107	(Net sales revenue + capitalized own products and services + subsidies , grants, allowances , compensations and other revenue associated with products and services + other operating income) / (assets - long-term investments - short-term investments)
X108	(Net sales revenue + capitalized own products and services + subsidies , grants, allowances , compensations and other revenue associated with products and services + other operating income + financial income) / (current assets + ST accruals + 1)
X109	(Net sales revenue + capitalized own products and services) / (short-term operating liabilities + short-term operating liabilities to group companies + other current operating liabilities + 1)
X110	(Net sales revenue + capitalized own products and services + subsidies , grants, allowances , compensations and other revenue associated with products and services + other operating income) / (current assets + ST accruals + 1)

ID	Variable [exluded if HP]
X111	(Net sales revenue + capitalized own products and services + subsidies , grants, allowances , compensations and other revenue associated with products and services + other operating income) / (current assets + ST accruals - ST financial investments + 1)
X112	(Net sales revenue + capitalized own products and services) / (short-term receivables + 1)
X113	Interest Expenses / (EBIT+min(EBIT)) [Y]
X114	EBIT / (Interest Expenses+1)
X115	Interest Expenses/(Ebitda+min(ebitda)) [Y]
X116	bank_debt/(ebitda+min(ebitda)) [Y]
X117	ebitda/(bank_debt+1)
X118	Gross operating income/(bank_debt+1)
X119	bank_debt/(Gross operating income+min(Gross operating income)) [Y]
X120	(net_income+amortization)/(LT_liabilities+current_liabilities+1)
X121	(Net profit + depreciation) / (liabilities + long- term liabilities + ST accruals + 1)
X122	(Net sales revenue + capitalized own products and services) / (current liabilities + 1)
X123	(Net profit + depreciation) / (current liabilities + 1)
X124	operating_cashflow/(ST_bank_debt+1)
X125	ST_bank_debt/(operating_cashflow+min(operating_cashflow)) [Y]
X126	operating_cashflow/(bank_debt+1)
X127	bank_debt/(operating_cashflow+min(operating_cashflow)) [Y]
X128	operating_cashflow/(interest_expenses+1)
X129	interest_expenses/(operating_cashflow+min(operating_cashflow)) [Y]
X130	Net Sales / Assets
X131	Net Sales / Assets*
X132	net_sales/(ST_operational_liabilities+1)
X133	net_sales/(inventory+ST_receivables-ST_operational_liabilities+1)

The highlighted variables are the variables that were chosen in the final logistic regression model 1.

Appendix H: Abbreviations

ANN – artificial neural network
RF – random forest
DE – differential evolution (optimisation algorithm)
AR – accuracy ratio
AUC – area under the ROC curve
KS – Kolmogorov-Smirnov test
KL – Kullback-Leibler Divergence
HSLS – Hosmer-Lemeshow Chi-square test
SPGH – Spiegelhalter test
TTC – through-the-cycle rating philosophy
PIT – point-in-time rating philosophy
$k=1,2,....K$ … features/independent variables
$i=1,2,....N$ … borrowers/observations
$r \epsilon \{1,....,R\}$ … rating grades
y_i … an indicator variable that takes the value *1* if borrower *i* defaulted and *0* otherwise
$\widehat{PD_{rt}}$ or PD_i … predicted probability of default (proxy for true default rate in the future year)
$DR_{rt}=\overline{\widehat{PD}_{rt}}$ … historical default rate (observed defaults)
N_{rt} … number of borrowers in the rating grade *r* in year *t*
$D_{rt}= N_{rt,y=1}$ … number of observed defaults in rating grade *r* ($y=1$) in year *t*
S_i … score of the borrower *i*
N_D … number of defaulted companies
N_{ND} … number of non-defaulted companies
F_D … cumulative probability of defaulters
F_{ND} … cumulative probability of non-defaulters
$HR(C)$ … hit rate - number of defaulters predicted correctly (with the cut-off value *C*)
$FAR(C)$ … false alarm rate – number of false alarms (non-defaulters classified incorrectly as defaulters)
P_D … probability of bads (defaulters)
P_{ND} … probability of goods (non-defaulters)
c_α c_{low} c_{high} … critical value (refers to D – number defaults) D_α
α … confidence level (i.e. 95 %)
F… common/systematic risk factor
ε_i … idiosyncratic individual risk factor
ρ … asset correlation
δ … default correlation
$\phi(x)$… density of the standard normal distribution
Φ … cumulative distribution function of the standard normal distribution (CDF)
Φ^{-1} … inverse cumulative distribution function for the standard normal distributions
$\Phi_2(,,\rho)$ … CDF of the bivariate standard normal distribution with asset correlation ρ

Appendix I: Summary in Slovenian (povzetek v slovenskem jeziku)

Naloga prikazuje razvoj in validacijo celotnega cikla bonitetnega sistema za ocenjevanje neplačil podjetij. Tematika je aktualna, saj so v polnem razmahu priprave na nov mednarodni standard računovodskega poročanja 9 (angl. *International Financial Reporting Standards 9*), ki daje poseben poudarek modelom za ocenjevanje kreditnega tveganja. Prav tako se v slovenskem bančnem prostoru pojavljajo premiki, ki so spodbujeni s strani centralne banke, t.j. smernice za izračunavanje stopenj stečajev in stopenj izgub, ki banke silijo v zbiranje podatkov kot osnovo za naprednejše modeliranje kreditnega tveganja. V pripravi je centralni kreditni register, ki bo v prihodnosti lahko sestavni del validacije kreditnih modelov (za potrebe benchmarkinga). Slovenske banke imajo sedaj na voljo dovolj dolge časovne vrste, ki so lahko osnova za razvoj naprednih pristopov k ocenjevanju kapitalskih zahtev. Poleg tega je v zadnjem obdobju na spletu mogoče zaslediti vse polno objav s področji podatkovne analitike, umetne inteligence in strojnega učenja. Ta področja so z razvojem metod in razmahom velikih količin podatkov postala pomembna področja zanimanja v praktično vseh gospodarskih panogah. Še posebej v zadnjem obdobju so močno prodrla tudi v finance in bančništvo. V nalogi smo testirali njihovo uporabnost v posameznih korakih modeliranja verjetnosti stečajev podjetij.

V nalogi se v vsakem koraku nanašamo tudi na zakonodajo in probleme, ki se pojavljajo v praksi. Aktualni regulatorni okvir predstavlja direktiva Evropske Komisije (angl. *European Commission*, v nadaljevanju EC; 2013). Odmevna so tudi poročila Evropskega bančnega organa (angl. *European Banking Authority*, v nadaljevanju EBA; 2013a; 2013b; 2013c), ki poročajo o pomanjkljivostih in nekonsistentnostih zakonodajnega okvira za nadzor nad implementacijo naprednega pristopa (angl. *Internal ratings-based approach*, v nadaljevanju IRB). Kot odgovor na ta poročila bo EBA v prihodnjih letih pripravila posodobitve smernic (EBA, 2015). Še najbolj problematična se zdi kalibracija na dolgoročne stopnje stečajev, ki je potrebna za razvoj IRB modelov. Banka za mednarodne poravnave (angl. *Bank for International Settlements*, v nadaljevanju BIS; 2006) v članku ugotavlja neuspešnost validacijskih testov kalibracije na trgih v razvoju in poziva na nadaljnje raziskave na tem področju. Zelo težavno je torej predvsem področje validacije modelov. BIS je sicer že v letu 2005 objavila zelo pomemben dokument (BIS, 2005), ki služi kot priporočilo za izbor trenutno najboljših validacijskih testov.

Jedro naloge predstavlja celoten cikel razvoja bonitetnega sistema, od začetnih priprav podatkov do izbora modela za napovedovanje neplačil podjetij, implementacije in validacije eno leto po njegovi implementaciji. Cel cikel je zajet tudi na diagramu (v Prilogi A), ki ima oštevilčenih 8 glavnih korakov. Z rdečo barvo so označene metode in pristopi, ki smo jih obdelali v empiričnem delu naloge. Posebna skrb je namenjena pripravi in obdelavi podatkov, ki vključuje tudi pripravo več vzorcev: vzorec za oceno modela, vzorec za testiranje rezultatov, vzorec izven časa, vzorec za validacijo ter vzorci za sprotno spremljavo uspešnosti modela. Temeljimo na podatkih o neplačilih (stečaji, prisilne poravnave in 90-dnevne zamude) manjše slovenske banke. Uporabimo več kot 130 računovodskih kazalnikov, ki jih pravilno oblikujemo, nato pa transformiramo s substitucijo logaritmiranih obetov v 8 razredih in jih zgladimo z uporabo Hodrick-Prescott filtra.

Za oceno verjetnosti stečajev smo poleg logistične regresije uporabili modela naključnih gozdov in nevronskih mrež. Kot metodo za izbor spremenljivk, ki so vstopale v logistično regresijo smo uporabili metodo naključnih dreves, ki se je zelo dobro izkazala že v članku Ortl in Gorenjec (2015). S primerjavo različnih metod izbora spremenljivk sta se na slovenskih podatkih ukvarjala tudi že Brezigar-Masten in Masten (2012).

Za primerjavo z novejšimi metodami strojnega učenja je ključen članek avtorjev Lessman, Seow, Baesens, & Thomas (2015), v katerem avtorji primerjajo veliko število metod strojnega učenja. Najboljše rezultate dobijo z metodami HCES-Bag, slučajnimi drevesi, nevronskimi mrežami in z logistično regresijo. Več metod primerja tudi Falavigna (2006). Njeni rezultati so mešani, vendar nevronske mreže kažejo dobre rezultate. Mešane rezultate dobi tudi Kraus (2014). Tudi mi smo primerjali rezultate metod strojnega učenja z rezultati pridobljenimi z uporabo logistične regresije. Ugotovili smo, da nam algoritmi strojnega učenja dajo primerljive (metoda naključnih dreves) ali pa celo boljše (nevronske mreže) rezultate od logistične regresije kljub dejstvu, da je naša podatkovna baza razmeroma majhna. Omenjeni algoritmi imajo svoje prednosti in slabosti. Kot prednosti lahko izpostavimo njihovo uporabnost kot dopolnilo logistični regresiji za dosego boljših rezultatov (npr. pri izboru spremenljivk). Glavna slabost je težja interpretacija rezultatov in težavna kalibracija parametrov.

V nalogi smo poleg ocene modela prikazali tudi nadaljni razvoj bonitetnega sistema. V nadaljevanju smo temeljili na modelu logistične regresije s katerim smo ocenili verjetnosti stečajev. Napovedane verjetnosti stečajev smo nato kalibrirali na pričakovano stopnjo neplačila, s pomočjo algoritma diferencialne evolucije pa le-te uvrstili v osem bonitetnih razredov. Eno leto po hipotetični implementaciji modela izpeljemo kvantitativno validacijo diskriminacije in kalibracije modela. Stabilnost in razlikovanje dolžnikov modela smo preverili z merami kot so AUC, AR, KS, KL, Kendall tau in Brier score. Napovedane stopnje neplačila smo primerjali z realiziranimi stopnjami neplačil s pomočjo kalibracijskih testov kot so binomski test, Hosmer-Lemeshow test, Spiegelhalter test in enofaktorski binomski test. Rezultati kažejo, da smo bili pri implementaciji modela uspešni, saj ima model stabilno diskriminatorno moč, obenem pa je bila kalibracija uspešna.

Menimo, da naloga predstavlja obsežen pregled obravnavane tematike, izboljšave pa vidimo predvsem v nadgradnji validacijskih testov v smeri TTC (angl. *through the cycle*) bonitetnega sistema, saj je naš sistem temeljil na PIT (angl. *point in time*) pristopu. Dodatno skrb bi bilo potrebno nameniti kalibraciji pričakovane stopnje stečajev, bodisi z vključitvijo makroekonomskih spremenljivk ali pa z izdelavo ločenega makroekonomskega modela napovedovanja agregatne mere stopnje neplačil v državi. Skozi nalogo smo na vsakem koraku predstavili možne alternative (npr. različne transformacije podatkov, različne načine izbora spremenljivk ipd.). Le-te bi veljalo testirati in poiskati najboljše med njimi. Pred implementacijo je potrebno še metodološko opredeliti delovanje modela in njegovo interakcijo z analitskimi ocenami.