CREDIT RISK SCORECARD ESTIMATION BY LOGISTIC REGRESSION

Aleksandr Peussa

Pro gradu
May 2016

DEPARTMENT OF MATHEMATICS AND STATISTICS
UNIVERSITY OF HELSINKI

Tiivistelmä — Referat — Abstract

The major concern of lenders is to answer the next question: "Who we lend to?"

Until 1970s the traditional schema was used to answer this question. Traditional credit assessment relied on "gut feel", which means that a bank clerk or manager analyses a borrower's character, collateral and ability to repay. Also, some recommendations from the borrower's employer or previous lender are used.

The alternative approach is credit scoring, which is a new way to approach a customer. Credit scoring is one of the most successful applications of statistics in finance and banking industry today. It lowers the cost and time of application processing and gives flexibility in making trade off between risk and sales for financial institution.

Credit scorecards are essential instruments in credit scoring. They are based on the past performance of customers with characteristics similar to a new customer. So, the purpose of a credit scorecard is to predict risk, not to explain reasons behind it.

The purpose of this work is to review credit scoring and its applications both theoretically and empirically, and to end up with the best combination of variables used for default risk forecasting. The first part of the thesis is focused on theoretical aspects of credit scoring - statistical method for scorecard estimation and measuring scorecard's performance. Firstly, I explain the definition of the scorecard and underlying terminology. Then I review the general approaches for scorecard estimation and demonstrate that logistic regression is the most appropriate approach. Next, I describe methods used for measuring the performance of the estimated scorecard and show that scoring systems would be ranked in the same order of discriminatory power regardless the measure used.

The goal of the second part is empirical analysis, where I apply the theoretical background discussed in the first part of the master's thesis to a dataset from a consumer credit bank, which includes variables obtained from the application forms and from credit bureau data, and extracted from social security numbers.

The major finding of the thesis is that that the estimated statistical model is found to perform much better than a nonstatistical model based on rational expectations and managers' experience. This means that banks and financial institutions should benefit from the introduction of the statistical approach employed in the thesis.

Contents

Chapter 1

Introduction

As soon as human beings started to cooperate and create a community, there are no doubts that they started to borrow and repay as well. The first known case of lending is presented by ancient Babylon about 5000 years ago: "Two shekels of silver have been borrowed by Mas-Dchamach, the son of Adadrimeni, from the Sun priestess Amat-Schamach, the daughter of Warad-Enlil. He will pay the Sun-God's interest. At the time of the harvest he will pay back the sum and the interest upon it." (Lewis 1992, vi.)

So, it is noticeable that both credit and interest rate was practiced since ancient times. On the other hand the attitude of society to interest rate varies significantly in different historical eras. For instance, in Medieval Age interest rate was forbidden, but nowadays interest rate is an extremely important mechanism, which drives modern economics and compensates lender for his risk.

The major concern of lenders is to answer the next question: "Who we lend to?" From lender's perspective the loan should be granted only to those customers, who are going to return the loan back. But how is the lender able to reveal the nature of a potential customer 'a priori'?

Until 1970s the traditional schema was used to answer this question. Traditional credit assessment relied on "gut feel", which is as old as borrowing itself. In other words, it means that a bank clerk or manager analyses borrower's character, collateral and ability to repay. Also, some recommendations could be needed from the borrower's employer or previous lender. (Thomas 2002, 9.)

Such a schema inevitably decreases supply because each lender faces only one customer. Furthermore, the overall process was slow, inconsistent and subjective. In order to receive satisfactory results, the manager should have years of experience in such work.

Over the the last quarter of the twentieth century lending to consumers has exploded. The introduction of mail-order companies, IT-booms and general population's facility to access into car market led to a drastic increase in credit demand.

Astounding growth of banks portfolios, which was affected by exogenous shocks in credit demand, led to a deeper penetration of the risk pool of new customers with unpredictable behavior. Moreover, the development of information technology opens possibilities to apply for a loan via telephone or internet, which makes the traditional schema almost useless. The need to process applications rapidly and effectively became as a major aim for financial institutions, which are trying to maximize their profits. (Lewis 1992, 17.)

Finance companies with the traditional approach are no longer capable of facing growing demand, which indicates a rise of the new era in credit assessment.

The history of credit scoring is only 60 years. Credit scoring is an essentially new way to approach a customer. Its philosophy is pragmatism and empiricism. Credit scorecards are based on the past performance of customers with characteristics similar to a new customer. So, the purpose of the credit scorecard is to predict risk not to explain reasons behind it. The credit-scoring system is not able to identify the future behavior of the customer based on his individuality, but it provides statistical odds that an applicant with a given score will be "good" or "bad". (Siddiqi 2006, 5.)

The opposition of the credit scoring came from people, arguing that credit decisions should be based only on a reasonable explanation (Capon 1982) why certain variables affect the risk this way, and from lawyers, who state that it is illegal to use some characteristics such as race, religion and gender. The most interesting fact that some countries reject gender as a scorecard variable is that they believe it will discriminate against women. In fact, most studies (Chandler and Ewert 1976) have shown that usage of scorecard with gender as an explanatory variable increases the number of loans for women.

In spite of the critics, credit scoring is one of the most successful applications of statistics in finance and banking industry today. It lowers the cost and time of application processing and gives flexibility in making trade off between risk and sales for financial institution. So, the introduction of credit scoring changes objectives of companies from minimizing the loss from every distinct customer to maximizing the overall profit. (Nevin and Churchill 1979, Saunders 1985.)

Furthermore, companies, which took credit scoring into account, experienced more than 50 % drop in default rates, which was a significant indicator that credit scorecards are much better in the risk separating than any judgemental scheme. (Myers and Forgy 1963, Churchill 1977.)

The purpose of this work is to review credit scoring and its applications both theoretically and empirically.

The first part of the thesis focuses on theoretical aspects of credit scoring - statistical methods for scorecard estimation and measuring its performance. Firstly, I explain the definition of scorecard and underlying terminology. Then I review the main approaches for scorecard estimation, both statistical and non-statistical. Next, I show methods for

measuring scorecard performance.

The goal of the second part is empirical analysis, which is based on the theoretical part and data from a consumer credit bank. I use the logistic regression and Gini coefficient as the performance measuring instrument for reasons discussed in the first part of the work. For security reasons I am not able to reveal the description of the data and name of the bank. Variables are presented as capital letters but are not explained.

Chapter 2

Basics of credit scoring

2.1 Definition of scorecard

The credit-granting process leads to two choices - grant a loan to a new customer or decline his application. The purpose of the scorecard is to assist this decision. So, credit scoring is a tool used to evaluate the level of risk associated with a certain applicant. This tool consists of a group of variables, statistically significant to be predictive in separating goods and bads. Scorecard variables may be selected from any sources of information available to the lender at the time of application. (Thomas 2004, 95.)

More formally credit scorecard is a statistical model, which predicts a probability of default for an applying customer with certain characteristics. In banking, default is failure to meet the legal obligations (or conditions) of a loan, for instance when a customer fails to make a payment. The scorecard attributes a score (number) to a customer or the estimated probability that the person will default his loan.

Mathematically, scorecard can be represented as

$$(2.1) \qquad\qquad p = f(X_1, X_2, \ldots, X_m),$$

where the probability (score) p is a function of the variables $X_1, X_2, \ldots, X_m$. For instance, marital status, age, occupation, type of accommodation, income and gender are candidates for credit scorecard variables.

The credit scoring is based on the behaviour of previously accepted accounts and can also be used to determine the initial credit limit a customer is offered.

2.2 Sources of information

Financial intermediary has to collect all data, which is relevant for risk management. The characteristics available to discriminate between the good and the bad are of three types - those derived from application form, those available from a credit bureau search, and, for behavioural scoring only, those describing the transaction history of the borrower. Firstly, I concentrate on the application characteristics in subsection 2.2.1 and then deal with the credit bureau data in the subsequent subsections.

2.2.1 Application data

Application data is information provided by the potential customer. The application form could contain such characteristics as salary, occupation, number of children, other loans and so on. When designing an application form, a financial institution faces a trade-off between simplicity of the form and the quantity of the information.

A detailed application form, with a great variety of variables is attractive from the risk management perspective. Clearly, the more variables available a lender has, the better risk separating scorecard can be constructed.

However, a complicated application process decreases the probability of its completion, which cuts sales. Therefore, there is a pressure on the lender to make the form as simple as possible. Some variables are not permitted for legal reasons. For instance, the U.S. Equal Credit Opportunity Acts of 1975 and 1976 made it illegal to discriminate in the granting of credit on the grounds of race, colour, religion, national origin, sex, marital status, or age. (Thomas 2002, 124.)

Having considered the application questions, the bank has to decide on the appropriate answers. For instance, in case of occupation, should the bank allow the customer to answer by himself or should he face a limited set of responses? The negative consequence of leaving an answer open is that the bank could face almost unlimited variations of answers, which makes risk analysis impossible. The reason behind this is that the lack of the observations per each variable's outcome leads to biased and inconsistent results. For instance, "working with Dad" is a good example of a badly defined occupation type. The limited set of responses could be something like this:

- permanent

- part time

- pensioner

- entrepreneur

- student

- unemployed.

An applicant has to choose between several options the one, which reflects his current occupation status best. Now financial intermediary has a sufficient quantity of observations per each outcome, which allows to use the empirical data for the risk analysis.

On the other hand, there are questions, where restricted answers are not necessarily the best way to handle the problem - income questions, for instance. Here the lender is supposed to be very careful in the wording of the question in order to clarify what type of income is needed. There is a big variety of incomes: monthly or yearly, net or gross, applicant's or household's, and so on.

The application data is a significant instrument from the risk management perspective but it contains problems as well. First of all, the data should be carefully looked through and validated. The most frequently arising problems with the application data are frauds and errors. An applicant could unintentionally violate information, putting wrong income, for example. This problem could be partly solved by revealing and eliminating impossible or inconsistent answers.

The frauds are a more serious problem. Here, an applicant intentionally violates application data in order to receive a positive loan decision or better offer. Moreover, if intermediary has economic incentives for sending good customers, they might temp to advise the applicant on suitable answers. So, from risk perspective, the usage of only application data could lead to major problems in the long run. This is the reason why credit bureau data is a significant part of risk management and credit scoring.

2.2.2 Credit bureau data

A credit bureau or credit reference agency, or consumer reporting agency is an organization, which collects data from various sources and provides consumer credit information on individual consumers. The data can be information about applicant's borrowing and bill-paying habits. The purpose of the credit bureau is to reduce the impact of asymmetric information between borrowers and lenders. Most of the banks sign agreements with credit bureaus in order to receive this information. Otherwise, banks would face problems of adverse selection and moral hazard, unintentionally granting loans to high risk applicants. Generally, financial institutions prefer to have a contract with one credit bureau only in order to minimize their expenses.

Though credit bureaus exist in many countries, their role and legislative framework can vary significantly in different countries (Thomas 2002, 126). In other words, the stage of integration of credit reference agencies is not identical from country to country. The higher the stage of integration, the larger the customer base and better information

quality provided by credit bureaus. For instance, Western Europe has experienced a stage of development, whereas in Eastern Europe the credit bureau integration issue has only recently started and their role is relatively small (Thomas 2002, 15).

In the U.S. and U.K. credit bureaus are well established (Thomas 2002, 15). This means that credit reference agencies are state owned or there is a small number of very large players in the market. This is attractive from lender's perspective, who desires to sign an agreement with only one credit agency. Otherwise, financial institution has no other choice than to sign multiple agreements with different credit bureaus in order to cover total customer base, which is much more expensive.

But, in spite of the high level of establishment, regimes in the U.S. and U.K. start from the opposite ends of the legislative framework spectrum. Roughly, in the U.S., information is available unless there is a good reason to restrict it, whereas in the U.K., information is supposed to be restricted unless there is a good reason to make it available. So, it means that financial institutions in the U.S. have much more valuable information than in the U.K. (Thomas 2002, 16.)

From the scorecard perspective, the position of a credit agency is very important. The score could be based on credit bureau data, which contains millions of applications and historical records. There are two advantages for such data: credit bureau data is validated and customer base is much larger than bank's own portfolio. So, a lender protects himself against violated or fraud information, which could happen if he uses application data. Moreover, if a financial institution is a new and small player in the market, its own portfolio is too small to be used for estimating a robust and consistent scorecard model.

Data provided by credit bureau could be divided into the following main categories:

- publicly available information

- previous searches

- shared contributed information

- aggregated data

- fraud warnings

- generic scorecard.

Publicly available information

Publicly available information consists of address, demographic data, tax returns, taxable income and any non-payment records that might be registered on the applicant.

Previous searches

When a financial institution makes an applicant check via credit bureau, the check is recorded on consumer's file of the credit agency. So, when another lender makes an inquiry on the same consumer, a record of any previous searches will be visible for the lender. This type of information contains dates and details of companies which made checks on this particular consumer.

Previous search data has limitations, which decrease its value. For instance, a bank is not able to check the final decision of the previous lender. On the other hand, a great number of inquiries could indicate that the consumer's profile represents a high risk from lender's perspective. For example, he is desperately short of money and applies to every financial institution and takes everything that he can get.

Shared contributed information

Like casinos, which in spite of rivalry, reveal to each other information about cardsharpers, lenders realized that it is profitable to share information on how consumers perform on their accounts. This data can be viewed and used by other financial institutions both from risk and marketing perspective. So, if an applicant has loans in other banks, a credit bureau will provide the payment behavior history and current account status of this consumer.

Unfortunately this is a simplified version of the reality. Actually, the information from other lenders is not complete. For example, financial institutions could reveal information only about particular products or consumer's payment history could be limited. Furthermore, a lender could provide information to a credit bureau only about such accounts, which are in pre-collection or collection stages. In the latter case, it is an issue of a default notice or, in other words, a payment remark.

Aggregated data

Having data from different lenders and other sources of information, such as Population Register Centre, provides credit bureau with a powerful tool in creating new measures that could be used in credit scoring. One of them is the possibility to create variables at the postcode level or at an even lower level. In other words, a credit bureau could divide country into bricks, where every brick contains a few dozens of households. Aggregating all relevant information by bricks, a credit agency is able to create measures such as:

- purchasing power

- education level

- life stage

- type of residential area

- ownership of housing

- risk of payment defaults.

So, if a lender knows the address of an applicant, then a credit bureau could provide the financial institution with all information based on the brick, where the applicant lives. Although aggregated information doesn't represent the applicant's characteristics by certainty, it gives the lender an estimate of his nature assuming that individuals are homogeneous within the particular brick.

Fraud warnings

One of the main disadvantages of application data is a violation of information, particularly frauds. In order to eliminate this problem, credit bureaus collect and record information about fraud incidences. For instance, frauds could be related to impersonation or address quoted. Such violations are a high risk from lender's perspective and bank's attitude to such applicants could vary significantly. Most lenders use this type of information in order to increase the level of applicant analysis, particularly concerning the application data. Moreover, from some banks' perspective it could be a reason for application rejection.

Generic scorecard

Credit bureaus also act as intermediaries between different lenders. Banks contribute to the credit agency details of customers' payment behaviour. This information received is used for creating a generic scorecard.

A generic scorecard is a model based on all relevant information provided by credit bureaus and not related to specific lender's experience. There are at least three reasons why a generic scorecard is extremely useful for financial institutions:

- Bank is too small to be able to construct its own scorecard. In order to keep his risk under control, the lender uses a package of generic scorecards and his own experience.

- The introduction of a new product. In this case, a scorecard provided by a credit bureau could be interesting even for a larger financial intermediary because of lack of information about the product.

- Lender's model has omitted bias problem. Credit bureaus, also acting as intermediaries between different lenders, contribute details of customers' payment behaviour into generic scorecard, which may add explanatory power to bank's own model.

On the other hand, blind usage of a generic scorecard could have its own disadvantages because the lending institution has no clue about the variables and weights used for generating the score. In the worst scenario, if bank's customer base completely differs from credit bureau's base, the generic scorecard could provide the lender with inconsistent and biased results.

2.2.3   Behavioural data

Behavioural data is a history of existing customers' transactions and cash flows. In other words, the data is a conglomeration of customers' payment characteristics and habits. The most common ones are minimum, maximum and average balance, and total value and regularity of both debit and credit transactions. Other characteristics, which are extremely important, are number of reminder letters and quantity of consecutive missed payments.

The payment behaviour data is a very significant tool when predicting future loan performance. For example, data received from credit information institutions could be used by a lender for risk-based pricing. This means that applicants face different interest rates depending on their previous payment behaviour. Customers with poor credit quality face a higher interest rate, whereas customers with excellent bill-paying habits face lower interest rate, which seems reasonable.

On the other hand, such risk-based strategy is not necessarily the best way for a lender, because a non-paying customer will not be profitable regardless of the interest rate. So, the only profit the lender receives is from customers with a good performance, which means that risk-based pricing lowers the bank's revenue.

2.3   Goods and bads

The important part of the scorecard development is to decide how good and bad applicants can be distinguished. Words "good" and "bad" are theoretical terms, which refer to bank's view on customers' paying performance. For simplicity, a good customer could be defined as creditworthy, paying the loan back, while a bad customer as lacking payment performance.

Defining a bad does not necessarily mean that all other cases are good. Actually, by credit scoring theory, accounts are normally divided into goods, bads and non-determinates. The last ones contain customers that can't be classified in a simple way as good or bad.

Generally, there are two main reasons for this - either customer's performance is between "goods" and "bads" or information about payment performance is lacking. By the former I mean customers who repeatedly miss payments and by the latter I mean customers, who have just signed an agreement. Therefore, there is no behavior information available for these. Both types of customers are normally excluded from the scorecard building process. (Siddiqi 2006, 32.)

From a bank's perspective:

- Goods are generally customers that have operated the account satisfactorily without being bankrupt and have been never or rarely delinquent. The term delinquent commonly refers to an individual with a contractual obligation to make payments against a loan in a timely manner but payments are not made on time. In other words, a delinquent customer is one, who missed his payment.

- Bads are customers, who miss three consecutive payments and the bank breaks their agreements. As a result, the customers' loans are sold to a collection agency, which turns to a profit decrease for the bank.

The division between non-determinates and goods is not so straightforward. Generally, non-determinates might be those that have one or two payments missing. Thus, they cause problems and additional pre-collection activity - perhaps repeatedly if these customers are delinquent several times but have never become three payments down.

2.4   Collection and pre-collection

Collection is a process when a bank breaks a contract with a non-paying customer and sells his loan to a collection agency for 60-80 percent of the loan's value. From bank's perspective, breaking a contract is equivalent to a loss, so financial institutions try to prevent customers from going into collection.

The purpose of pre-collection is to prevent clients from going to the collection. Pre-collection is a department or division within a bank, which contacts the customers with missed payments and investigates their status. There are three different ways contacting the customer - by telephone, email or message. Those, who eventually paid the loan back, belong to non-determinates category. Goods have rare or no missed payments at all by definition and bads will inevitably end up with the collection. Non-determinates could be, for instance, customers, who repeatedly forget to pay bills or who face several bills, but have limited budget constraint. In the latter case, efforts of pre-collection could turn a customer to pay the bill.

2.5 Credit policy and scorecard

Credit policy is a set of rules and models with techniques that aid lenders to determine who will get a credit, how much consumer credit a customer should get and what operational strategies will enhance the profitability of borrowers. Regardless of the techniques used, it is necessary to have a large sample of previous customers with their applications and behavior data available. Most of the techniques use the previous customers' sample to identify the connection between the characteristics of the consumers picked from the sources of information discussed in section 2.2 and their subsequent performance. One of the most powerful techniques, which meets these requirements is credit scorecard.

From the bank's perspective there is always a trade-off between sales and risk. A financial institution is able to increase sales by increasing the number of approved customers, which could be achieved by mild credit policy. But this would lead to a drop in customer quality. So, the bank should accept a larger portion of not only good, but also bad customers. On the other hand, too tight credit rules will negatively affect the sales and market share of the bank.

So, scorecard is used for determining desired trade-off between risk level and approval rate. Approval rate is the ratio of customers, who are approved by the bank's credit policy and is calculated as quotient of the quantity approved to the total number of customers applied, so that approval rate represents sales. The desired trade-off between sales and risk could be achieved by setting up an appropriate score limit. In the credit scoring context, such limit is called a cut-off.

2.6 Definition of cut-off

Most organizations that use scorecards set minimum score levels at which they are willing to accept applicants (or qualify them for any subsequent account treatment with behaviour scorecards). This minimum score is referred to as a "cut-off", and can represent a threshold risk, profit, or some other level, depending on the organization's objectives in using the scorecard. (Siddiqi 2006, 146.) A simple cut-off strategy function is

$$D(s) = \begin{cases} reject, & \text{if } s \geq s_{min}, \\ accept, & \text{if } s \ < s_{min}, \end{cases}$$

where $D(s)$ is a decision function, s is an applicant's score and s_{min} is a desired cut-off.

In this example, if applicant's score is equal to or below s_{min}, he is approved automatically and if his score is above s_{min}, he is rejected.

Chapter 3

Theory of scorecard building

Methods for creating credit scorecards can be divided into two groups: statistical and non-statistical. Until the 1980s, the only approaches used were statistical methods like contingency tables or linear regression models. Still today statistical analysis is the most general approach in constructing credit scorecards. The reason behind this is that one could benefit from knowledge of the sample estimators and their properties, confidence intervals and hypothesis testing in the credit-scoring context. This knowledge could be used for evaluating relative importance of different characteristics: both for ensuring the significance of relevant variables and removing unimportant ones.

The idea in constructing credit scorecards is to use a statistical approach to a sample of past customers in order to reveal the existing or new applicants who are likely to be satisfactory.

The aim of this chapter is to describe categorical data and focus on logistic regression for constructing a scorecard. My decision to use the logistic regression was based on a publication of Desai (1997), where the author compared statistical and non-statistical approaches to credit scoring, and demonstrated that logistic regression is the most appropriate approach.

3.1 Categorical data

A categorical variable has a measurement scale consisting of a set of categories. For instance, occupation could be measured as "permanent", "fixed term", "unemployed", "student" and "pensioner".

The development of categorical data analysis was followed by an increasing need to analyse information collected for both the social and biomedical sciences. In the case of credit scoring, methods for categorical data analysis are pervasive because most of the

variables used in credit scorecard are categorical.

Categorical variables could be classified in different ways. In fact, statistical analysis may distinguish between a response variable and explanatory variables. From credit scoring perspective the response variable represents customer's quality (bad or good) and explanatory variables are a list of characteristics, which have discriminating power, or in other words, which strongly affect on being a good or a bad customer. So, a statistical model shows how the response variable is influenced by explanatory variables. A credit scorecard can use as predictors all variables collected from applications and credit bureau data.

From bank's perspective there are only two possible outcomes: accept or reject the customer. The terms of lending for all accepted customers are the same. So, there is no advantage in classifying this performance into more than two classes - the goods and the bads. The goods are those, who pay their loans back and the bads are those, who eventually miss payments and are transferred to collection. Variables with only two categories are called binary. So, the purpose of this chapter is to focus on modeling binary response variables.

3.2   Logistic regression model

In binary or dichotomous logistic regression the response variable $y \in \{0,1\}$ follows a Bernoulli distribution. The response variable reflects the creditworthy of the customer:

$$y = \begin{cases} 1, & \text{if a customer goes to collection,} \\ 0, & \text{if a customer continues to make payments.} \end{cases}$$

The observations $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ should be independent, where n indicates the number of the observations. This assumption holds for the bank's dataset: potential customers and their characteristics are independent from each other because only one person per household is able to receive a loan.

The logit link function, which is an inverse of the standard cumulative distribution function of the logistic distribution, is applied to each component of $E(\boldsymbol{y})$ that relates it to the linear predictor $X\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a $(m+1) \times 1$ regression coefficient vector and X is a $n \times (m+1)$ model matrix containing n observations of m explanatory variables and a constant term. (Agresti 2015, 2.)

Let $p_i = P(y_i = 1)$ be the probability of $y_i = 1$ and $\text{logit}(p_i) = \log(p_i \, / \, (1 - p_i))$ be the logit link function for observation i. The logistic regression model has a linear form for the logit:

$$(3.1) \qquad \text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{\beta}^T \boldsymbol{x}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi}$$

where $\boldsymbol{x_i}$ is a $m+1$ vector, which contains 1 and m categorical or continuous explanatory variables.

The probability p_i can be derived from equation (3.1) by using the exponential function:

$$(3.2) \qquad p_i = \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)}{1 + \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)}.$$

3.2.1 Estimation

Maximum likelihood estimation is the standard way to estimate a logistic regression model (Agresti 2007, 6). The likelihood function $L(\boldsymbol{\beta})$ is the probability for the occurrence of a sample configuration $y_1, ..., y_n$ given the Bernoulli probability density for y_i. In the case of logistic regression

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$$

where n is a number of observations, y_i is equal to 1 if the customer goes to collection and equal to 0 otherwise, and p_i is defined in equation (3.2).

The log-likelihood function is

$$l(\boldsymbol{\beta}) = logL(\boldsymbol{\beta}) = \sum_{i=1}^{n}[y_i log(p_i) + (n_i - y_i)log(1-p_i)].$$

Its first derivative $S_j(\boldsymbol{\beta})$ with respect to β_j is

$$S_j(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n}(y_i - n_i p_i)x_{ij} = \sum_{i=1}^{n}(y_i - \mu_i)x_{ij}$$

where $\mu_i = n_i p_i$. In matrix form $\boldsymbol{S}(\boldsymbol{\beta}) = X^T(\boldsymbol{y} - \boldsymbol{\mu})$, which consists of the derivatives $S_j(\boldsymbol{\beta})$, where $\boldsymbol{\mu} = E(\boldsymbol{y})$.

The negative inverse of the second derivative $I_{jk}(\boldsymbol{\beta})$ with respect to β_{jk} is

$$I_{jk}(\boldsymbol{\beta}) = -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_j \beta_k} = \sum_{i=1}^{n} n_i p_i(1-p_i)x_{ij}x_{ik}$$

where $I_{jk}(\boldsymbol{\beta})$ is the jk element of the expected information matrix $I(\boldsymbol{\beta})$.

Under standard regularity conditions (Cox and Hinkley 1974, 281), for large n the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ has an approximate normal distribution. The approximate covariance matrix is the inverse of the expected information matrix $I(\boldsymbol{\beta})$.

Generalizing from the typical element of the information matrix to the entire matrix, with the matrix of explanatory variables X, $I(\boldsymbol{\beta}) = X^T W X$, where W is the diagonal matrix with main-diagonal elements $w_i = n_i p_i (1 - p_i)$. Accordingly, $\text{Cov}(\boldsymbol{y}) = W$. In summary, $\hat{\boldsymbol{\beta}}$ has an approximate $N[\boldsymbol{\beta}, (X^T W X)^{-1}]$ distribution (Agresti 2015, 126).

To find the MLE $\hat{\boldsymbol{\beta}}$, the Newton-Raphson method can be used. The Newton-Raphson method iteratively solves nonlinear equations, so, the method can be used for maximizing the log-likelihood function in the case of logistic regression (Agresti 2013, 143). The current estimate of $\boldsymbol{\beta}$, $\boldsymbol{\beta}_{(k)}$ is used to compute an approximation of the expected value $\boldsymbol{\mu}_{(k)}$ and covariance matrix $W_{(k)}$. After that, the next approximation of the estimate is obtained as

$$\boldsymbol{\beta}_{(k+1)} = \boldsymbol{\beta}_{(k)} + (X^T W_{(k)} X)^{-1} X^T (\boldsymbol{y} - \boldsymbol{\mu}_{(k)}).$$

The process is repeated until the estimates stop changing, that is, until $\boldsymbol{\beta}_{(k+1)}$ is sufficiently close to $\boldsymbol{\beta}_{(k)}$. The MLE $\hat{\boldsymbol{\beta}}$ is the limit of $\boldsymbol{\beta}_{(k)}$ as $k \to \infty$.

For large samples, the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$ is $\hat{\text{Cov}}(\hat{\boldsymbol{\beta}}) = (X^T \hat{W} X)^{-1}$, where $\hat{W}$ denotes the $n \times n$ diagonal matrix having $n_i \hat{p}_i (1 - \hat{p}_i)$ on the main diagonal, where $\hat{p}_i$ is obtained by replacing $\boldsymbol{\beta}$ in (3.2) by $\hat{\boldsymbol{\beta}}$ (Agresti 2015, 127).

3.2.2 Hypothesis testing

The customer base of a bank contains thousands of customers, so it is possible to test hypotheses by a Wald test based on the large-sample distribution of the MLE $\hat{\boldsymbol{\beta}}$.

In general, I test the hypothesis $H_0 : \beta_j = 0$. The test-statistic asymptotically converges to normal distribution, when $n \to \infty$ and is obtained by computing the ratio of the MLE to its estimated standard error:

(3.3)
$$z = \frac{\hat{\beta}_j}{\sqrt{\hat{\text{Cov}}(\hat{\beta}_j)}} \underset{as}{\sim} N(0,1)$$

where $\hat{\text{Cov}}(\hat{\beta}_j)$ is the estimated variance of the MLE, and a two-tail p-value is obtained as $P(Z > |z|)$, where $Z \sim N(0,1)$.

Chapter 4

Gini coefficient and model selection criteria

Having estimated a credit scorecard, two obvious questions arise, "How to measure its performance?" and "How to compare potential models with each other?"

A perfect model distinguishes goods from bads, allowing the financial institution to accept only the good ones (creditworthy) and rejecting other applicants. Although, perfect models don't exist in practice, banks try to estimate a model, which maximizes the proportion of good customers approved and maximizes the proportion of bad customers rejected.

There are a number of different methods to measure scorecard performance and compare it with competing models like mean difference or Gini coefficients. These methods are based on the same principle, therefore the scoring systems would be ranked in the same order of discriminatory power regardless the measure used. (Thomas 2002, 86.)

First, I focus on one of the most popular methods to measure scorecard performance, the Gini coefficient. This method is useful in testing the model's efficiency from a business perspective. I have chosen this measure because of its intuitiveness. Next, I review the main types of criteria, AIC and BIC, used for model selection.

4.1   Lorenz diagram

In order to understand the logic behind the Gini coefficient, I first introduce the Lorenz diagram in Figure 4.1. In this diagram the horizontal axis presents the proportion of bads $Pr_b(s)$ rejected by the scorecard for every score $s \in [0, 1]$ the sample includes against the proportion of goods $Pr_g(s)$ presented on the vertical axis, where 0 is a lowest possible score and 1 is a highest possible score. By score I mean the value, which the scorecard generates
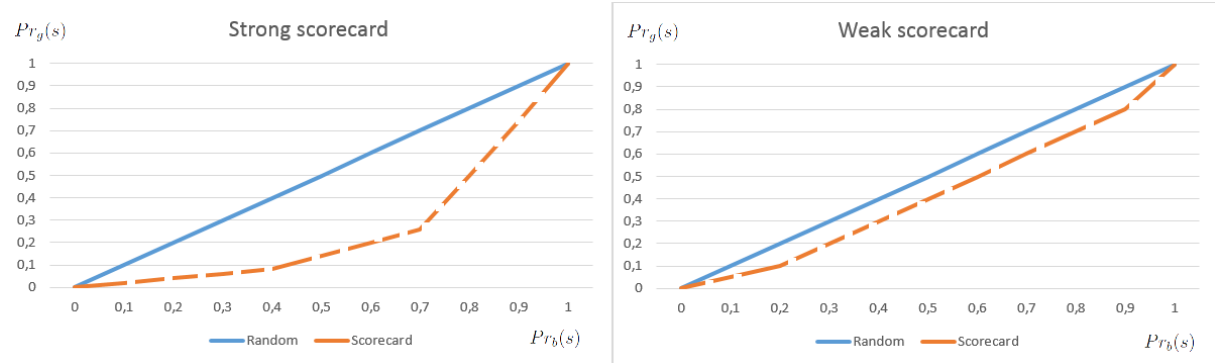
Figure 4.1: Lorenz diagram

for the customer. In case of my scorecard (based on logistic regression), the score is the estimated probability that the person will default his loan. The proportions $Pr_b(s)$ and $Pr_g(s)$ run from 0 to 1. Bottom left corner represents the lowest possible score 0, at which there are no rejections. In contrast, the upper right corner represents the highest possible score 1, at which scorecard rejects all customers. A straight line (diagonal) drawn from the bottom left corner of the chart to the top right corner represents a scenario, where the scorecard has no discriminating power at all, meaning that it is no better than classifying the cases randomly. So, at every score, the scorecard rejects the same percentage of goods and bads $Pr_b(s) = Pr_g(s)$.

The curve located below the straight line represents a set of proportions $Pr_b(s)$ and $Pr_g(s)$ for every score produced by estimated model and is called receiver operating characteristic (ROC). The better the scorecard, the larger is the difference between proportions of goods and bads for every score. Graphically, the further from the diagonal the ROC curve is, the better is the scorecard.

4.2   Gini coefficient

Let the area between the diagonal and ROC curve be DR and the area between ROC curve and bottom left axes AR. The area of the triangle defined by bottom left axes and diagonal is 0.5. The Gini coefficient is defined as the ratio of the area between the diagonal and ROC curve to the area of the triangle:

$$G = \frac{DR}{0.5} = 2(0.5 - AR) = 1 - 2AR$$

Mathematically, the area between ROC curve and bottom left axes AR can be defined as follows:

$$AR = \sum_{s=0}^{1} Pr_b(s)[Pr_g(s-1) + Pr_g(s)]/2$$

where $Pr_b(s)$ is the proportion of bads, $Pr_g(s)$ is the proportion of goods, 0 is the lowest score attained and 1 is accordingly the highest attainable score. The proportions $Pr_b(s)$ and $Pr_g(s)$ are aggregations of bads and goods for every score s, which means that within the formula each score s is unique. Accordingly, the Gini coefficient is defined as:

(4.1)
$$G = 1 - \sum_{s=0}^{1} Pr_b(s)[Pr_g(s-1) + Pr_g(s)]$$

The range of the Gini coefficient is therefore from 0 to 1. In case of a perfect scorecard, the Gini coefficient is equal to 1, which means that the scorecard drops all bads keeping all goods. In case of a weak scorecard, the Gini coefficient is close to 0, which indicates that the model does not pick the difference between goods and bads. (Siddiqi 2006, 125.)

The Gini coefficient is a powerful tool for measuring scorecard performance. I will use it for comparison of models.

4.3   Model selection criteria: AIC and BIC

The main aim of these methods is to check whether over-fitting occurs within a model. This problem generally arises when a model is excessively complex, having too many parameters relative to the number of observations. A model that has been over-fit will generally predict poorly, as it can exaggerate minor fluctuations in the data.

Given a set of candidate models, a model selection criterion allows us to select the most appropriate model, which has a balance between the goodness of the model, measured by the maximum value of the likelihood function, and simplicity. The model's simplicity is generally measured by counting the number of parameters in the model.

The Akaike information criterion (AIC) is a measure of the relative quality of statistical models for a given set of data. It is defined as

$$AIC = -2\ln(\hat{L}) + 2k$$

where $\hat{L}$ is the maximum value of the likelihood function for the model and k is the number of parameters in the model. The value of AIC increases ceteris paribus, when the number of parameters increases, and decreases ceteris paribus, when the likelihood increases. The most appropriate model from the selection criteria perspective will minimize the AIC value. (Akaike 1973.)

An alternative is the Bayesian information criterion (BIC):

$$BIC = -2\ln(\hat{L}) + k\ln(n)$$

The BIC generally penalizes for extra parameters more strongly than AIC, because it depends on the relative magnitude of n (Kass 1995). In other words, for AIC the coefficient for k is always 2, while for BIC the k's coefficient $\ln(n)$ is greater than 2 if $n \geq 8$, which holds always in practice.

Chapter 5

Scorecard development

This chapter presents the empirical part of my master thesis. First, I review the data and list the variables. Next, I estimate a few competing scorecards and present the estimation results. Then I provide comparison of the scorecards using AIC, BIC and the Gini-coefficient to end up with the final scorecard.

5.1 Data

The data contains 9 explanatory variables and the response variable. The response variable shows whether a customer defaults his loan within 8 months after disbursement:

$$y = \begin{cases} 1, & \text{if a customer is in collection within 8 months after loan paid out,} \\ 0, & \text{if a customer makes payments within 8 months after loan paid out.} \end{cases}$$

The explanatory variables, both categorical and continuous, are picked from the application, credit bureau data and social security number (age and gender). The data consists of 2458 independent observations from the year 2014, where every observation is a bank's customer. The observations are independent because only one person per household is able to receive a loan. For security reasons I'm not able to reveal the labels of the variables, so that all explanatory variables will be presented as capital letters:

- M - categorical variable (3 categories)

- G - categorical variable (2 categories)

- Y - continuous variable

- R - categorical variable (2 categories)

- P - categorical variable (2 categories)

- A - categorical variable (3 categories)

- O - categorical variable (3 categories)

- W - continuous variable

- I - continuous variable

5.2   Estimation

I examine which variables predict customers' probability of the default by estimating a logistic regression, on which the scorecard is based.

I regress the response variable on all explanatory variables:

Table 5.1: Estimation results : Scorecard 1

Variable	Coefficient	Std. Err.	p-value
Intercept	-1.630	0.448	0.002
M1	0.367	0.198	0.064
M2	0.629	0.207	0.003
G1	0.339	0.157	0.031
Y	-0.029	0.009	0.001
R1	0.698	0.149	0.000
P1	-0.596	0.183	0.001
A1	-0.290	0.222	0.192
A2	0.059	0.185	0.750
O1	0.440	0.452	0.330
O2	-0.120	0.240	0.616
W	-0.006	0.015	0.669
I	-0.005	0.115	0.960

The estimation process handles the variables' categories as dummies, which means that one category from every variable is used as a reference category. So, the coefficient estimates provided in Table 5.1 represent the direction and the level of the difference in probability of default between the underlying and reference categories.

The scorecard's evaluation analysis (Table 5.2) includes 3 different measures: AIC, BIC and the Gini coefficient. The most relevant measure from bank's perspective is the Gini coefficient because it describes the performance of the scorecard. According to the

Table 5.2: Evaluation analysis

Measure	Value
AIC	1384.3
BIC	1459.8
Gini	44.86

credit scoring framework, the Gini coefficient for the scorecard generally lies between 10% and 50%, which means that the performance of my scorecard (44.8%) is relatively high. Moreover, comparing with the bank's current qualitative (non-statistical) scorecard, whose Gini coefficient is equal to 20 %, the performance of the statistical model is approximately 2 times higher.

The last column of Table 5.1 (p-value) records the observed level of significance of the estimates. I use the p-value to drop statistically insignificant variables.

It is reasonable to exclude first the statistically insignificant variables I and W from the scorecard because they are based on application form. Moreover, these variables are unrestricted, which means that an applicant should put the value himself, which increases the risk of unintentional mistakes. I continue the iteration without variables I and W:

Table 5.3: Estimation results : Scorecard 2

Variable	Coefficient	Std. Err.	p-value
Intercept	-1.634	0.439	0.002
M1	0.367	0.198	0.064
M2	0.629	0.214	0.003
G1	0.338	0.153	0.027
Y	-0.030	0.008	0.000
R1	0.701	0.148	0.000
P1	-0.605	0.182	0.001
A1	-0.295	0.221	0.184
A2	0.056	0.184	0.761
O1	0.422	0.442	0.341
O2	-0.122	0.240	0.610

Comparing Scorecard 2 with Scorecard 1, Scorecard 2 has lower AIC and BIC values, and quite a similar Gini coefficient (Table 5.4), which means that Scorecard 2 is preferred to Scorecard 1. Still, I have variables with a large p-value, which are candidates for exclusion (Table 5.3). On the other hand, according to the bank's business perspective,

Table 5.4: Predictive power

Measure	Value
AIC	1380.5
BIC	1444.4
Gini	44.84

it is important to have at least 5 variables in the scorecard. The reason is that the bank would like to avoid the situation, where scorecard is too dependent on a particular variable. Also, if the scorecard has a limited number of variables it could lead to a clusters' problem.

Clusters are groups, which include observations with homogeneous characteristics. The idea is that these groups contain observations that are more similar to each other than to those in other groups. In case of the credit scoring, customers having exactly the same score provide a cluster.

The less variables I have within the scorecard, the higher the probability to have more customers with exactly the same score. This means that the size of a cluster is dependent on the number of the characteristics included into the scorecard. If the clusters are too big, it means that the scorecard is not flexible in controlling the trade off between the risk level and the number of applications approved. A minor change of the cut-off will decrease or increase approval rate rapidly. In other words, if I face a big portion of customers with the same score, a small change in a minimum acceptance score will decrease the number of approved customers greatly.

The last model includes 7 variables, which means that I'm going to exclude two additional variables. I would like to keep variables G, Y, P, R because these variables are from the credit report, which means that an applicant is not able intentionally or unintentionally to violate the information. The signs of the coefficient estimates for these variables seem logical and in line with bank's expectations.

Then, the possible candidates for removal are the application variables M, A and O. The signs of the coefficient estimates for variables O and A seem illogical. This can be explained by the natural correlation between these and other variables included in Scorecard 2. Moreover, the categories of A and O are not significant at 10 % risk level.

The categories of variable M are significant at 10 % risk level and their signs are in line with business expectations. Thus, I exclude variables A and O from the regression and produce a new iteration:

Table 5.5: Estimation results : Scorecard 3

Variable	Coefficient	Std. Err.	p-value
Intercept	-1.819	0.374	0.000
M1	0.383	0.197	0.052
M2	0.662	0.211	0.002
G1	0.341	0.153	0.026
Y	-0.028	0.008	0.000
R1	0.689	0.148	0.000
P1	-0.626	0.167	0.000

Table 5.6: Predictive power

Measure	Value
AIC	1377.5
BIC	1418.1
Gini	44.02

Based on Tables 5.5 and 5.6, I'm able to conclude that the final selection is Scorecard 3. There are several reasons behind this decision. First of all, the scorecard minimizes the values of both the BIC and AIC information criteria. The performance of Scorecard 3, measured by the Gini coefficient, is 1.8 % lower compared with Scorecard 2, but from the bank's perspective such a drop in the performance is acceptable. Next, all coefficient estimates for Scorecard 3 are significant at 10 % risk level, while Scorecard 2 contains insignificant ones too. And most important, in contrast to Scorecard 2, all signs of the coefficient estimates of Scorecard 3 are in line with logical expectations.

Scorecard 3 gives us an estimate of the probability of the default for a customer based on his characteristics. In other words, every customer's characteristic influences the customer's creditworthiness. The sign of the characteristic's coefficient gives us the direction of the influence. If the sign is positive, the characteristic increases the probability of the default, while the negative sign decreases it. The absolute value of the coefficient gives us understanding about the relative power of the characteristic compared with other characteristics.

Given the sample, the lowest score is equal to 0.008, which indicates that a customer with this score is best from the creditworthiness perspective with probability of the default equal to 0.8 percent. In contrast, the highest score is equal to 0.356, which means that this customer is worst from the creditworthiness perspective with probability of the default equal to 35.6 percent. The score of the typical customer is equal to 0.09, which means that probability of the default is about 9 percent. So, the worst customer has approximately

4 times higher probability of the default comparing with the typical customer.

5.3 Cluster overview

The purpose of this section is to provide a cluster overview of the final model, Scorecard 3. I introduce the cluster overview in Figure 5.1. This diagram presents the share of the observations (y-axis) from the particular score (x-axis) to the total sample. In other words, every point in a graph corresponds to a cluster, which includes customers with exactly the same score. On the x-axis, I see what is the customers' score for this particular cluster. On the y-axis, I see the share of the observations belonging this cluster to the total number of the observations.
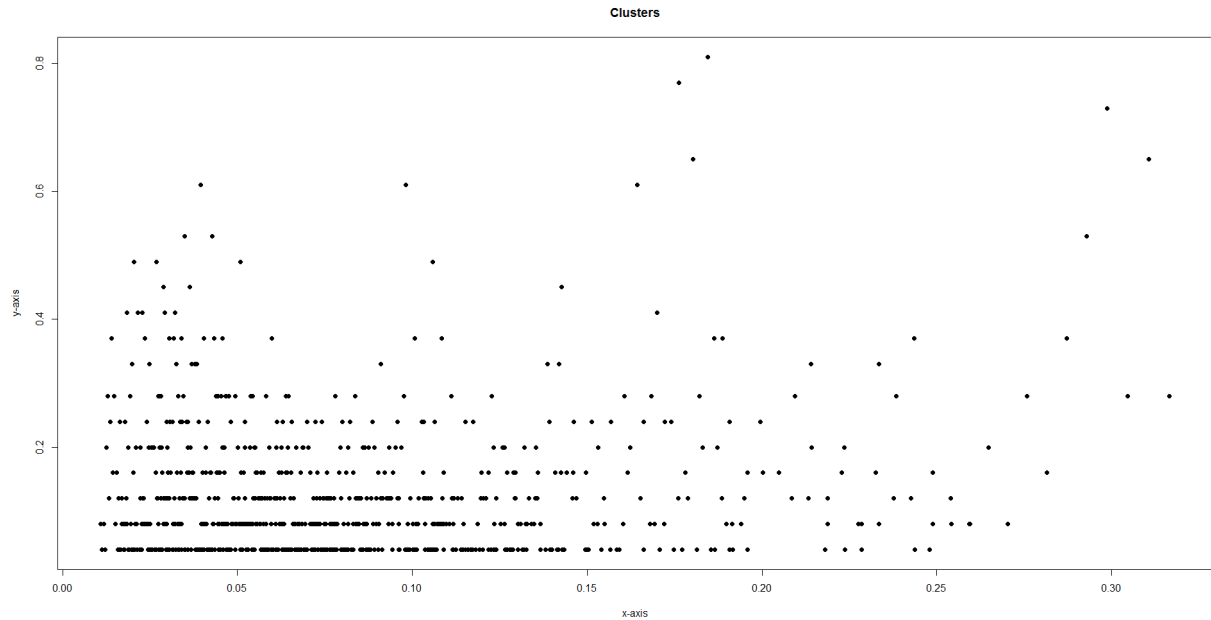

Figure 5.1: Clusters

The size of the acceptable cluster should be at maximum 2 % from the total sample. If the size of the cluster is more than 2 %, then an additional variable should be added to the scorecard. The size of the largest cluster is approximately 0.80 %, which is in line with the requirements.

5.4 Validation

The final model, Scorecard 3, should be validated on a non-estimation sample. The validation sample consists of 800 independent observations, where every observation is the bank's customer. I haven't used this sample for the estimation. The Gini coefficient for the validation sample is equal to 37.22 %, which is less than 44 % for the estimation sample. The drop of the performance is natural for the validation process. Still, the Gini coefficient for Scorecard 3 is much higher compared with the current scorecard's Gini (20 %). As well, I verify the scorecard by presenting the cluster diagnostics on the validation sample.
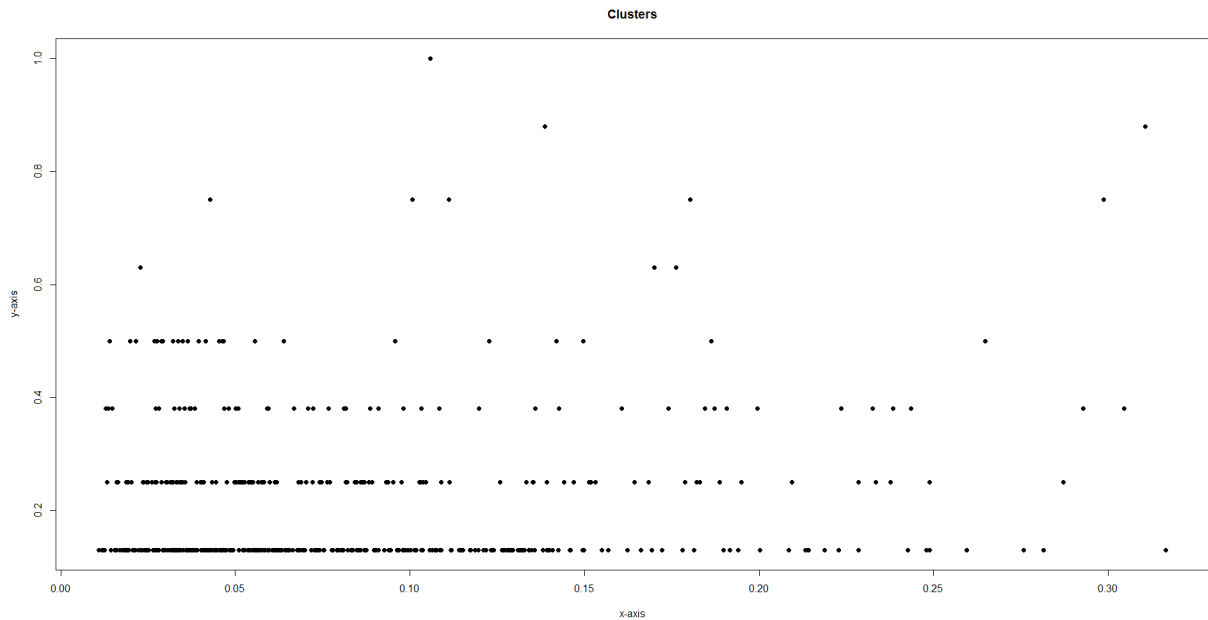

Figure 5.2: Clusters

From Figure 5.2 I see, that the biggest cluster share is 1 % from the validation sample. It is slighter higher than obtained with the estimation sample but still it is in line with the requirements.

5.5 Diagnostics

The following table presents diagnostics for the final model:

Table 5.7: Diagnostics

Measure	Value
McFadden R^2	0.085
Significance of the overall model	< 0.001

While no exact equivalent to the R^2 of the linear regression exists, the McFadden R^2 index can be used to assess the model fit. It is defined as

$$R^2_{McFadden} = 1 - \frac{L_1}{L_0}$$

where L_1 is the maximized log likelihood for a given model and L_0 for the null model containing only an intercept term. (McFadden 1973.)

The McFadden R^2 index is equal to 0.085 for the final scorecard. The value seems quite low compared with the value of R^2 usually obtained in linear regression models but it is rather common in logistic regressions. The value of $R^2_{McFadden}$ is higher than 0, which gives us information that the model fit the data. The value of $R^2_{McFadden}$ as such is not informative in case of logistic regression.

An alternative measure of model fit is the significance of the overall model. This measure asks whether the model with predictors fits significantly better than a model with just an intercept (i.e., a null model). In case of Scorecard 3, the p-value associated with the test is less than 0.001, which tells us that our model as a whole fits better statistically than an null model.

Chapter 6

Conclusion

The purpose of this work was to review credit scoring and its applications both theoretically and empirically, and to end up with the best combination of the variables for the default risk forecasting.

The first part of the thesis focused on theoretical aspects of credit scoring - statistical method for the scorecard estimation and measuring its performance. Firstly, I explained the definition of scorecard and underlying terminology. Then I reviewed the general approaches for scorecard estimation and demonstrated that logistic regression is the most appropriate approach with all assumptions and statistical inference included. Next, I described methods for measuring scorecard performance and showed that scoring systems would be ranked in the same order of discriminatory power regardless the measure used.

The goal of the second part was empirical analysis, where I applied the theoretical background discussed in the first part of the master thesis to the dataset from a consumer credit bank, which includes variables presented in the application form, from credit bureau data and extracted from social security number.

The estimation results showed that variables, which are from credit bureau or extracted from social security number are the most informative ones, having most significant coefficient estimates with the observed level of significance less than 3 %. I assume, this is due to the fact that a customer is not able to affect these variables.

In contrast, the variables customer could affect, which are from the application form, have a poor quality. The variables with unrestricted responses are even less informative compared with the restricted ones, which could be explained by unintentional violation of the information, like typos or misunderstanding of the application form.

The major finding of the master thesis is that comparison of the performance between a non-statistical model, based on rational expectations and managers' experience, and the estimated statistical model, shows that the statistical model performs much better. This means that banks and financial institutions should benefit from the introduction of the

statistical approach, employed in the thesis.

Bibliography

[1] A. Agresti (2007): *An introduction to categorical data analysis*, Wiley, New Jersey.

[2] A. Agresti (2013): *Categorical data analysis*, Wiley, New Jersey.

[3] A. Agresti (2015): *Foundations of linear and generalized models*, Wiley, New Jersey.

[4] H. Akaike (1973): *Information theory and an extension of the maximum likelihood principle*, Symposium on Information Theory, Budapest.

[5] N. Capon (1982): Credit scoring systems: A critical analysis, *Journal of Marketing*, 46, 82-91.

[6] G.G. Chandler and D.C. Ewert (1976): *Discrimination on Basis of Sex and the Equal Credit Opportunity Act*, Credit Research Center, Purdue University, West Lafayette.

[7] D.R. Cox and D.V. Hinkley (1974): *Theoretical Statistics*, Chapman & Hall, London.

[8] V.S. Desai, J.N. Crook and G.A. Overstreet (1996): A comparison of neural networks and linear scoring models in the credit environment, *European journal of operational research*, 95, 24-37.

[9] R.E. Kass and L. Wasserman (1995): A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion, *Journal of the American Statistical Association*, 90, 928-934.

[10] E.M. Lewis (1992): *An introduction to Credit Scoring*, Athena Press, San Rafael.

[11] D. McFadden (1973): Conditional Logit Analysis of Qualitative Choice Behavior, *Frontiers in Econometrics*, 105-142.

[12] J.H. Myers and E.W. Forgy (1963): The development of numerical credit evaluation systems, *Journal of the American Statistical Association*, 58, 799-806.

[13] J.R. Nevin and G.A. Churchill (1979): The Equal Credit Opportunity Act: An evaluation, *Journal of Marketing*, 42, 95-104.

[14] J. Saunders (1985): This is credit scoring, *Credit Management*, September, 23-26.

[15] N. Siddiqi (2006): *Credit risk scorecards: developing and implementing intelligent credit scoring*, Wiley, New Jersey.

[16] L.C. Thomas (2002): *Credit scoring and Its Applications*, SIAM Monographs on mathematical modeling and computation, Philadelphia.

[17] L.C. Thomas, D.B. Edelman and J.N. Crook (2004): *Readings in credit scoring*, Oxford University Press, New York.