MODELLING PROBABILITY OF DEFAULT IN THE NORDICS

CARL GÖRANSSON AND EGIL NORDGREN

Master’s thesis
2018:E27

LUND UNIVERSITY

Faculty of Engineering
Centre for Mathematical Sciences
Mathematical Statistics
Modelling Probability of Default in the Nordics

Carl Göransson & Egil Nordgren

Division of Mathematical Statistics
Faculty of Engineering at Lund University

June 2018
Abstract

Credit risk is one of the greatest risks facing financial institutions, and it is therefore very important that models with good predictive power are used in order to better capture this risk. This thesis proposes logistic regression models for modelling risk-drivers of the probability of default in a financial institution active in the Nordics. The thesis focuses on handling and analysing large amounts of data in an attempt to find significant risk-drivers for each of the countries in the Nordics, as well as for a pan-Nordic model. The conclusion is that a pan-Nordic model performs very poorly, while some individual countries can be modelled quite well.

Keywords: Credit risk, Probability of default, Logistic regression, Risk-drivers.
Acknowledgements

We would like to thank our supervisor, Nader Tajvidi, for his guidance and feedback during the process of writing this thesis.

Lund, June 2018
Carl Göransson
Egil Nordgren
## Contents

### 1 Introduction 1
  1.1 Background ......................................................... 1
  1.2 Problem formulation ............................................ 2
  1.3 Scope and limitations ........................................... 2
  1.4 Chapter outline .................................................. 3

### 2 Risk definitions and regulations 4
  2.1 The Basel Capital Accords ..................................... 4
  2.2 The three pillars of Basel II .................................. 5
  2.3 Types of risks in the banking sector ......................... 5
  2.4 Credit risk ....................................................... 6
    2.4.1 Probability of default .................................... 6
    2.4.2 Loss given Default ......................................... 7
    2.4.3 Exposure at Default ....................................... 7
    2.4.4 Expected Loss .............................................. 7
    2.4.5 Internal Ratings Based Approach .......................... 7

### 3 Modelling probability of default 9
  3.1 Logistic regression ............................................. 10
    3.1.1 Parameter estimation ...................................... 12
  3.2 Weight of Evidence (WoE) and Information Value (IV) ...... 14
  3.3 Goodness-of-fit, and Pseudo-$R^2$ ........................... 15
    3.3.1 The $R^2_L$ measure ...................................... 16
    3.3.2 The $R^2_{McF}$ measure .................................. 16
    3.3.3 The $R^2_{CS}$ measure .................................... 17
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3.4</td>
<td>Summary</td>
<td>17</td>
</tr>
<tr>
<td>3.4</td>
<td>ROC curve</td>
<td>17</td>
</tr>
<tr>
<td>3.5</td>
<td>Kolmogorov-Smirnov test statistic</td>
<td>19</td>
</tr>
<tr>
<td>3.6</td>
<td>AIC</td>
<td>19</td>
</tr>
<tr>
<td>3.7</td>
<td>Likelihood ratio test</td>
<td>20</td>
</tr>
<tr>
<td>3.8</td>
<td>Wald test</td>
<td>21</td>
</tr>
<tr>
<td>4</td>
<td>Methods</td>
<td>22</td>
</tr>
<tr>
<td>4.1</td>
<td>PD modelling</td>
<td>22</td>
</tr>
<tr>
<td>4.2</td>
<td>Variable selection</td>
<td>22</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Manual variable selection</td>
<td>23</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Manually created variables</td>
<td>23</td>
</tr>
<tr>
<td>4.2.3</td>
<td>Statistical variable selection</td>
<td>23</td>
</tr>
<tr>
<td>4.2.4</td>
<td>Variable transformations</td>
<td>24</td>
</tr>
<tr>
<td>4.3</td>
<td>Model fitting</td>
<td>24</td>
</tr>
<tr>
<td>4.4</td>
<td>Model validation</td>
<td>25</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Out-of-sample</td>
<td>25</td>
</tr>
<tr>
<td>4.4.2</td>
<td>Out-of-time</td>
<td>25</td>
</tr>
<tr>
<td>4.5</td>
<td>Model selection</td>
<td>25</td>
</tr>
<tr>
<td>4.6</td>
<td>Outlier analysis for selected model</td>
<td>26</td>
</tr>
<tr>
<td>5</td>
<td>Results</td>
<td>27</td>
</tr>
<tr>
<td>5.1</td>
<td>Brief explanation of all significant risk-drivers</td>
<td>27</td>
</tr>
<tr>
<td>5.2</td>
<td>Results from the Nordic countries</td>
<td>29</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Sweden</td>
<td>29</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Denmark</td>
<td>30</td>
</tr>
<tr>
<td>5.2.3</td>
<td>Finland</td>
<td>30</td>
</tr>
<tr>
<td>5.2.4</td>
<td>Norway</td>
<td>31</td>
</tr>
<tr>
<td>5.2.5</td>
<td>The Nordics</td>
<td>34</td>
</tr>
<tr>
<td>5.3</td>
<td>Results for models fitted on outlier-adjusted data</td>
<td>36</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Sweden</td>
<td>36</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Denmark</td>
<td>36</td>
</tr>
<tr>
<td>5.3.3</td>
<td>Finland</td>
<td>38</td>
</tr>
</tbody>
</table>
6 Analysis and Discussion

6.1 Analysis of each country’s respective models
   6.1.1 Sweden ........................................... 50
   6.1.2 Denmark ........................................ 51
   6.1.3 Finland ......................................... 51
   6.1.4 Norway .......................................... 52
   6.1.5 The Nordics ................................... 52

6.2 Comparison between models from the same country
   6.2.1 Denmark ........................................ 53
   6.2.2 Finland ......................................... 54
   6.2.3 Norway .......................................... 54
   6.2.4 Nordics ......................................... 55

6.3 Comparison between results from the different countries ........... 55
   6.4 Discussion about validity of the models .......................... 56

7 Conclusions ............................................ 57

Bibliography ............................................. 59
Notation and Abbreviations

Below is a list of the different abbreviations used throughout this thesis.

- **PD**: Probability of default
- **LGD**: Loss given default
- **EAD**: Exposure at default
- **EL**: Expected loss
- **IRB**: Internal ratings based approach
- **SSE**: Squared sum of errors
- **ML**: Maximum likelihood
- **MLE**: Maximum likelihood estimator
- **WoE**: Weight of evidence
- **IV**: Information value
- **ROC**: Receiver operating characteristic curve
- **FPR**: False positive rate
- **TPR**: True positive rate
- **AR**: Accuracy ratio
- **KS**: Kolmogorov-Smirnov statistic
<table>
<thead>
<tr>
<th></th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>Akaike’s information criterion</td>
</tr>
<tr>
<td>OOS</td>
<td>Out-of-sample</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

In this chapter some background to the concept of credit risk and its regulatory landscape is given. Furthermore, the goal and limitations of this thesis are presented. Lastly, an outline of the contents of the different chapters is shown.

1.1 Background

Credit risk has traditionally been the greatest risk facing financial institutions and has therefore required the most regulatory capital - that is, the minimum capital the institution must hold to be able to properly cover large losses in the event of e.g. a financial crisis, see Hull [2015]. It is therefore crucial to financial institutions that they calculate these capital requirements correctly.

The capital requirements are regulated by the Basel Capital Accords, which are regulatory standards for financial institutions. Albeit being only regulatory standards, they have been incorporated into European Union law, meaning all banks and financial institutions within the European Union must comply with certain demands. The largest of these accords, the second Basel Capital Accord, *Basel II*, specifies among other things the minimum capital requirements and is reviewed further in Chapter 2.

One important aspect of Basel II is that it allows, if cleared by supervisory enti-
ties, sufficiently large financial institutions to use internal models to model credit risk, as opposed to using standardized external models. This approach is called the Internal Ratings-Based Approach, which is examined in Chapter 2. Furthermore, as holding too much capital affects the profit, it is of great importance that regulatory capital is optimized - not only to better handle future financial crises, but also to best utilize the free capital of the financial institution.

The third Basel Capital Accord, Basel III, which was developed in response to the financial crisis of 2007-2008, has made the regulatory landscape even stricter. This stricter framework, in combination with the lessons learned from the crisis, makes accurate credit risk models even more important for banks and financial institutions.

1.2 Problem formulation

The aim of this thesis is to suggest statistical models for estimating the probability of default based on a number of explanatory variables, or covariates, from customer data. Several models will be analysed, for each of the countries in the Nordics as well one for all of the Nordics.

The models should be viable from a statistical perspective, as well as from a business perspective, which is derived from common practices in the industry.

1.3 Scope and limitations

The scope of this thesis will be on fitting a logistic regression model for each of the four countries in the Nordics and one for the whole Nordics - mainly focusing on finding the variables that are risk-drivers for credit risk. Other models were considered but due to a large amount of variables and data, the variable selection in itself becomes extensive, meaning the thesis focuses on just the logistic regression model.
1.4 Chapter outline

The chapter outline of the thesis is as follows:

Chapter 2 Credit risk - This chapter introduces the most important aspects of credit risk, as well as the current regulatory landscape.

Chapter 3 Mathematical theory - This chapter reviews logistic regression theory, as well as statistical theory used for testing and evaluating the different models.

Chapter 4 Methods - This chapter goes through the approach to modelling the logistic regression, looking at the variable selection and model validation methods used.

Chapter 5 Results - In this chapter the results achieved when using the methods discussed in Chapter 4 are presented.

Chapter 6 Analysis and discussion - This chapter discusses and analyses the results from Chapter 5.

Chapter 7 Conclusions - The thesis is concluded with the main points drawn from the discussion in Chapter 6 together with ideas for future studies.
Chapter 2

Risk definitions and regulations

In this chapter, the main types of risks for banks are presented - with credit risk being the most important. The regulatory landscape, mainly focusing on the Basel Capital Accords, as well as definitions for default are presented.

2.1 The Basel Capital Accords

The latest version of the Basel Capital Accord, *Basel III*, first published in December 2010, consists of reforms and regulations that aim towards financial stability and a banking sector that is resilient against financial downturn; see Banking Supervision 2011 for more details. The third version is a development of Basel II, published in June 2004, which in turn is a development of the first Basel Accord, published in 1998. As stated in Banking Supervision 2004, Basel II consists of three main pillars: *Minimum Capital Requirement*, *Supervisory Review*, and *Market Discipline*, which will be presented in more detail below as this is such an important framework. In Banking Supervision 2017a it is explained how Basel III mainly strengthens the capital requirements as well as addresses a number of shortcomings from Basel II.


2.2 The three pillars of Basel II

**Minimum Capital Requirement** Hull [2015] explains how this pillar deals with maintenance of regulatory capital that is required to safeguard against the three major risks facing banks - credit risk, market risk, and operational risk. It sets out threshold levels for capital that needs to be held for the different risk classes. These risks will be reviewed further in the next section.

**Supervisory Review** As discussed in Hull [2015], this pillar covers the supervisory review process. It treats both quantitative and qualitative aspects of risk management by the banks. Supervisors are required to ensure that banks have processes in place for ensuring that capital requirements are met. This is to be able to handle fluctuations in capital requirements and to quickly be able to raise capital should the need arise. Emphasis in this pillar is to intervene early and to encourage the use and development of better risk management techniques.

**Market Discipline** This pillar requires banks to disclose more information about capital allocation and which risks they take. Hull [2015] argues that this is to incite banks to make sounder risk management decisions as shareholders will have more information about these decisions.

2.3 Types of risks in the banking sector

Three main risks are examined in the first pillar of Basel II, namely operational risk, market risk, and credit risk - with credit risk being the most crucial for banks, according to Hull [2015].

**Operational risk** This is the risk of losses where the bank’s internal procedures do not work as expected, or the risk of losses from external events such as earthquakes and fires. It is quite loosely defined, but it can be seen as risks that are neither market risk nor credit risk.

**Market risk** Market risk is the risk primarily generated by the bank’s trading operations. Generally it is the risk that financial instruments in the bank’s
trading book will decline in value and incur losses.

Credit risk This is defined as the risk that counterparties in loan and derivatives transactions will default - that is, not be able to pay back their loans. This has traditionally been the largest risk facing banks and also the one that requires banks to hold the most regulatory capital.

2.4 Credit risk

As credit risk is such an important risk for banks, and is also the underlying risk that this thesis is modelling, some more concepts regarding this are presented below. In the scope of this thesis, the probability of default is the most relevant factor, while the others are presented for completeness.

2.4.1 Probability of default

Default risk is defined as the probability that a default event occurs, also called the probability of default (PD). This is a probability measure, with 0 being no risk of a default occurring, and 1 being a guaranteed default. Financial institutions have some freedom in how they define default events. Generally, being more than three months late on payments counts as a default event. The Basel Accords, see Banking Supervision 2004, define a default event as the following:

The Basel definition of default
A default is considered to have occurred with regard to a particular obligor when either or both of the two following events have taken place:

- The bank considers that the obligor is unlikely to pay its credit obligations to the banking group in full, without recourse by the bank to actions such as realising security (if held).

- The obligor is past due more than 90 days on any material credit obligation to the banking group. Overdrafts will be considered as being past due once the customer has breached an advised limit or been advised of a limit smaller than current outstandings.
2.4.2 Loss given Default

Loss given default (LGD) is defined as the share of the exposure that is lost when a counterparty defaults. The size of the LGD is affected by several factors. For example, the industry of the obligor can greatly affect the recovery rate. According to Schuermann [2004], there is usually a significantly higher recovery rate for industries with tangible assets than in the service sector.

2.4.3 Exposure at Default

Exposure at default (EAD) is defined as the amount outstanding that can be lost in the event of a default. For example, this can be the nominal value of a housing loan, or the credit card debt a customer has incurred during a month.

2.4.4 Expected Loss

The expected loss is the amount a bank can expect to lose given an actual default. It is defined as:

\[ EL = EAD \times LGD \times PD \]

EL in turn is the basis for the capital requirements that the bank needs to hold to safeguard against credit losses.

2.4.5 Internal Ratings Based Approach

As stated before, banks and financial institutions of sufficient size are, if given supervisory approval, allowed to calculate their own measures of the EL, EAD, LGD, and PD, as long as they meet certain demands. This is called the Internal Ratings Based (IRB) Approach, and effectively allows institutions to formulate their own credit risk models. This is opposed to using the other approach called the Standardized Approach, which uses external approximations of the risk measures presented above.

When given clearance to use the IRB Approach, the institutions can choose between using two approaches. The first one, called Foundation IRB, lets them
provide their own internal estimations of their PD, while all other relevant risk measures come from the supervisor. The second one, Advanced IRB, lets them provide their own estimations of all relevant risk measures. See Banking Supervision [2017b] for more details.

For this thesis, the institution that supplied the data used the IRB Approach, which allowed for the logistic regression to be made.
Chapter 3

Modelling probability of default

This chapter contains statistical theory and concepts regarding mathematical models and concepts used. As mentioned above, it is important for financial institutions to estimate the credit risk that they are exposed to. The more exposed they are, the more capital they are required to hold, which in turn affects the profits and losses. One major component of estimating this risk is calculating the probabilities that the borrowers of the institutions will be unable to repay their debts, that is - calculating the PD.

One common approach to PD modelling is to use historical data of earlier defaults in order to find common traits among the borrowers that have defaulted previously, see Bandyopadhyay [2006] for example. From this the institutions are then able to estimate a current borrower’s likelihood of defaulting. The historical data consists of a binary response variable, ”did the borrower default, yes or no?” and customer specific variables that are used as explanatory variables. Based on this setup it is natural to consider a model where the response variable is categorical. In this section the main model for modelling this is presented - the logistic regression model.
3.1 Logistic regression

In the following sections, the mathematical specification for the logistic model is presented, together with some related test statistics.

The logistic function

The logistic function, \( \sigma(t) \), is a function that takes in any real value \( t, t \in \mathbb{R} \) and outputs a value \( \in [0,1] \). Because of this, it can be interpreted as a probability:

\[
\sigma(t) = \frac{1}{1 + e^{-t}}. \tag{3.1}
\]

Now, it is possible to assume a linear relationship between \( t \) and some explanatory variables, or covariates, \( [x_1, x_2, \ldots] \), e.g.:

\[
t = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots. \tag{3.2}
\]

Using this it is possible to rewrite the logistic function, i.e. the probability of an event, as

\[
\sigma(t) = p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots)}}. \tag{3.3}
\]

Odds

Odds is a statistical measure that describes how likely it is that an event occurs. The higher the odds, the higher the probability that the event occurs. Using the probability of an event \( p(x) \) odds are defined in the following way.

\[
\text{odds}(x) = \frac{p(x)}{1 - p(x)} \tag{3.4}
\]

Using the expression for the probability \( p(x) \) in equation Equation (3.3) above, one can derive an expression for the odds in the following way:

\[
\frac{p(x)}{1 - p(x)} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots)}} = \frac{1}{e^{-((\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots)}}} = e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots)}. \tag{3.5}
\]
The logit-function, log-odds

The logit function, \( g \), is defined as taking the natural logarithm of the odds.

\[
g(p(x)) = \ln\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... 
\] (3.6)

From this one can see that the logit function - taking the natural logarithm of the odds, so called log-odds, is equivalent to the linear regression for \( t \) defined in equation (3.2) above. In similar fashion to James et al. [2013] a more precise definition of the logistic regression model can now be presented.

Formal definition

Using the concepts discussed above, it is now possible to give a more formal definition of the logistic regression model.

The goal is to model the binary outcome \( Y_i \) of an event \( i \) using some explanatory variables \( x_{1,i}, ..., x_{m,i} \). It is assumed that the outcomes \( Y_i \) are Bernoulli-distributed, which means that each outcome depends on an unobservable probability \( p_i \). More formally this can be written as

\[
\mathbb{E}[Y_i | x_{1,i}, ..., x_{m,i}] = p_i. \quad (3.7)
\]

Now, applying the expression from equation (3.3) to this probability \( p_i \) (i.e. assuming that it is possible to model the probability with the logistic function) the following linear prediction function for the probability is obtained.

\[
f(p_i) = logit(p_i) = logit(\mathbb{E}[Y_i | x_{1,i}, ..., x_{m,i}]) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... \quad (3.8)
\]

This is usually written more compactly as

\[
f(p_i) = logit(p_i) = \boldsymbol{\beta} \cdot \mathbf{x}_i, \quad (3.9)
\]

where \( \boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, ...] \) and \( \mathbf{x}_i = [x_{0,i}, x_{1,i}, x_{2,i}, ...] \). The value of \( x_0 \) is 1 for all observations. It is added as a covariate so that it is possible to write the regression
using dot multiplication between the vectors containing the parameters and the explanatory variables.

Now it is possible to find estimates for the set of parameters, \( \boldsymbol{\beta} \), using iterative methods.

### 3.1.1 Parameter estimation

When estimating the parameters for the logistic regression model, one assumes a model of the following form

\[
    f(p_i) + \epsilon_i = \boldsymbol{\beta} \cdot \mathbf{x} + \epsilon_i, \tag{3.10}
\]

where \( \epsilon_i \) is an error-term that has a logistic distribution, with \( \mathbb{E}[\epsilon_i] = 0 \) and \( \text{Var}(\epsilon_i) = p_i(1 - p_i) \).

The logistic regression model is a model that is non-linear in its parameters, and with heteroscedastic errors, which makes it trickier to estimate its parameters compared with a linear regression with homoscedastic error terms.

#### Iteratively reweighted least squares

The iteratively reweighted least squares algorithm aims to solve an optimization problem of the following form

\[
    x = \operatorname{arg\,min}_x \sum_{i=1}^{n} |Y_i - f_i(x)|^p, \tag{3.11}
\]

where usually the goal is to minimize the squared sum of errors (SSE), by using an iterative method where a weighted least squares problem is solved in each step.

The weighted least squares problem is of the following form

\[
    x^{(t+1)} = \operatorname{arg\,min}_{x^{(t)}} \sum_{i=1}^{n} w_i(x^{(t)})|Y_i - f_i(x^{(t)})|^p, \tag{3.12}
\]
where \( w_i \) are weights assigned to each observation. Equation (3.12) can be solved (minimised) using a numerical minimisation algorithm, such as the Gauss-Newton algorithm for example. The minimisation algorithm is run iteratively, using the previous iterations solution as a starting point.

In Rawlings et al. [1998] it is shown that an iteratively reweighted least squares estimate for the parameters in the logistic regression can be obtained by minimising

\[
\sum_{i=1}^{n} \frac{1}{\hat{p}_i(1 - \hat{p}_i)} \left( Y_i - \frac{\exp(\beta \cdot x_i)}{1 + \exp(\beta \cdot x_i)} \right)^2,
\]

where \( \hat{p}_i \) is the calculated value of \( p_i \) given the current set of parameters \( \beta \). If one compares this expression with (3.12) it can be seen that the weights \( w_i \) assigned are the same as the inverse of the calculated variance. Furthermore, the expression \( \frac{\exp(\beta \cdot x_i)}{1 + \exp(\beta \cdot x_i)} \) is the same as \( \hat{p}_i \), since using the expression from Equation (3.6) one can show

\[
\logit(p_i) = \log \left( \frac{p_i}{1 - p_i} \right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \iff \frac{p_i}{1 - p_i} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots} \iff (1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots}) p_i = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots} \iff p_i = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots}} = e^{\beta \cdot x_i}.
\]

**Maximum likelihood estimation**

An alternative to using the iterative reweighted least squares method discussed above is the maximum likelihood estimator (MLE) for the parameters. ML is, as the name suggests, focused on finding the parameters that maximizes the likelihood of having the observations occur. More formally this can be written as

\[
\hat{\Theta} = \arg\max_{\Theta} \mathcal{L}(\Theta),
\]
where $\mathcal{L}(\Theta)$ is the likelihood function. When variables are independent, the likelihood function simply becomes

$$\mathcal{L}(\Theta) = \prod_{i=1}^{n} f(x_i|\Theta).$$  

(3.16)

Depending on the distribution, (3.16) might be difficult to differentiate in order to find the maxima. It is therefore often times convenient to consider the log-likelihood function which is the natural logarithm of the likelihood function. Since taking the logarithm is a non-decreasing transformation, the set of parameters that maximizes the log-likelihood also maximizes the likelihood. The log-likelihood is defined as

$$\ell(\Theta) = \sum_{i=1}^{n} \log (f(x_i|\Theta)).$$  

(3.17)

In Agresti 2002, p.192-194, the MLE for the logistic regression model is presented. It is shown that the maximum likelihood estimator of the parameters, $\hat{\beta}_{ML}$, is the set of parameters that maximize

$$\ell(\beta) = \sum_{i=1}^{n} Y_i \beta \cdot x_i - \sum_{i=1}^{n} \log(1 + e^{\beta x_i}).$$  

(3.18)

Furthermore, Agresti presents iterative methods of obtaining $\hat{\beta}_{ML}$, as well as showing that the estimator is approximately normally distributed with mean $E[\hat{\beta}_{ML}] = \beta$ and covariance $\text{Cov}(\hat{\beta}_{ML}) = [x \text{diag}(\hat{p}_i(1 - \hat{p}_i))x^T]^{-1}$.

### 3.2 Weight of Evidence (WoE) and Information Value (IV)

Weight of evidence is a measure that quantifies to what extent a variable or a certain attribute supports or undermines a hypothesis, see Computing the Weight of Evidence and Information Value 2018. It can be used as a tool during screening of possible covariates to limit the set of variables to only those that have sufficiently high WoE, and hopefully also good predictive power. The WoE measure is defined
as

\[
\text{WoE}_{\text{variable}} = \log \left( \frac{p_{\text{non-event}}^{\text{variable}}}{p_{\text{event}}^{\text{variable}}} \right),
\]

where \(N_{\text{event}}^{\text{variable}}\) is the number of observations that possess the attribute and had an event. \(N_{\text{event}}^{\text{total}}\) is the total number of observations that had an event. \(N_{\text{non-event}}^{\text{variable}}\) and \(N_{\text{non-event}}^{\text{total}}\) are defined in an analogous fashion.

To deal with continuous variables it is common to group them into bins, where the attribute for an observation would be “had this variable in a certain range”.

Information value is a weighted sum of the WoE. It calculates the total WoE of all bins for a variable, and weighs them according to the difference in conditional probability of an attribute given an event, and the conditional probability of that variable given a non-event. More formally

\[
IV = \sum_{i=1}^{m} \left( \frac{N_{\text{event}}^{\text{variable}}}{N_{\text{event}}^{\text{total}}} - \frac{N_{\text{event}}^{\text{variable}}}{N_{\text{event}}^{\text{total}}} \right) \text{WoE}_i,
\]

where \(m\) is the number of bins (and variables). Generally speaking, the higher \(IV\), the greater the probability that the binned variable will be a good predictor.

### 3.3 Goodness-of-fit, and Pseudo-\(R^2\)

In ordinary linear regression, the coefficient of determination, or \(R^2\), is commonly used as a measure of goodness-of-fit. Rawlings et al. [1998] explains that it represents the proportion of variance in the dependent variable that can be explained by the covariates. Unfortunately, as explained in Allison [2014], there is not an analogous measure for the logistic regression model that is as unanimously adopted as \(R^2\). Instead, there exist several competing measures. A selection of those are
presented below.

3.3.1 The $R^2_L$ measure

The $R^2_L$ measure (often called the likelihood ratio) uses the deviance expressions

\[
D_{null} = -2 \log (L(M_{null})) \\
D_{model} = -2 \log (L(M_{model})) ,
\]

(3.21)

where $M_{null}$ is the model where only an intercept has been fitted, and $M_{model}$ is the model that is being evaluated. Using these two expressions, the $R^2_L$ is defined as

\[
R^2_L = \frac{D_{null} - D_{model}}{D_{null}} .
\]

(3.22)

As can be seen, this measure is quite similar to the normal $R^2$ in the sense that it relates the proportion of deviance explained when using the model compared to when it is not used. One drawback of the measure is that the value of the likelihood ratio does not necessarily change when the odds ratio changes.

3.3.2 The $R^2_{McF}$ measure

In McFadden [1975], the $R^2_{McF}$ measure is presented. Instead of utilising the deviance, as the measure above, it uses the log-likelihood of the model and null model. It is defined as

\[
R^2_{McF} = 1 - \frac{\ell(M_{model})}{\ell(M_{null})} .
\]

(3.23)

When examining this expression one sees that in the case of the fitted model performing no better than the null model, then the value of $R^2_{McF}$ will be close to zero, since the ratio between the log-likelihoods will be close to 1. In the opposite case, when the model is fitted perfectly the value of $R^2_{McF}$ will be close to 1. This because if the observations are modelled perfectly, then $P(Y_i) = 1$ for all observations that had $Y_i = 1$, and $P(Y_i = 0) = 1$ for all observations where $Y_i = 0$. Since $\log(1) = 0$ one see that $\ell(M_{model}) = 0$, which gives $R^2_{McF} = 1$. 

16
3.3.3 The $R^2_{CS}$ measure

Another related measure is the $R^2_{CS}$ measure, defined in Cox et al. 1989. It is defined as

$$R^2_{CS} = 1 - \left( \frac{L(M_{model})}{L(M_{null})} \right)^{2/n}, \quad (3.24)$$

where $n$ is the number of observations. However, as discussed in Allison 2014, there are some issues with this measurement as well. Due to its construction there exists a maximum upper bound of 0.75.

3.3.4 Summary

It is worth noting that none of the Pseudo-$R^2$ measures discussed above function in the same way as $R^2$ does in normal linear regression. The pseudo-$R^2$ measures do not represent the same proportional reduction in errors as $R^2$ for linear regression. Furthermore, since the errors are heteroscedastic in the logistic regression model the error variances are different. Therefore, there might be a different proportional reduction in variance for each observation.

3.4 ROC curve

A Receiver Operating Characteristic-Curve (ROC-curve) is a graphical tool to evaluate the diagnostic ability of a binary classification system, given certain discrimination thresholds - that is, the threshold at which a value between 0 and 1 is counted as either a 0 or a 1. The curve is generated by plotting the true positive rate (TPR), or the sensitivity, against the false positive rate (FPR), or 1 - specificity, at various thresholds, thus generating a curve. Essentially, the shape of the curve reflects the accuracy of the diagnostic test. Random guessing leads to a diagonal line from the bottom left corner to the top right, with TPR = FPR. A perfect ROC-curve is just a point at the top left corner - with the TPR rate being 1, and the FPR being 0.

Furthermore, the area under the curve (AUC) is used for evaluating the accuracy as well. As explained in Hanley et al. 1982, the AUC can be seen as a measure of
the probability of correctly identifying observations. A perfect curve has an AUC of 1, while a random curve (the worst case) has an AUC of 0.5. A good ROC should therefore have an AUC close to 1 as possible, meaning a curve that goes from the bottom left to the top right as close to the top left as possible. See Figure 3.1 for three different ROC-curves visualising a bad, an okay, and a perfect ROC-curve. As in the okay example, an AUC equal to about 0.8 can be interpreted to mean

![Figure 3.1: Example of ROC-curve with three cases](image)

that a randomly selected individual from the positive group has a test value larger than that for a randomly selected individual from the negative group 80 percent of the time.

This statistic is related to another statistic called the Gini Coefficient or the Accuracy Ratio, which is defined as

\[
Gini = 2 \times \text{AUC} - 1, \tag{3.25}
\]
meaning it takes values between 0 (worst) and 1 (best) instead of between 0.5 and 1.

### 3.5 Kolmogorov-Smirnov test statistic

Similar to the ROC-curve and the Gini coefficient, the Kolmogorov-Smirnov (KS) statistic captures discriminatory power in a model by looking at the difference in distribution between good and bad observations (ones or zeros). As explained in Řezáč et al. [2011](#), the KS statistic essentially looks at the cumulative distributions of ones and zeros respectively, calculating the KS statistic as the maximum separation between these two cumulative distributions. It is defined as

\[
\text{KS} = \sup_{a \in [0, 1]} |F_{m,\text{good}}(a) - F_{n,\text{bad}}(a)|,
\]

where

\[
F_{m,\text{good}}(a) = \frac{1}{m} \sum_{i=1}^{m} I(s_i < a \land D_i = 1)
\]

\[
F_{n,\text{bad}}(a) = \frac{1}{n} \sum_{i=1}^{n} I(s_i < a \land D_i = 0)
\]

\[
D_i = \begin{cases} 
1, & \text{if observation } i \text{ is "good"} \\
0, & \text{if observation } i \text{ is "bad"}
\end{cases}
\]

It can also be seen as the maximum vertical distance from the ROC-curve to the diagonal line (worst case). The KS statistic goes from 0 to 1, with values of above around 0.2 being considered as acceptable.

### 3.6 AIC

Information criteria offer a way of comparing different models with each other. They are based on the idea of looking at the relative information lost when modelling a process using some model. Because of that it handles the trade-off between having a simple model, and a model that fits the data. In Akaike [1973](#), a criterion
is proposed that is defined in the following way.

\[
AIC = 2k - 2\log(\hat{L}),
\]  

(3.28)

where \(k\) is the number of parameters estimated in the model, and \(\hat{L}\) is the estimated likelihood for the model. Looking at (3.28) the lower the AIC value a model, the "better" the model can be considered when compared to other models with higher AIC. This because a model with lower AIC either has a higher likelihood, or fewer parameters, when compared with other models.

It is important to note that the AIC value does not say anything about the objective quality of a model, only the relative performance when compared with other models.

### 3.7 Likelihood ratio test

Another method for model comparison is the so called likelihood ratio test. Similar to the AIC value discussed above, this is also a test between two (or more) competing models. When the models to be compared are nested, that is one of the models contains a subset of the other models parameters, the test statistic receives some nice properties.

The idea of the test is to check whether adding more parameters to the so called null model results in a significantly higher likelihood or not. If it does not, then it is better to keep the model with fewer parameters. The test statistic, \(D\) is defined as

\[
D = -2\log\left(\frac{\mathcal{L}(M_0)}{\mathcal{L}(M_1)}\right) = 2(\ell(M_1) - \ell(M_0)),
\]

(3.29)

where \(M_1\) is the alternative model, the model with added parameters, and \(M_0\) is the null model. The statistic \(D\) will be approximately \(\chi^2\) distributed with degrees of freedom equal to the difference in parameters between the two models. Based on this distribution it is then possible to determine whether or not the increase in likelihood with the alternative model is significant.
3.8 Wald test

The Wald test, sometimes called Wald $\chi^2$ test, is a statistical test that can be used to perform hypothesis testing regarding parameter estimates. More formally, in the univariate case, the hypotheses tested are

$$H_0 : \beta = \beta_0.$$ 
$$H_1 : \beta \neq \beta_0. \quad (3.30)$$

If one lets $\beta_0 = 0$, then it is possible to use the Wald test to determine whether a parameter can be deemed insignificant or not (if the null hypothesis cannot be rejected then the value of the parameter might be zero, and therefore not have a significant impact statistically).

As explained in Engle [1984], the Wald test statistic is constructed using the asymptotic normality of the maximum likelihood estimates. The statistic is defined as

$$z^2 = \frac{(\hat{\beta} - \beta_0)^2}{\text{Var}(\hat{\beta})}. \quad (3.31)$$

The value of $z^2$ will be $\chi^2$ distributed with 1 degree of freedom. Because of this the P-value is the right tail probability of the $\chi^2$ distribution above the observed value of $z^2$. 

21
Chapter 4

Methods

This chapter reviews the methods used to derive the results presented in Chapter 5. Starting with the logistic regression modelling, it also looks at variable selection and reviews the methods used for model validation.

4.1 PD modelling

The modelling was performed using the statistical software SAS. The process began with cleaning the data, selecting the most important variables, running the logistic regression, and finally validating the various models. The data examined is from a time period of 5 years, with defaults being examined one year forward - meaning for example the data for year 5 is used when looking at defaults occurring during year 6. I.e. the response variable for the logistic regression is whether or not the customer has defaulted at any time in the coming year.

4.2 Variable selection

Due to confidentiality reasons, the actual number of observations and variables cannot be disclosed in this thesis. Although, what can be said is that the data contains a substantial amount of variables, meaning it would be computationally impossible to perform a logistic regression with all of them included. More specifically, the selection step would become very time-demanding due to the amount
of possible combinations. Because of this, a lot of the variables had to be removed manually before running the analysis. The method used for this is presented below.

### 4.2.1 Manual variable selection

Due to the data being real-life customer data, some peculiarities were to be expected, due to for example manual errors or outdated information. Redundant and missing variables were therefore removed from the data set, as well as variables having a substantial amount of missing values. Variables that were known to be irrelevant, but still contained information, were also removed after discussions with senior staff.

Furthermore, the decision to remove customers that had not been active for the past two years was made. The rationale behind this decision was that the purpose of the analysis is to find the factors that drive credit risk, and logically customers that are not active should not hold any risk. Including their data in the analysis could therefore skew the results, and possibly affect the found risk-drivers.

### 4.2.2 Manually created variables

Some custom variables not present in the original data were created, mostly aggregate variables for variables that represent product usage behaviour that seemed logical to be in the same category. Furthermore, some variables were transformed manually so that they were consistent for the different years in the data set.

### 4.2.3 Statistical variable selection

After removing all the variables that could be cleaned out manually, a statistical analysis of the variables was performed in SAS to further decrease the size of the data set. As stated in [Developing Credit Scorecards Using Credit Scoring for SAS Enterprise Miner 12.1](2012), the approach used places variables in bins and then uses decision trees to sort out, based on their Information Value (3.20) and Gini coefficient (3.25). The cut-off values were chosen such that a reasonable amount of variables were selected, and that those variables most likely would have good
predictive power. Common practice in the industry is to consider variables with an IV of over 0.1 as "medium" predictive power, and above 0.3 as "good" predictive power, see Siddiqi 2006 for example. Doing this results in some ten variables, which can be properly handled by the logistic regression.

4.2.4 Variable transformations

After the variable selection step described above, certain transformations were applied to the covariates. Depending on the nature of the variable, different transformations were used. Two examples of transformations tried are taking the logarithm of the variable, and taking the square root of the variable.

As described in Rawlings et al. 1998, Chapter 12, variable transformations can sometimes be useful when fitting models. If the magnitude of the values of the variables is too large it might lead to ill-conditioned matrices which in turn affect the quality of parameter estimates. Transformations can also be used to stabilise the variance of covariates.

4.3 Model fitting

Using the variables selected above a logistic regression model was fitted to the data. In the fitting process the maximum likelihood estimation method discussed in Equation (3.18) was used. Furthermore, in order to find a good combination of explanatory variables a stepwise approach was used. The variables were added one at a time and a check was performed to see if the AIC value increased significantly when a variable was added to the model. If it was not significant, then that variable was dropped from the model, and another was added. This stepwise selection ran iteratively until there were no longer any significant increase in AIC by including another variable.
4.4 Model validation

After fitting the models, they are to be validated to ensure that the risk-drivers that are found are indeed relevant and statistically significant. Additionally the validation is done to ensure that the model has not been overfitted to the training data. The goal is to find drivers of credit risk that are both consistent and significant over time, and therefore they should not only be significant for a small time frame. Some validation methods are presented below:

4.4.1 Out-of-sample

Out-of-sample (OOS) validation is done by using a subset of the available data to fit the model, and then testing the performance of that model on a validation set. This is done to verify that the model has captured the underlying behaviour of the data, and that it is not over-fitted to the modelling data. For the modelling purposes of this report, the division was 80% modelling data, and 20 % validation data, where the selection of the validation data was done randomly.

4.4.2 Out-of-time

Out-of-time (OOT) validation is done by modelling on all data, and then validating on a time period outside of the modelling data. According to industry standards, this is preferred over OOS when data is scarce, which is not the case here. Therefore, the OOS validation method is used instead of OOT.

4.5 Model selection

To compare the performance of the fitted models that were validated some of the measures discussed in Chapter 3 were used. Essentially the model with the best predictive power was determined by the AUC, $R^2_{\text{Mcf}}$ and AR - values on the validation data set.
4.6 Outlier analysis for selected model

After a model had been chosen, the observation data for the variables in that model were analysed to see if there were any significant outliers in the data set that could skew the regression. If so, the data was trimmed at the upper and/or lower quantiles, depending on where the outliers were in the data. The quantile level at which the data was trimmed was determined on a per-variable basis, as the distributions for the variables were quite different at times. The reason for considering this after the analysis is that ideally, the empirical odds curve of the model should be a straight trend, either upwards or downwards. If the empirical odds curve exhibits for example a parabola-shape then outliers might be the cause of this, which indicates that the model might be improved if these are removed.

Instead of doing the outlier-adjustment based on the empirical odds curve, this sort of analysis could have been done to all variables, before any other analysis of the variables. However, as mentioned above, the sheer amount of variables present in the data set lead to this process being left out in favour of other analyses. In theory it is possible to automatically trim the data set with an arbitrarily chosen quantile before any analysis. However, this can be risky to do, since there is a possibility that observations with outliers are those that exhibit the characteristic modelled. If those observations are removed, then the quality of the overall regression might be severely impacted.
Chapter 5

Results

In this chapter the results obtained from applying the methods and frameworks discussed above on the data are presented.

5.1 Brief explanation of all significant risk-drivers

After fitting logistic regression models to the data, a number of different risk-drivers were found. Due to confidentiality reasons, the risk-drivers are grouped on a less granular level than what was used in the actual analysis. Furthermore, by presenting them in an aggregated fashion, it is simpler to compare results among different models. This is because all models might contain variables that are unique, but of the same type as variables in other models. A brief description of the variable groupings used is presented below.

**Past due behaviour** The variables in this grouping represent the customer’s behaviour regarding late payments. It could for example be the amount of times in the last year the customer was past due on a payment.

**Past due amount** An example of a variable in this grouping could be the maximum amount the customer was past due in the last 3 months.

**Delinquency history** This represents the information regarding the customer’s delinquency history. The more recent the customer has been delinquent, the more the variables in this grouping are affected.
Demand-chain Depending on how much is past due and on how long the customer has been delinquent there are different stages in the demand-chain the customer can be at. This is represented in this variable grouping. One example of a possible variable in this group is how many times the customer has received a reminder letter regarding their late bills.

External information This grouping consists of data from external companies. It could for example be an external credit score of the customer.

Suspension Variables that represent if the customer has been suspended from using the product are found in this grouping. An example of such a variable could be how many times in the past three years the customer has been suspended.

Balance The variables in this grouping represent the customer’s balance, i.e. how much debt the customer has. An example of a variable in this group could be the total amount of debt a customer had the past 6 months.

Product usage This grouping consists of variables representing the amount of usage of the underlying product. One example of such a variable could be how many times the product has been used the last month.

Product usage behaviour In this variable group are variables that reflect the nature of the usage of the product. An example of a variable could be how many times in the last year the product was used in a certain way.

Payment behaviour These variables represent the customer’s behaviour in paying off their debt. An example of a variable in this group could be the number of payments the customer has made in the past year.

Customer information In this grouping there are variables that contain background information about the customer. One such information could be how long the customer has been a client of the financial institution.
5.2 Results from the Nordic countries

With the variable categories explained, the results from each model is presented below.

5.2.1 Sweden

First, the Swedish model. The significant variable groups and their weights can be seen in Figure 5.1. The weights correspond to the percentage of the Wald $\chi^2$ score of the model, excluding the intercept parameter, that each variable grouping amounts to. As explained in the theory, a statistical measure that is commonly used to gauge logistic regression is the Accuracy Ratio and the Area under the ROC-curve (AUC), which can be seen in Table 5.1. Another important measure is the $R^2_{McF}$ statistic. Note that this is only calculated for the training data, while the tables for each model only contain the results for the validation data. The $R^2_{McF}$ statistic for the training data is instead presented in Table 5.14.

![Figure 5.1: Pie chart of significant variables (grouped) for the Swedish model](image)

Figure 5.1: Pie chart of significant variables (grouped) for the Swedish model
5.2.2 Denmark

In similar fashion as above, the results for the Danish model can be seen in Figure 5.2 and Table 5.2 below.

![Pie chart of significant variables (grouped) for the Danish model](image)

Figure 5.2: Pie chart of significant variables (grouped) for the Danish model

<table>
<thead>
<tr>
<th>AUC</th>
<th>Accuracy Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.82</td>
<td>0.64</td>
</tr>
</tbody>
</table>

Table 5.2: AUC and Accuracy Ratio for the Danish model

5.2.3 Finland

The results from Finnish logistic regression model can be seen in Figure 5.3 and Table 5.3 below.
Figure 5.3: Pie chart of significant variables (grouped) for the Finnish model

<table>
<thead>
<tr>
<th>AUC</th>
<th>Accuracy Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.90</td>
<td>0.79</td>
</tr>
</tbody>
</table>

Table 5.3: AUC and Accuracy Ratio for the Finnish model

5.2.4 Norway

Norway was modelled using three different models, as there were significant differences in the behaviour of the customers within the Norwegian product portfolio. For confidentiality reasons these customer groups will be called Type 1 and Type 2. Noteworthy information about the Type 2 customers is that they made up only about a fifth of the customers but accounted for a significantly larger portion of defaults than this. This therefore warranted using a separate model to model the risk-drivers for these customers. Nevertheless, for completeness, a model was also created looking at both Type 1 and Type 2.
Norway all customers

The results for the model concerning all the Norwegian customers can be found in Figure 5.4 and Table 5.4 below.

Figure 5.4: Pie chart of significant variables (grouped) for the total Norwegian model

<table>
<thead>
<tr>
<th></th>
<th>Norway</th>
<th>Deficiency history</th>
<th>External Information</th>
<th>Past due behaviour</th>
<th>Demand chain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Balance</td>
<td>37%</td>
<td>1%</td>
<td>27%</td>
<td>22%</td>
<td>13%</td>
</tr>
</tbody>
</table>

Table 5.4: AUC and Accuracy Ratio for the total Norwegian model

<table>
<thead>
<tr>
<th>AUC</th>
<th>Accuracy Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.88</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Norway Type 1

In Figure 5.5 and Table 5.5 below, the results for the model of type 1 customers can be seen.
Figure 5.5: Pie chart of significant variables (grouped) for the Norwegian Type 1 model

<table>
<thead>
<tr>
<th></th>
<th>AUC</th>
<th>Accuracy Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.90</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Table 5.5: AUC and Accuracy Ratio for the Type 1 Norwegian model

Norway Type 2

In Figure 5.6 and Table 5.6 below, the results for the model of type 2 customers can be seen.
Figure 5.6: Pie chart of significant variables (grouped) for the Norwegian type 2 model

<table>
<thead>
<tr>
<th></th>
<th>AUC</th>
<th>Accuracy Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.89</td>
<td>0.77</td>
</tr>
</tbody>
</table>

Table 5.6: AUC and Accuracy Ratio for the Type 2 Norwegian model

5.2.5 The Nordics

Lastly, the results for the model concerning all the Nordic customers are presented. They can be found in Figure 5.7 and Table 5.7 below.
Figure 5.7: Pie chart of significant variables (grouped) for the Nordic model

<table>
<thead>
<tr>
<th>AUC</th>
<th>Accuracy Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.77</td>
<td>0.54</td>
</tr>
</tbody>
</table>

Table 5.7: AUC and Accuracy Ratio for the Nordic model
5.3 Results for models fitted on outlier-adjusted data

As discussed in Section 4.6, the analysis above was also performed on datasets where outliers in the covariate variables had been removed.

5.3.1 Sweden

As seen in Figure 5.8, the empirical odds curve for this country was deemed to be already adequate, meaning no outlier-analysis was performed.

Figure 5.8: Empirical odds curves for the Swedish model

5.3.2 Denmark

In Figure 5.11 and Table 5.8 below the results for the Danish model with outlier-adjusted covariates are presented. In Figure 5.9 and Figure 5.10 a slight change in the shape of the empirical odds curve can be seen.
Figure 5.9: Empirical odds curves for the Danish model

Figure 5.10: Empirical odds curves for the Danish outlier-adjusted model
Figure 5.11: Pie chart of significant variables for the Danish outlier-adjusted model

<table>
<thead>
<tr>
<th></th>
<th>AUC</th>
<th>Accuracy Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.72</td>
<td>0.44</td>
</tr>
</tbody>
</table>

Table 5.8: AUC and Accuracy Ratio for the Danish outlier-adjusted model

5.3.3 Finland

The results of the Finnish outlier-adjusted model can be seen in Figure 5.14 and Table 5.9 below. In Figure 5.12 and Figure 5.13 a significant change in the shape of the empirical odds curve can be seen.
Figure 5.12: Empirical odds curves for the Finnish model

Figure 5.13: Empirical odds curves for the Finnish outlier-adjusted model
Figure 5.14: Pie chart of significant variables for the Finnish outlier-adjusted model

<table>
<thead>
<tr>
<th>AUC</th>
<th>Accuracy Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.91</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Table 5.9: AUC and Accuracy Ratio for the Finnish outlier-adjusted model

5.3.4 Norway

Norway, all customers

Similar to above, the results for the Norwegian outlier-adjusted model are presented in Figure 5.17 and Table 5.10 below. In Figure 5.15 and Figure 5.16, a significant change in the shape of the empirical odds curve can be seen.
Figure 5.15: Empirical odds curves for the Norwegian model

Figure 5.16: Empirical odds curves for the Norwegian outlier-adjusted model
Figure 5.17: Pie chart of significant variables for the Norwegian outlier-adjusted model

<table>
<thead>
<tr>
<th>AUC</th>
<th>Accuracy Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.91</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Table 5.10: AUC and Accuracy Ratio for the Norwegian outlier-adjusted model

**Norway, customer type 1**

In Figure 5.20 and Table 5.11, the results for the outlier-adjusted model for Norwegian customers of type 1 are presented. In Figure 5.18 and Figure 5.19, a significant change in the shape of the empirical odds curve can be seen.
Figure 5.18: Empirical odds curves for the Norwegian Type 1 model

Figure 5.19: Empirical odds curves for the Norwegian Type 1 outlier-adjusted model
Figure 5.20: Pie chart of significant variables for the Norwegian outlier-adjusted model for type 1 customers

<table>
<thead>
<tr>
<th></th>
<th>AUC</th>
<th>Accuracy Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.91</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Table 5.11: AUC and Accuracy Ratio for the Norwegian outlier-adjusted model for type 1 customers

Norway, customer type 2

The results of the outlier-adjusted model for Norwegian type 2 customers can be seen in Figure 5.23 and Table 5.12 below. In Figure 5.21 and Figure 5.22, a significant change in the shape of the empirical odds curve can be seen.
Figure 5.21: Empirical odds curves for the Norwegian Type 2 model

Figure 5.22: Empirical odds curves for the Norwegian Type 2 outlier-adjusted model
Figure 5.23: Pie chart of significant variables for the Norwegian outlier-adjusted model for type 2 customers

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AUC</strong></td>
<td><strong>Accuracy Ratio</strong></td>
</tr>
<tr>
<td>0.88</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Table 5.12: AUC and Accuracy Ratio for the Norwegian Type 2 outlier-adjusted model

5.3.5 The Nordics

Finally, the results for the outlier-adjusted model for the Nordics are presented. They can be seen in Figure 5.26 and Table 5.13 below. In Figure 5.24 and Figure 5.25, a significant change in the shape of the empirical odds curve can be seen.
Figure 5.24: Empirical odds curves for the Nordic model

Figure 5.25: Empirical odds curves for the outlier-adjusted Nordic model
Figure 5.26: Pie chart of significant variables for the outlier-adjusted Nordic model

<table>
<thead>
<tr>
<th></th>
<th>AUC</th>
<th>Accuracy Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.73</td>
<td>0.46</td>
</tr>
</tbody>
</table>

Table 5.13: AUC and Accuracy Ratio for the outlier-adjusted Nordic model
5.4 Summary of results

For readability, the AUC and Accuracy Ratio of each respective model is summarised in Table 5.14. The measures are presented for both the training data and the validation data.

<table>
<thead>
<tr>
<th></th>
<th>Training data</th>
<th>Validation data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC</td>
<td>AR</td>
</tr>
<tr>
<td>Sweden</td>
<td>0.82</td>
<td>0.64</td>
</tr>
<tr>
<td>Denmark</td>
<td>0.82</td>
<td>0.64</td>
</tr>
<tr>
<td>Denmark OA</td>
<td>0.77</td>
<td>0.55</td>
</tr>
<tr>
<td>Finland</td>
<td>0.92</td>
<td>0.83</td>
</tr>
<tr>
<td>Finland OA</td>
<td>0.90</td>
<td>0.79</td>
</tr>
<tr>
<td>Norway all</td>
<td>0.90</td>
<td>0.80</td>
</tr>
<tr>
<td>Norway all OA</td>
<td>0.89</td>
<td>0.78</td>
</tr>
<tr>
<td>Norway Type 1</td>
<td>0.88</td>
<td>0.76</td>
</tr>
<tr>
<td>Norway Type 1 OA</td>
<td>0.90</td>
<td>0.80</td>
</tr>
<tr>
<td>Norway Type 2</td>
<td>0.95</td>
<td>0.89</td>
</tr>
<tr>
<td>Norway Type 2 OA</td>
<td>0.92</td>
<td>0.84</td>
</tr>
<tr>
<td>Nordics</td>
<td>0.78</td>
<td>0.56</td>
</tr>
<tr>
<td>Nordics OA</td>
<td>0.72</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Table 5.14: AUC, Accuracy Ratio, and $R_{McF}^2$ for all models. OA indicates outlier-adjusted.
Chapter 6

Analysis and Discussion

In this chapter the results from Chapter 5 are analysed and discussed.

6.1 Analysis of each country’s respective models

In the following sections, the results for each country’s models are analysed and discussed.

6.1.1 Sweden

After performing the model fitting it can be seen that the variable grouping that explains the risk of default in Sweden the most is External information, with the group accounting for 64% of the predictive power of the model. This might be seen as the variable being over-predictive - but it is worth noting that the external information available about customers in Sweden is very comprehensive and is known to be accurate. The other significant variable groups also make sense from a credit risk perspective. Past due behaviour and delinquency history logically seem like good indicators of whether a customer is at risk of defaulting. Lastly, the customer information, being for example how long a customer has been with the financial institution, also makes sense as newer customers are known to be more risky.

As for the Accuracy Ratio of this model, it seems to perform quite poorly, with a
value of only 0.61 for the validation data. Note that a value of 0.5 is only random
guessing while a value too close to 1 might indicate over-fitting. An interesting
area of further study would be to exclude external information and see the effects
of the predictive power of this model. The poor performance of the model can also
be seen in the low $R^2_{McF}$ value.

6.1.2 Denmark

It is noteworthy that Denmark is the only country that does not have its exter-
nal information variable grouping as significant. This is largely due to the fact
that the external information about customers is very decentralized and not as
comprehensive as other countries. Instead, customer information, product usage,
and delinquency history make up 60% of the predictive power, followed by differ-
ent behavioural and balance variables. This is interesting as not having access to
good external information leads to the model relying on only internal variables to
predict the default.

This model also has quite a low Accuracy Ratio (for both training and valida-
tion data), indicating poor predictive performance. Furthermore, the $R^2_{McF}$ value
also indicate a poor fit.

6.1.3 Finland

Due to the external information available in Finland is comprehensive, this is a
very significant variable, followed by balance and behavioural variables. It is note-
worthy that being delinquent and past due is not at all as significant as having a
high balance for this model.

The Accuracy ratio of this model is high with a value of 0.79, which indicates
that it has a high predictive power. Furthermore, the fit seems to be about the
same for both training and validation data. The $R^2_{McF}$ value is also a bit higher
than for the Swedish and Danish models.
6.1.4 Norway

Looking first at the predictive power using the Accuracy Ratio on validation data, it is interesting to note that the Type 1 and Type 2 models have higher values of 0.81 and 0.77 compared to 0.75 from the model with both types. This indicates that it indeed is better to separate the customer groups into two different models than to have one model for both. It can be noted that the $R^2_{McF}$ value for the Type 2 model is significantly higher than both the Type 1 model and the overall model. This could be an indication that the quality of fit is higher for this model. However, the predictive power does not seem to be better than the Type 1 model, which has a significantly lower $R^2_{McF}$ value.

As for the significant variables, it can be seen that external information is the most significant for both the overall model and the Type 1-model, followed by balance and past due behaviour. It is not all to surprising that these are similar as the Type 1 customers make up about four fifths of all customers in Norway. The Type 2-model on the other hand, has balance as more significant, and does not include the external information. This indicates that the external information is not comprehensive enough for the Type 2-customers - which might be reasonable as the product and subsequently possibly the customers are different from the customers in Type 1.

6.1.5 The Nordics

Overall, the Nordic model has the worst predictive power out of all of the models, with an Accuracy Ratio of 0.54 on the validation data set. The value of the $R^2_{McF}$ measure is also quite low. This is not all too surprising as a reasonable hypothesis is that the types of customers and customer behaviour differs quite a lot between countries. Furthermore, rules and conditions are not necessarily the same for the products that the financial institute offers in the different countries. One can also note that the suspension variable is only significant in the Nordic model. The reason for this might be that this factor only becomes significant when paired with other countries, while other previously more significant risk-drivers might be too different between customers in the different countries and therefore no longer be
It is interesting to note that an aggregated external information variable grouping was created for the Nordic model, but as noted this does not become significant. This was created by transforming the different countries’ external information to a uniform scale and assigning each customer a new score based on their domestic score. All in all, the Nordic model does not perform at all well, and the true reasons for this are hard to know. An area of further study could therefore be to more deeply examine the different product portfolios, target groups, and customer segments in the different countries, and then categorise the different customers to see if other constellations of Nordic customers can be significant.

6.2 Comparison between models from the same country

As the empirical odds curve of all models except the Swedish one exhibited non-optimal behaviour (that is, not in a straight trend), outlier-adjustments were made in an attempt to improve this. This led to fewer variables being significant than in the non-outlier-adjusted case, which is due to the fact that the analysis was rerun with only the previously significant variables as input, but with outlier-adjusted values. This is obviously an simplification of the process, but as stated before, it would be unreasonable to do this for all variables as there were so many. This is instead an interesting field of further study where one can study the effect of removing outliers for all variables beforehand.

6.2.1 Denmark

The outlier adjustment for the Danish model proved to be significantly worse, with an Accuracy Ratio of only 0.44, compared to a previous value of 0.64 - which is still quite poor. It can be noted that the $R^2_{McF}$ value has decreased from 0.13 to 0.082. This decrease in model performance might be due to the fact that significant observations were actually present in the outliers - and thus removing these also removes significant information about customer behaviour.
Furthermore, the empirical odds curve did not seem to improve much after outlier-adjustments were made, especially since there is still no clear trend throughout the curve.

6.2.2 Finland

In this case, the outlier-adjusted model actually improved somewhat, with an Accuracy Ratio of 0.81, compared to 0.79 (validation data). The $R^2_{McF}$ has decreased a bit, but does not seem to impact the predictive power of the model that much. This indicates that the outliers were in fact observations that previously skewed the results, and when removed did not contain too much significant information about customer behaviour. Furthermore, the empirical odds curve seemed to exhibit a somewhat more even trend. Even though it is not strictly linear, it is at least not parabola-shaped.

6.2.3 Norway

For the overall model in Norway, the largest increase can be observed when adjusting for outliers, with an increase in Accuracy Ratio on the validation data set from 0.75 to 0.81. The Type 1 model increases almost nothing, while the Type 2 model actually decreases in performance. The reasons for this might be due to the same reasoning as previously - either significant information is removed in the outliers, or the significant information is not concentrated to the outliers. It can also be noted that the empirical odds curve for the overall Norwegian model shows a much more even trend when removing outliers, which reasonably reflects in the improved performance of the model. However, the empirical odds curves of the Type 1 and Type 2 Norwegian models improved somewhat, even though the performance for the Type 1 was unchanged while the performance for the Type 2 became worse. This shows that the empirical odds curve is only one tool of many that can be used to assess whether a model might perform well.
6.2.4 Nordics

As for the Nordic model, the second largest decrease in performance is observed out of all models. The Nordics being already a poor performing model, gives even poorer results when outliers are removed. Once again, this is possibly due to too much information being in the outliers - or due to the fact that the countries are too different and thus provide no good predictive power when aggregated. Interestingly, the empirical odds curves shows a very linear trend for the outlier-adjusted model, even though the overall performance of the model is so poor. Once again, the empirical odds curves are only one tool out of many used to assess model performance. In this case, even though the empirical odds curves look good, they do not reflect the over goodness-of-fit of the model.

6.3 Comparison between results from the different countries

Overall, the significant parameters for all countries pertain to being late on payments, how much debt the customer has, or how the customer uses the different products. It is hard to draw any conclusions to exactly why the weights differ as they do, but what can be said is that these risk-drivers are quite logical as late payments and having a lot of debt seems to be factors that drive defaulting behaviour.

In essence, the customers in the countries are not the same, and therefore it makes sense that the significant factors differ. Possibilities to why the order is different for the countries might be due to how the different late payments and demand chain variables are defined, or that the average balance is more evenly distributed in some countries compared to others. Due to confidentiality reasons the actual balances and demand chain specifications cannot be presented more in detail in this thesis. Otherwise this could be an interesting area of further study.
6.4 Discussion about validity of the models

After examining the actual model performance, it makes sense to also discuss the validity of the model based on the data and the information available.

As for the amount of data, it differs quite a lot between countries as the business in these countries is not equally large. Due to confidentiality reasons the actual numbers cannot be disclosed - but this is reasonably an error factor in some countries as the business here is smaller and therefore has less data to perform the regression on.

Regarding the accuracy of the data observed, there are known errors with manual insertion and human error in for example the demand chain and the customer information. These observations were however deemed to be of not significant proportion and should therefore not skew the analysis. This could however be studied further and all strange observations could be removed. However, as stated before, the sheer amount of data makes this process very time-consuming.

A large problem with fitting models against defaults is the fact that defaults in general are very rare and thus the response variable contains significantly more zeros than ones. However, after discussing with senior staff and examining common knowledge in these areas of credit modelling, the amount of defaults was deemed to be of adequate size for the regression to be made.
Chapter 7

Conclusions

Based on the results and discussions in Chapter 5 and Chapter 6 above, one can conclude that the models for Norway and Finland seems to perform the best. The Swedish, Danish, and Nordic model all perform significantly worse. In some cases adjusting for outliers in the explanatory variables improves the performances of the model (e.g. the Norwegian model), in other cases it significantly worsens the performance (e.g. the Danish model).

As mentioned before, the pan-Nordic model fitted to this data does not show good predictive power. One possible explanation for this is that even though the countries are geographically close, there might be differences in regulation and/or product offerings in the countries that significantly alter the default behaviour, and its risk-drivers.

There are several further studies around the subject discussed in this paper that are possible. As mentioned previously, it is possible to more thoroughly perform analysis of the explanatory variables before any model fitting. By for example outlier analysis on a larger set of explanatory variables, a different set of variables might turn significant, and possibly result in models with better predictive power.

Another possibility for further studies is to consider a model other than the logistic regression model considered in this thesis. For example, in Efron et al. 2004 a new
model selecting method is presented. It could be interesting to compare results between that algorithm and the one used in this thesis.
Bibliography


