Aalto University
School of Economics

The Determinants of Default in Consumer Credit Market

Finance

Master's thesis

Marjo Hörkkö

2010

Department of Accounting and Finance
Aalto University
School of Economics

Aalto University
School of Economics
Master's Thesis
Marjo Hörkkö

Abstract
June 4, 2010

THE DETERMINANTS OF DEFAULT IN CONSUMER CREDIT MARKET

PURPOSE OF THE STUDY

This thesis uses empirical observations on consumer credit behavior to study the determinants of default in Finland. The main objective is to investigate if both socio-demographical and behavioral variables have effect on default. In the thesis I construct three different models to show which variables have predictive power the most. The models are compared in terms of efficiency and power to discriminate between low and high risk customers. The purpose of this study is also to provide practical information for credit companies to create more up-to-date and reliable credit scoring models. I also illustrate how such a model can be constructed to achieve the strategic objectives of the credit institution.

DATA & METHODOLOGY

The data set of this paper is from an anonymous consumer credit company who offers loans to retail customers. I have 14 595 observations of customers of which 29% turned out to default their loan. All the applications were received between May 2008 and September 2009 and the default information was captured in December 2009. Out of 30 explanatory variables 23 were socio-demographical and the rest, 7 were behavioral. There are several unique and important features of this data set that enables me to test the impact of both socio-demographical and behavioral variables. The analyses are performed using logistic regression, forward and backward stepwise analysis and several tests with SPSS program.

RESULTS

The main findings are that both socio-demographical and behavioral variables have a notable effect on default. Consistent with previous literature the most significant socio-demographical variables are income, time since last moving, age, possession of credit card, education and nationality. Some behavioral variables seemed to have even more predictive power. Those are the amount of scores the customer obtained, loan size and the information if customer has been granted a loan earlier from the same company. Interestingly, the results have variation to some extent when excluding few of the variables outside the model. The predictive power of all three models is adequate and thus can be employed as a reliable credit scoring model for the credit institutions.

KEYWORDS

default, failure, consumer credit, consumer loan, credit scoring model, consumer credit market, socio-demographical, behavioral, relationship

Aalto Yliopisto
Kauppakorkeakoulu
Pro Gradu -tutkielma
Marjo Hörkkö

Tiivistelmä
4.6.2010

ASIAKKAIDEN MAKSUKYKYYN VAIKUTTAVAT TEKIJÄT KULUTUSLUOTTOMARKKINOILLA

TUTKIELMAN TAVOITTEET

Tämä pro gradu – työ tutkii kuluttajien maksukykyyn vaikuttavia tekijöitä suomalaisilla kulutusluottomarkkinoilla. Päätavoitteena on tutkia onko sekä sosio-demograafisilla, että asiakassuhteeseen liittyvillä muuttujilla vaikutuksia luoton laiminlyönnissä. Tässä työssä muodostan kolme eri mallia osoittaakseni millä muuttujilla on eniten ennustuskykyä. Mallien hyvyyttä vertaillaan niiden tehokkuudella ja ennustuskyvyllä erottaa matala- ja korkeariskiset asiakkaan toisistaan. Tämän pro gradu – tutkielman tavoitteena on myös antaa käytännön hyötyä ja uutta informaatiota kulutusluottoa tarjoaville yrityksille päivitetyn ja luotettavan credit scoring – mallin rakentamisessa.

LÄHDEAINEISTO

Käytän tutkimuksessa suomalaisen anonyymin kulutusluottoa tarjoavan yrityksen aineistoa. Minulla on havaintoja 14 595 kuluttajasta, joista 29 % jätti maksamatta luottonsa takaisin. Kaikki hakemukset rekisteröitiin toukokuun 2008 ja syyskuun 2009 välillä, ja takaisinmaksuinformaatio otettiin ulos joulukuussa 2009. Tutkimuksessa käytettiin 30 selittävää muuttujaa, joista 23 oli sosio-demograafisia ja 7 asiakassuhteeseen liittyviä muuttujia. Aineistossa on useita tärkeitä ominaisuuksia, jotka edesauttavat selvittämään sekä sosio-demograafisten, että asiakassuhteeseen liittyvien tekijöiden vaikutuksia maksukyvyn ennustamisessa. Analyysi toteutettiin logistisen regressioanalyysin avulla SPSS -ohjelmalla.

TULOKSET

Tulosten mukaan sekä sosio-demograafisilla, että asiakassuhteeseen liittyvillä muuttujilla on vaikutusta maksuvaikeuksiin kulutusluottomarkkinoilla. Tutkimuksessa esille tulleet merkittävimmät muuttujat vastasivat aiempaa kirjallisuutta ja ne olivat: tulot, aika edellisestä muutosta, ikä, luottokortin omistaminen, koulutustaso ja kansallisuus. Jotkin asiakassuhteeseen liittyvät muuttujat olivat vieläkin merkitsevämpiä. Näitä olivat luottoyhtiön antamien pisteiden määrä, lainan koko ja tieto siitä, onko asiakkaalla ollut yrityksestä aiemmin luottoa. Tuloksissa oli jonkin verran eroavaisuuksia kun osa merkityksettömimmistä tai tärkeimmistä muuttujista poistettiin mallista. Kaikkien kolmen mallin selittämiskyky on riittävä muodostaakseen luotettavan credit scoring –mallin luottoyhtiöille.

AVAINSANAT

maksukyky, maksuhäiriö, kulutusluotto, credit scoring – malli, kulutusluottomarkkinat, sosio-demograafinen, käyttäytyminen, asiakassuhde

Table of contents

List of tables and figures

1 Introduction

The introduction section familiarizes the reader with the topic of this thesis and gives an overview of the main issues, which will be covered in the following chapters. Firstly, I will explain the background and motivation to this study. Secondly, I will introduce the objective and the main findings as well as the contribution to existing literature and to the industry. Thirdly, I will mention limitations regarding to this topic. At the end of the introduction chapter, I will briefly explain the structure of the rest of the paper.

1.1 Motivation to the study

Consumer credit and default prediction have been studied relatively little - if at all - in Finland. We have several companies who offer consumer credit or small loans. No wonder, consumer credit has become more popular than ever (e.g. Brown et al., 2005). Sudden change in income level, unemployment and other unexpected occasions are reasons[1] to apply for a consumer loan to maintain the consumption at the same level. There has also been intense conversation about the nature and morality of consumer credit due to the high costs related to it. The real annual interest rates can reach up to 300%[2] but which are nowadays more transparent due to actions taken by Finnish Consumer Agency[3]. This may come as a surprise for some customers who are not familiar with the terms and conditions of consumer credit and might thus lead to increased level of insolvency, payment troubles and default. Brown et al. (2005) document that unsecured debt is associated with an increased level of psychological distress when compared to secured loans like mortgages, due to the loans' surprisingly high levels of interest.

The need of consumer credit today is at it's highest, but at the same time the default rates have risen and from the banks' perspective the riskiness of these loans is usually higher than that of a regular bank loan. As Kočenda and Vojtek (2009) show as much as 50% of the

[1] Compare to the use of small instant loan (Autio et al., 2009): typical purposes are buying alcohol, cigarettes, partying, buying food and repaying credit or interest.
[2] See www.lainatieto.fi/kulutusluotot.
[3] See Finnish Consumer Agency (Kuluttajavirasto in Finnish): www.kuluttajavirasto.fi.

granted loans they analyzed defaulted. For the lending institution such a default rate affects to its financial performance significantly. The phenomenon of consumer credit has shown rapid growth over the last years also in Finland. The total amount of consumer debt today[4] is 13,6 billion Euros showing a growth of 4,3% in comparison with last year's equivalent and 8% between the years 2007 and 2008[5]. At the same time more than 7% of Finnish capita had defaulted in 2009 (Suomen Asiakastieto[6], 2009). In the light of these numbers I can conclude that studying the default predictability is particularly important.

Credit risk measurement has evolved dramatically over the last 20 years in response to a number of secular forces that have made its measurement more important than ever before. According to Altman and Saunders (1997) these forces have been: a worldwide structural increase in the number of bankruptcies, a trend towards disintermediation by the highest quality and largest borrowers, more competitive margins on loans, a declining value of real assets (and thus collateral) in many markets and a dramatic growth of off-balance-sheet instruments with inherent default risk exposure. After launching the Basel II framework banks have started to upgrade their credit risk management approaches (Claenssens et al., 2005) and vendors have started to offer more and more improved models to banks for calculating their regulatory capital requirements. Especially in the consumer credit market no securities are needed when applying for a loan. Due to the nature of small loans there is a great amount of asymmetric information i.e. the lender has a risk of a customer defaulting the loan. Basel II framework was built based primarily for large commercial credits including credit card loans, mortgage loans, home equity lines of credit, auto loans, and other consumer loans. The implementation of Basel II was mainly due to retail lenders' great reliance on statistical models only. Regardless of the negative acceptance of Basel II the banks and other credit institutions have worked out and improved their risk management. Bofondi and Lotti (2006) state that the diffusion of credit scoring is likely to be boosted by the introduction of the New Basel Capital Accord[7], which encourages improvements in banks' risk assessment capabilities by closely linking capital requirements with portfolios' risk level.

[4] Based on statistics 30.9.2009.
[5] See Statistics Finland, 2010 and Federation of Finnish Financial Services, 2010.
[6] A private held credit bureau that keeps record of both Finnish corporate and private credit defaults.
[7] See Basel Committee on Banking Supervision (2001) for details.

The loan granting decision is carried out by banks and other credit institutions. Traditional methods of deciding whether to grant loan to an individual are based on human judgment and experience of previous decisions. However, to consider every small loan as a separate loan is time consuming and expensive. Usually the lender doesn't have information about the solvency or credit behavior of a new potential customer and especially in consumer credit business customers are often persons who are applying for a loan for the first time. Thus, to determinate the customer's expected probability of default the lender must estimate his ability to pay back from his current characteristics, as default can only be observed afterwards. To evaluate the customers' solvency banks often use behavioral and demographical characteristics as predictors of default. The most common variables are often income, age and education[8]. Also determinants that characterize the relationship between the lender and the customer, like the amount of resources and length of the relationship, are seen to have a clear connection with default.

Allen et al. (2004) notes that the trend in retail credit decision-making is strongly toward increased reliance on statistical, databased models of credit risk measurement. Retail lending has gradually shifted from relationship lending to transactional (portfolio-based) lending. To measure the level of risk managers in banks and credit companies use loan default predicting models or credit scoring models (CSMs), which tend to be the easiest and most common methods to utilize. CSM is an analytic technique, which combines the current and historical information of the customer to make predictions whether the customer will repay the debt. In CSM customers are given points by their socio-economic features and behavior and thus their default probabilities are estimated based on the default behavior of previous customers who have either paid their loan full or defaulted. After all characteristics are given points and the managers have decided the cut-off value the new customers are either accepted or rejected based on their total points. The goal of CSM is to predict default in order to make a rational decision in approving or rejecting a new loan application and to apply a suitable pricing policy. The CSM should optimize the likelihood of bad obligor being rejected and the good one being accepted. Not being able to optimize this can lead to underpricing the bad loans and overpricing the good loans.

[8] See Table 1.

Being able to define which characteristics are those that affect default and picking up the customers who perform well is relatively difficult. Warren (2002) shows that most of the people who file for bankruptcy in the US come from a middle-class family. She emphasizes that only 30% of the defaulted Americans in her sample were from the lowest income quintile and the rest were the so-called "nearby neighbors", as she illuminates. The characteristics are thus not obvious in a sense and cannot be estimated and scored based on pure intuition. The performance of the credit company depends on how successful it is in predicting customer default based on behavioral and demographic characteristics of the customer. From a lenders perspective it is highly important to study the determinants of default in order to minimize the credit losses. The difficulty of constructing a suitable model can be pointed out with an example of actual default rates: the dataset in question is from a company who received 103 037 applications during the observation period[9], accepted only 14% of those based on a CSM of good quality and yet faces a default rate of 29%.

It can be seen that it is also in the interest of customers who are not granted the loan they couldn't afford and thus ending up to a national default register (Suomen Asiakastieto). The trustworthiness of the whole industry is also partly based on the evaluation systems of its actors. Thus, the core of this thesis is not only to investigate the socio-demographical and behavioral determinants that have an effect on the customers' ability to pay but also to help lenders consider their scoring models more carefully.

1.2 Objective

While the improving of the prediction accuracy and comparison of different methods has been the prime mover of bankruptcy and default prediction studies, this study focuses on analyzing the predictive power of variables. The objective of this thesis is to study the determinants of default; which behavioral and socio-demographical variables have effect on default, how important are they and how do the results change when I exclude some of the irrelevant or most significant variables to create a new model. This paper concentrates especially on socio-demographical variables as determinants of default. By constructing three different logistic models employing a large dataset I am able to provide reliable results and proposals for a new

[9] Observation period was May 27th, 2008 to September 1st, 2009.

CSM for the company. A logistic regression model is used to develop a numerical scoring system for consumer credit.

This thesis answers the following research questions:

1. Can both socio-demographical and behavioral variables predict default behavior?

2. Which characteristics are to be used in the scoring model as variables that can discriminate between a "good" and a "bad" loan?

3. How to obtain the score for each characteristic?

In addition to these three main questions this thesis compares the results to previous studies and gives attention to both socio-demographical and behavioral variables in creating practical CSM. This paper answers to the company's needs to improve its scoring model and provides practical and up-to-date information.

The dataset employed in this study consists of 14 595 observations, of which 4 191 were defaulted or "bad" loans and the rest, 10 404 were non-defaulted or "good" loans. The unique data is provided by one of the largest consumer credit companies in Finland that wishes to stay anonymous for the thesis.

The initial sample consisted of 31 variables of which 30 were employed to the analysis. In this thesis I use parametric logistic regression, which has given reliable results (Arminger et al., 1997 and Hand & Henley, 1997) in creating CSMs. In addition, I construct two other models that focus on the drawbacks of the initial model.

The objective of this study is both to be one of the first studies accomplished in Finland in the area of consumer credit and to evaluate different alternatives to the traditional scoring system the company in question uses.

1.3 Results

The main findings are that both socio-demographical and behavioral variables have an effect on default. Consistent with previous literature the most significant socio-demographical variables are income, time since last moving, age, possession of credit card, education and nationality. Some behavioral variables seemed to have more predictive power than others. Those are the amount of scores the customer obtained, loan size and the information whether the customer had been granted a loan earlier from the same company. Interestingly, the results have variation to some extent when excluding few of the variables outside the initial model. The predictive power of all three models is adequate and thus each of them can be employed as a reliable credit scoring model for the credit institutions.

1.4 Contribution to the existing literature

This thesis contributes to existing literature in various ways and with a two different point of view: from the perspective of research and from the perspective of the whole credit industry. The following two sections describe the benefits for both of them.

1.4.1 Contribution to the industry

This study gains added value by using data from an older and stable EU country with more matured markets (compare to Kočenda and Vojtek, 2009). By the means of the new information and results it can also be adapted to European and Nordic countries, in which no similar analyzes have been produced in the area of consumer credit before.

Credit companies often buy the CSM they use outside. The model is usually very expensive and above all, the information goes out-of-date because of macro-economical changes, possible recessions and general economic conditions. It would be important for the credit companies to use an up-to-date scoring model that has been conducted with real historical customer data from the same nationality and population. This paper provides topical information of socio-demographical variables that affect default. By utilizing the models I build, the companies are able to minimize the default rate that could be caused by asymmetric

information. Jaffee and Russel (1976) as well as Stiglitz and Weiss (1981) studied the traditional loan market from the perspective of asymmetric information and adverse selection and found them to have significant negative impact on default rates. This paper provides important information for credit companies encouraging them to take the socio-demographic variables into more precise analysis.

1.4.2 Contribution to the research

This thesis contributes to the existing default literature in the following ways. First, due to the fact that no similar documents have been written on the Finnish markets in the area of CSM, I am able to provide interesting results with sensitive dataset. Consumer credit and default are relatively new areas of research in Finland and also in the Nordic countries. However, in the United States, CSM and default have been studied to some extent, or at least more widely than in Europe of Nordic. This is mainly due to the more matured consumer credit market in U.S, the availability of sensitive data and the size of the customer base. The data and information required for these kinds of studies are difficult to obtain.

Second, this study is comprehensive and broad with a large dataset: 14 595 observations and 31 variables, which is more than many of the previous studies, have been used[10]. The empirical analysis is also very detailed: I study default predictors with three different models as most of the studies focus only on one.

Thirdly, most of the studies have focused on finding the best possible technique to build a CSM. However, no major differences have been found and using logistic regression for example has proven to give just as reliable results as the other techniques such linear discriminant analysis, neural networks or CART analysis. In my opinion it is more important to study the significance and predictive power of different variables rather than the techniques themselves. Therefore the issue of variable selection is a crucial and challenging problem to solve before different credit scoring techniques are used to develop the best performing model. Hence, to provide reliable results the number of input variables has to be adequate.

[10] See Table 1 for details.

1.5 Limitations

The limitation of the study is the data that contains default information only from the first year. Default information used in this study has been captured on December 15th, 2009. The customers can choose the repayment period to be as long as 4 years, which means that default can also occur later on. To be able to have default rates from the full period this analysis should be remodeled when all the loans have expired. However, having a significantly large amount of data, I am able to draw reliable conclusions. In addition, I can make the assumption that customers who will not default during the first year are considered "good" ones.

As the data includes only the observations where customers were granted credit, there is a sample selection bias when not taking all the applications in to examination. However, this is common in the literature. It has been studied (Banasik et al., 2003) that the difference between rejected and accepted customers is small and thus has no large effect when analyzing the characteristics of customers. In addition, several variables could not be used as an explanatory variable because no data on these variables is available for rejected applicants.

1.6 The structure of the study

The first chapter of the research introduced the topic, the research objectives and background. The rest of this paper is divided into five sections. In the second chapter I describe the consumer credit market in practice and summarize the previous studies. Chapter three introduces the data used in the paper and presents the variables with descriptive statistics. The fourth chapter begins the empirical part of the study by justifying the choices of techniques and presenting the methods and tests. Chapter 5 presents the findings of the research, including the empirical results, as well as their analysis and interpretation. The last chapter discusses the main implications and concludes. It also gives suggestions for managerial implementation and for further research in the area of default and consumer credit.

2 Theoretical and practical background for consumer credit, CSM and default

This chapter presents an overview of the main issues related to consumer credit. The concept of consumer credit, application process and the nature of Finnish consumer credit markets will be covered in this chapter. Also the previous literature related to determinants of default, consumer loan markets and CSM are discussed here.

2.1 General information about consumer credit

This section familiarized the reader with the concept of consumer credit and the process of applying loan. It also presents the idea of constructing a credit scoring model.

2.1.1 Special features of consumer credit

The concept of consumer credit is broad and in a sense unclear. In general, consumer credit is granted to finance the purchase of commodities and services. Financing of car, home appliance, traveling and furniture is often understood as consumer credit. Instant loans are sometimes ambiguously understood as consumer credit. However, in this study they are treated separately and the division is made based on the amount, maturity and the application process.

The application for consumer credits requires, unlike the one for instant loans, a bank account and more information of the customer. Some companies that offer instant loans accept an application sent via mobile phone. Consumer credit cannot be obtained based on an SMS application but requires registration to the lender's web pages or a personal phone call. The instant loans are often smaller, amounting up to a maximum of 1 000 euros and have a maturity of few months, while in consumer loans the loan period can usually be as long as four years[11]. The process of applying for a consumer loan is, however, made easier than

[11] See www.kulutusluotto.org for details (service available only in Finnish).

applying for example a mortgage. The concept of consumer credit is more commercial and more available, making it possible for customers to obtain a loan outside office hours and without collateral. This flexible nature of consumer credit reflects to the higher interest rates. Whereas mortgages and other traditional loans involve fixed amounts and payment schedules, in consumer credit market the customer has extensive authority on deciding the debt repayment with the minimum monthly repayment being a fixed percentage of the total balance.

Consumer loans are granted by banks, financial and credit institutions, credit card companies, commercial stores and mail-order firms.

2.1.2 The application process

The process of applying consumer credit is quite straightforward. First, the applicant logs in to the lenders web page with his bank username and password so that the lender can identify the applicant's identity. Through logging in the lender is also able to have the applicant's social security number in order to define if the applicant has credit standing in the Finnish credit register or whether the applicant has had credit before and has default notification in the company's own database. Also the applicant's address can be confirmed from the Finnish Population Register[12]. Once the applicant has logged in, he fills in an application online. The application includes several questions about the applicant's identity, which are further treated in this study as variables. The service is available 24/7 and the credit decision can be given instantly. If applicant is granted the credit he will have the money on his account in few minutes[13]. No collateral or guarantor is needed. To be able to continue he must agree on the terms and conditions and thus promise to provide true information. Customers are also able to apply for a loan by phone, which is less common nowadays. The credit is a type of annuity loan where the customer can define the maturity, however, not exceeding the maximum of four years.

[12] See www.vaestorekisterikeskus.fi for details.
[13] In accordance with the new Finnish Law of Consumer Protection (Kuluttajansuojalaki in Finnish) the credit companies are not allowed to transfer any money between 11pm and 7am since the first of February 2010.

2.1.3 The concept of credit scoring model

Consumer loans are relatively small and granted to unrated borrowers. Therefore it is not usually cost effective to evaluate each loan on an individual loan-by-loan basis. The small size of each consumer credit implies that the absolute size of the credit risk of a one loan is minimal. Due to economies of scale associated with information gathering, risk management and loan monitoring, limited resources are devoted to analyzing the risk for an individual loan. Hence, lenders typically rely on scoring models and automation for approving loans. Mester (1997) documents that 97 percent of banks use CSM to approve credit card applications, whereas 70 percent of the banks use CSM in their small business lending.

Although the first credit scoring system implemented for banks and mail order firms occurred already in the fifties in the U.S, in housing finance the turning point was not until in the 1990s with the growth of automated statistical credit and mortgage scoring as a method for underwriting and approving loans (Straka[14], 2000). Automated underwriting was previously used in credit card and auto lending but after 1995 also mortgage business and consumer credit started to benefit from it. The oldest and most commonly used traditional scoring model was the multiple discriminant credit scoring analysis for companies pioneered by Altman (1968). Since then also other techniques have been employed widely. For example Allen et al. (2004) summarize the four suitable methods to create a CSM; 1) the linear probability model, 2) the logit model, 3) the probit model, and 4) the multiple discriminant analysis model. All of these models identify financial variables that have statistical explanatory power in differentiating defaulting firms from non-defaulting firms.

The objective of such models is to minimize the credit risk and default rates and to prevent granting loan to ”bad” customers and to avoid giving false rejection to ”good” customers. Scoring models use historical data combined with a statistical technique to identify which customer characteristics such as age, income and marital status are the ones that distinguish between customers who default and those who perform well. Credit score is not a percentage

[14] Straka provides a comprehensive study of moving to automated credit evaluations in 1990s.

nor is there an amount presenting the cut-off value for proper scoring. Each credit bureau, bank and other lending institution can determine its own CSM being used.

The modeling of CSM is not definite and for example Basel II does not impose any standards on the process. A lender can purchase the model or construct one itself. In general, the modeling is based on historical information. Creditors can construct the classification rules based on the data of the previous accepted and rejected applicants. First, the old customers are divided into two groups: those who defaulted the loan and those who did not. Second, their socio-demographic and behavioral characteristics are evaluated with the help of empirical modeling. Information such as income or age can be kept as continuous variable but most often is transformed into categorical value. After deciding suitable thresholds each variable or category is given scores. Every new customer is evaluated based on these subscores and the summed score value is compared to the cut-off value. The managers need to determine a suitable cut-off value to correspond their business and risk management. The value indicates how much risk they can adopt and what their presumption of the default rate is. If a customer is given more points than the fixed cut-off value he is admitted credit.

2.2 The nature of Finnish consumer credit market

The importance of the study may be illustrated by the expenditure of consumer loans. Finnish consumers have become more open to the use of credit. Federation of Finnish Financial Services[15] (2010) reports that the amount of consumer credit in euros rose 4,3% from 2008 to 2009[16] totaling up to 13,6 billion euros (see Figure 1). During November 2009 and October 2009 new consumer credit was granted worth 222 and 305 million euros, respectively. The decrease is explained by the seasonality of applications; new consumer loan is applied mostly from spring to autumn. The proportion of consumer credit from the total household debt was 13% in the end of November 2009.

[15] Finanssialan Keskusliitto in Finnish.
[16] Granted by financial institutions (here: credits from banks, credit card companies and other financial institutions). Observation period is 30.11.2008-30.11.2009.

According to a survey conducted by Federation of Finnish Financial Services (2010) as much as 28% of 18-74 year olds had consumer credit. The mean loan amount in consumer credit market is between 1000 to 4000 euros with a repayment period from one month to four years.

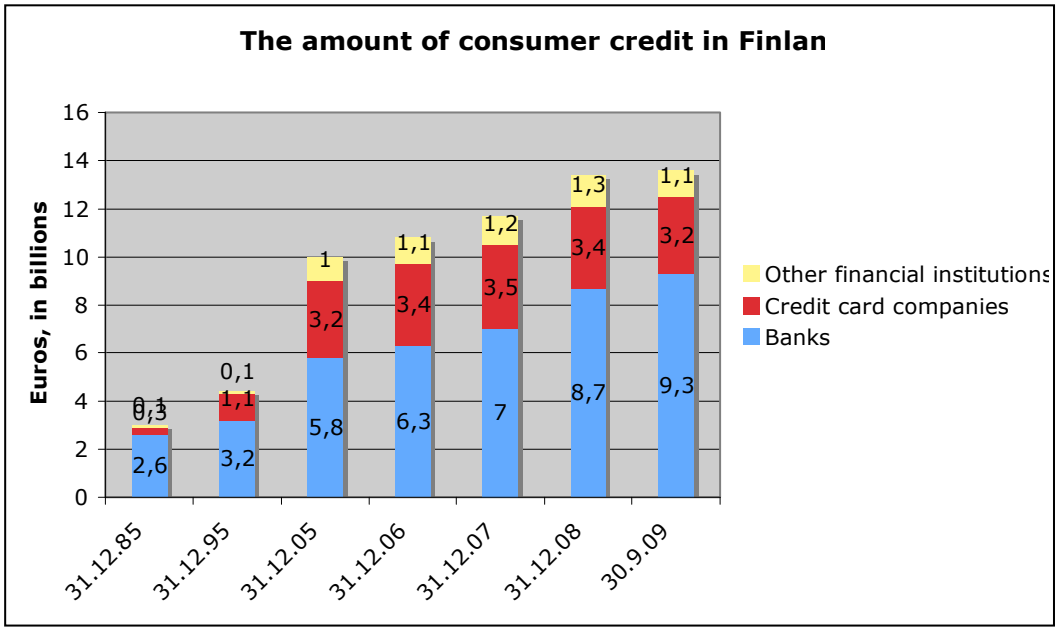

Figure 1: The amount of consumer credit in Finland

Source: Finanssivalvonta (2010).
The total amount of consumer credit has increased continuously. At the time of the study the year-end information for 2009 was not published but it can be seen that the amount at the end of September 2009, anyhow, was higher (13,6 billion Euros versus 13,4 billion Euros) compared to the year before. The amount of loan granted by credit card companies and other financial institutions has remained quite the same while banks have increased their market share constantly.

Suomen Asiakastieto reports that in the end of December 2009 305 000[17] private persons[18] had defaulted. This corresponds to 4,4% more than in 2008. The amount of new payment troubles was even larger, amounting up to 645 000 defaults[19] during 2009, which is 18% more than a year earlier (547 000) and twice the amount of the year 2000 (see Figure 2).

[17] Includes often several notes to one person. On average one person had six defaults at the same time.
[18] Persons over the age of 18.
[19] Including both retail market and companies.

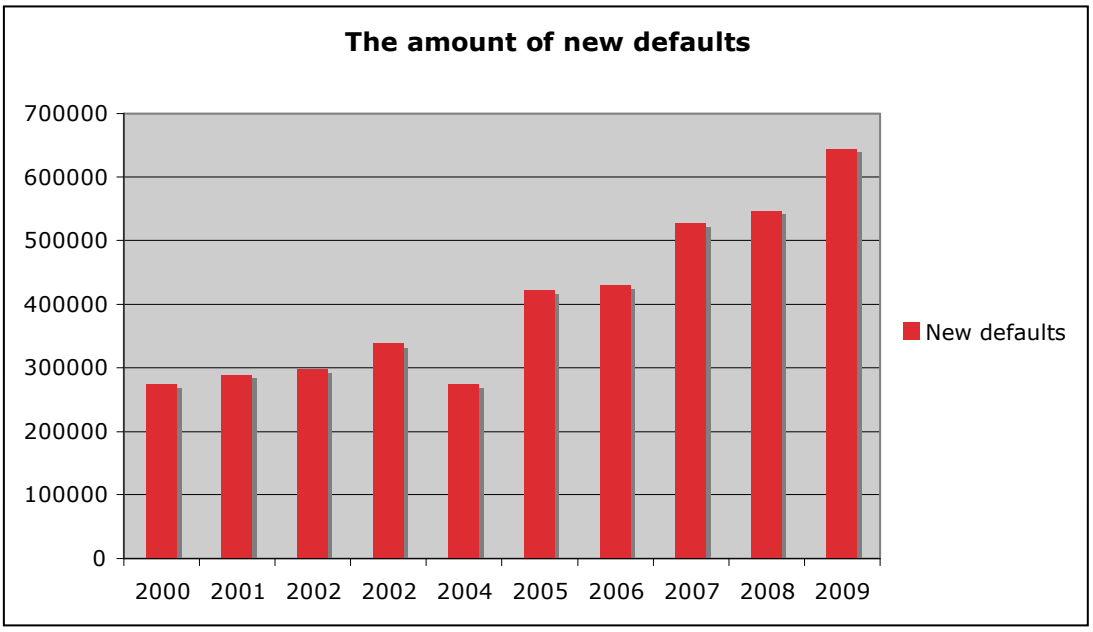

Figure 2: New defaults

Source: Suomen Asiakastieto (2010).
The amount of new defaults in Suomen Asiakastieto's register. It is common that most of these defaults will mass to same persons. On average a person in the register has six defaults and only every sixth has only one default. The amount of new payment troubles increased with 100 000 defaults (18%) from year 2008. On average two third of defaulted are men.

The interest rates differ significantly depending on the lender. Collateralized consumer credit granted by banks had an average interest rate of 7,82% in the end of November 2009. Other credit companies have a broad spectrum of interest, fluctuating between 7 and 15%. The most expensive consumer loans are the ones granted by mail-order companies having an average interest of 20 to 30%. Worth noticing is the fact that the real annual interest rate can reach up to hundreds of percentages.

2.3 Related studies

Consumer credit markets have been studied relatively little due to the confidential nature of the customer data and the difficulty of measuring the risk appropriately. Most of the studies conduct U.S data but the literature and research is evolving also in Asia. The evidence from Finnish and even European markets is minimal.

However, the literature of credit scoring systems and default rate approximation started already when the first consumer loans were granted. To study default risk is extremely important due to the automation of decision-making process and the easiness of applying for the loan. According to Straka's (2000) study of automated credit evaluations, development of CSMs has proven to reduce defaults.

In this chapter I will go through the most relevant early literature divided into four different markets facing default: corporate loan markets, credit card industry, mortgage markets and consumer credit markets as itself. Table 1 in section 2.4 shows the most employed variables of some of the studies presented next.

There are several studies that concentrate on comparing different techniques to create CSM. Those are not discussed comprehensively here but are covered in Chapter five.

2.3.1 Default in corporate loan markets

Quantitative CSMs were developed for consumer credit purposes much later than those for corporate credit mainly due to problem of availability of data. In many countries legal and other reasons prevented the buildup of publicly available databases. Data were limited to the own databases of financial institutions. Nowadays, the data on personal loans is still highly delicate but information on corporate defaults is often publicly available to help institutions and researchers to develop quantitative CSMs. Most of the credit risk literature (Altman, 1968, Neophytou & Charitou, 2000, Carvalho & Dermine, 2003 and Altman et al., 2007) deals with corporate loans where it is possibly to define the size, asset turnover, solvency, leverage and other historical key ratios of the company and construct a reliable CSM based on historical performance of the companies. In consumer markets the case, however, is more difficult. In addition to the information the credit bureaus offer the lenders have to trust the information a customer gives in the application.

Laitinen and Kankaanpää (1999) are one of the only Finnish authors who have discussed the default behavior. They assessed six alternative methods[20] (LDA, LR, RPA, survival analysis,

[20] See section 4.1 for details.

NN and HIP) that have been applied to financial failure prediction. The main objective was to study whether the results stemming from the use of alternative methods differ from each other. They used only three financial ratios (total debt to total assets, the ratio of cash to current liabilities and the operating income to total assets) due to methodological issues. The results of 76 randomly selected Finnish small and medium sized failed firms indicate that no superior method has been found but the predictive power of logistic analysis was best resulting a 89,5% prognostic accuracy. Laitinen continued the work with Laitinen (2000) by testing whether Taylor's series expansion can be used to solve the problem associated with the functional form of bankruptcy prediction models. To avoid the problems associated with the normality of variables, the logistic model to describe the insolvency risk was applied. Several financial ratios were employed with estimation sample including 400 firms and the results suggest that the cash to total assets, cash flow to total assets, and shareholder's equity to total assets ratios operationalize the factors affecting the insolvency risk. The usefulness of Taylor's model in bankruptcy prediction was evaluated applying the logistic regression model to the data from the Compustat database.

Allonen (2010) analyzed the key ratios of companies similar to Laitinen and Laitinen (2000), with logistic regression to make assumptions about the determinants of default in business. His master's thesis confirmed the findings of earlier literature: the insolvencies of small and medium-sized companies are able to predict with a rather high level of confidence using logistic regression and employing financial ratios (describing profitability, indebtedness, liquidity and operational magnitude). He employed 1 094 Estonian companies' default information from 2001 to 2009 taking into consideration the economic recession and its consequences to default prediction. The results suggest that the predictive power of the model weakened slightly at the time of economic downturn.

Wilson et al. (2000) studied payment behavior prediction of 7 034 UK companies with logistic regression. They found that history of payment behavior is more predictive than accounting data. The evaluation was implemented with two aspects; that of predicting future payment behavior and that of corporate failure prediction.

2.3.2 Default in credit card and instant loan markets

Agarwal et al (2009) assessed the role of individual social capital information characteristics on household default and bankruptcy outcomes. They used monthly panel data set of more than 170 000 credit cardholders for a period of over 24 months. With the observations of each borrower's default and bankruptcy filing status they were able to find distress factors such as riskiness, spending, debt, income, wealth, economic conditions, legal environment and socio-demographical characteristics that affect default (see Table 1). The study was conducted with Cox proportional hazard model. The results show that borrowers who migrate from their state of birth default more. Another suggestion was that a borrower who is married and owns a house of his own has a lower risk of default. With respect to age, observation was that the youngest (30 years or younger) and oldest (60 years or older) groups of consumers had the lowest bankruptcy risk. Income and wealth were also relatively significant indicating that cardholders with high income and high wealth are 17 and 22 percent, respectively, less likely to default on their debt.

Dunn and Kim (1999) studied household credit card use to investigate the determinants of default with a monthly random household telephone survey conducted by the Center for Survey Research Center at the Ohio State University in each of the 12 months per year from the period February 1998 through May 1999. The sample consists of at least 500 households throughout the state of Ohio. It focused on the relationship between default and the outcomes of financial choices consumers make within the constraints of the contract terms set by credit card issuers. They found the three most significant variables to be: 1) the ratio of total minimum required payment from all credit cards to household income, 2) the percentage of total credit line which has been used by the consumer, and 3) the number of credit cards on which the consumer has reached the borrowing limit. They also found that socio-demographical variables like age, marital status and number of children are strongly related to default whereas income, education and home-ownership did not have expected effect (see Table 1).

There are few studies whose main focus is not in minimizing the misclassifications between "good" and "bad" accounts but in the profit maximization. Boyes et al. (2002) document that traditional view of default probability is too narrow. According to Boyes et al. the goal of

credit assessment should be to provide accurate estimates of each applicant's probability of default and the pay-offs that will be realized in the event of default or repayment. They demonstrate in credit card lending how maximum likelihood estimates of default probabilities can be obtained from a bivariate censored probit framework using a choice based sample originally intended for discriminant analysis. Through this framework they were able to obtain a more meaningful model of credit assessment. Out of 4 632 credit card applicants 1 773 (47,8%) turned out to default their loan. Variables that were significant (at 5% level) were age, number of dependants, education, home ownership, expenditures to income ratio, finance company reference and several credit bureau variables. Lieli and White (2008) had also doubts about the CSM as such and they examined a profit- or utility-maximizing lender's decision about extending or denying credit in consumer credit markets. Lieli and White suggest that lenders should measure the probability of a loan by its net present value (NPV) defined as the revenue stream of the loan, discounted at an appropriate rate, minus the amount of the loan.

Autio et al. (2009) conducted a comprehensive study of the use of small instant loans[21] in Finland among 1 951 young adults. An open online survey for 18- to 29-year-olds included questions about age, gender, financial situation, such as income, employment and occupational status, and family structure. They were also asked what kind of credit they have: a credit card, a mortgage, a student loan, small loans etc. Their attitudes towards borrowing were also examined. The results showed that the 18- to 23-year-olds use small instant loans more than the 24- to 29-year-olds. The latter group, on the other hand, use consumer credit more, because of their higher income and occupational status. Gender does not seem to have an effect on the number of loans taken, but occupational status, income and household structure do.

2.3.3 Default in mortgage markets

Default has been mostly studied in the area of corporate loans followed by mortgages. No generalization between these markets and consumer credit markets can be done since for example mortgages usually require collateral, which is not the case in the Finnish consumer

[21] Autio et al. uses the concept of consumer credit when studying instant loans.

credit market. In addition, the default of mortgage-related loans is not as straightforward as in consumer credit. Customers are often affected by the volatility in prices and interest rates when it comes to mortgages and other long-term loans (Zorn and Lea, 1989).

Vasanthi and Raja (2006) estimated the likelihood of default risk associated with income and other factors with Australian data (Australian Bureau of Statistics, ABS 2001) in a sample of 3 431 households. The goal was to establish the relationship between the default risk of homeowners and their socio-economic and housing characteristics. The repayment rate is substantially high compared to consumer credit, amounting to 93,03%. Vasanthi and Raja find that the age of the head of the household is significant: the younger households tend to be adversely affected by the increasing burden of mortgage payments. Income as socio-demographic variable show to have predictive power: lower income is one of the major contributory factors for default. Another important factor was the loan to value ratio indicating that higher loan to value ratio would increase the probability of default. Also the educational level of the head of household and marital status had significance impact on default. Vasanthi and Raja draw a conclusion that the probability of default is higher with an uneducated, younger and divorced as head of the family compared to others. The other variables employed in the study can be found in Table 1.

2.3.4 Default in consumer credit markets

Credit default predictors have been studied through several financial models. One of the most common methods is logistic regression that was also employed by Kočenda and Vojtek (2009). They show that socio-demographic data is a useful predictor of future characteristics relevant to the loan granting process tested with both logistic analysis and CART analysis. Kočenda and Vojtek use dataset of 3 403 observations and 21 variables (see Table 1). With both methods the most important financial and behavioral characteristics of default behavior were: the amount of resources a client owns, the level of education, marital status, the purpose of the loan, and the years of having an account with the bank. The dataset Kočenda and Vojtek analyzed was relatively small for CSM with only 3 403 observations but showed that most variables had reliable information value and were able to give information to construct a good scoring model based on both the traditional parametric as well as the non-parametric method. This study provides evidence that non-parametric methods can also be successful and

are able to create good models. Following Kocenda and Vojtek this thesis constructs several models to evaluate the performance of also variables with lower predictive power.

Arminger, et al. (1997) analyzed the three different techniques; logistic discriminant analysis, classification tree analysis and a feedforward network in finding the best method in predicting the determinants of default. Their data consisted of 8 163 observations provided in 1991 and 1992 by a major bank in Germany specializing in consumer loans. The predictor variables they employed were sex, starting year of current job, year of birth, car ownership and marital status. They report that the predictive power is about equal for all techniques with logistic discrimination providing the best estimates. The logistic discriminant analysis suggests that the probability of paying back the loan without problems is greater for telephone owners and older people. People with longer employment at current job, car owners, female and people who are married rather than single are also less likely to default (see Table 1).

Another interesting prospect is defined by Musto and Souleles (2006) by taking a portfolio view of consumer credit. They used a unique panel dataset of credit files from one of the major U.S. credit bureaus, Experian which includes approximately 100 000 randomly sampled consumers monthly from March 1997 to March 2003, a total of 37 months. Unlike most of the default studies Musto and Souleles computed also the risk-adjusted returns, as lenders also need to know the covariances of the returns on their loans with aggregate returns. They measured the covariance risk of individual consumers, i.e., the covariance of their default risk with aggregate consumer default rates. This is to analyze the cross-sectional distribution of credit, including the effect of credit scores. Musto and Souleles found that consumers with high covariance risk tend to have low credit scores (high default probabilities) and that amount of credit obtained by consumers significantly increases with their credit scores and significantly decreases with their covariance risk. Covariance risk tends to be higher for younger and single consumers, lower-income consumers, those who rent rather than own, and those from states with higher raters of divorce and lower rates of health-insurance coverage (see Table 1).

Jacobson and Roszbach (2003) contribute to the existing literature by taking into account the sample-selection bias that credit scoring models are suffering from. Therefore the basic value-at-risk measure is not reliable enough but they suggest using unbiased scoring model such as bivariate probit approach that takes into account also the rejected loans. In their work they

used a data set consisting of 13 338 applications for a loan at a major Swedish lending institution between September 1994 and August 1995. All loans were granted in stores where potential customers applied for instant credit to finance the purchase of a consumer good. The dataset contained extensive financial and personal information on both rejected and approved applicants (Table 1). They had 57 variables available but employed only 16 because they lacked a univariabte relation with the variables of interest of displayed extremely high correlation with another variable. Income, age, change in annual income and amount of collateral-free credit facilities had significant impact on default.

Roszbach continues the study in 2004 with the same Swedish sample to show that not only the default on loan matters but also the timing of default. Roszbach (2004) contribute to the existing literature by presenting a multiperiod character in a credit default topic and at the same – questioning the usage of CSM. By illustrating this aspect, a loan is usually a multi-period contract and thus generates a flow of funds until it either is paid off or defaults. The net present value of a loan is thus not determined by whether it is paid off in full or not but by the duration of the repayments, amortization scheme, collection costs and possible collateral value. Roszbach emphasizes that it may still be profitable to provide a loan, even if the lender is certain that it will default since the goal of the lender is to maximize profit. Roszbach observed the exact survival time for the loans in the dataset by constructing a Tobit model with sample selection and variable censoring threshold. The results show that financial institutions are not acting rationally when taking into consideration both the default risk and higher returns. The lending policy of companies does not favor people that survive longer and thus would provide higher rates of return. Roszbach also found that lenders are indifferent between loans of different sizes. This study provides evidence that banks are behaving in a way that is not consistent with profit-maximization. By using Tobit model banks would be able to pick out future defaults and select applicants with longer survival times and thus create a more efficient policy.

Dinh and Kleimeier (2007) used a database of one of the Vietnam's commercial banks and had access on sample of 56 037 loans. They used forward stepwise selection to select among 22 variables (see Table 1). Applying stepwise methods, 16 variables were included in the model. Their paper addressed the lack of information on retail credit scoring by identifying which borrower characteristics a bank needs to collect. Dinh and Kleimeier developed a flexible approach that is built on the principles of transactional lending but leaves room for

relationship lending. The most important predictors they found were time with bank, followed by gender, number of loans, and loan duration. Dinh and Kleimeier suggest companies to update their CSMs regularly to answer to the economic changes.

Updegrave (1987) found that there were eight variables that affected consumer credit risk: the number of variables, the historic repayment record, bankruptcy history, work and resident duration, income, occupation, age and the state of savings account. Similar results were found by Steenackers and Goovaerts (1989) who collected data on personal loans in Belgian credit company. The loans dated from November 1984 till December 1986 and contained 995 good loand, 1257 bad loans and 693 refused loans. They were able to use 19 characteristics (see table 1) of which 11 were employed to construct a CSM. By using logistic regression and for the final selection a stepwise analysis they found following results: age, resident and work duration, the number and duration of loans, district, occupation, phone ownership, working in the public sector or not, monthly income and housing ownership have a significant relationship with repayment behavior.

Özdemir's work (2004) explored the relationship between consumer credit clients' credit default risk and some demographic and financial variables with a logistic binary regression. Data to examine this relationship was obtained from the customer records of a private bank in Turkey. Interestingly, Özdemir does not find significant relationship between any of the demographic variables and the risk of default. Residential status seemed to be the most important demographical variable with relatively high p-value, however. Instead, the financial variables had significant predictive power. Interest rate and maturity both positively affected the credit default risk, thus, the longer the maturity or the higher the interest, the higher the risk for clients not paying their loans on time.

Hand and Henley (1997) made a wide review of different statistical methods in consumer credit scoring. They compared LDA, OLS regression, LR, mathematical programming methods, recursive partitioning, expert systems, NN, nonparametric methods and time varying models[22]. Hand and Henley stated that there is no overall "best" method but it depended on the details of the problem: on the data structure, the characteristics used, the extent to which it

[22] The abbreviations are represented as LDA: Linear Discriminant Analysis (also known as DA: Discriminant Analysis), OLS: Ordinary Least Squares, LR: Logistic Regression, NN: Neural Networks.

is possible to separate the classes by using those characteristics and the objective of the classification. The variables employed in this study are presented in Table 1. Followed by Hand and Henley, Tsai et al. (2009) constructed a consumer loan default predicting model using dataset from a Taiwanese financial institution. They studied both the consumers' demographic variables and money attitude and constructed four predicting methods, DA, LR, NN and DA to compare the suitability of these methods. They found that the predictive efficiency with all these four methods was more than 75%.

2.4 Summary of the variables used in earlier studies

Table 1 shows the most used variables in previous literature. The most common socio-demographical variables seem to be age, time in current address, gender, income, marital status, occupation, number of other loans, residential status, time in present job and region. The correspondent customership-based or behavioral variables are length or relationship, loan size and duration of the loan.

From these age, marital status, number of other loans, residential status, loan size and duration of the loan seem to have the best predictive power in determining default and are the ones to form a reliable CSM for credit institutions.

Table 1: Variable comparison

This table presents the most common variables employed in previous literature. The table does not cover exactly the same variable expression as the authors have used but is meant to provide summary of the most significant variables. Neither is the division to socio-demographical and behavioral variables based on any of the mentioned studies but is for the purposes of this thesis. Csm stands for variables that are used in credit scoring model or considered as the most significant variables in the certain studies. Desai et al. as well as Hand & Henley and Lieli & White do not document the importance of variables but concentrate on investigating the best technique.

Socio-demographical variable	Agarwal et al. 2009	Desai et al. 1996	Dinh & Kleimeier 2007	Dunn & Kim 1999	Hand & Henley 1997	Jacobson & Roszbach 2003	Kocenda & Vojtek 2009	Lieli & White 2008	Steenackers & Goovaerts 1989	Vasanthi & Raja 2006	Özdemir & Boran 2004
age	csm	x	x	csm	x	csm	x	x	csm	csm	x
big city						x					
credit card ownership		x		csm	x						
credit history								x			
current address / time in current addres	x	x	csm		x			x	csm		
education			csm				csm			csm	
foreign worker								x			
gender			csm			x	x	x	x		x
government assistance										x	
income / change in income	csm	x	x	x	x	csm			csm	csm	x
marital status	csm		csm	csm	x	x	csm		x	csm	x
migrating out of state of birth	csm										
monthly expenses	x	x							x		
nationality									x		
nr of children			csm	x					x	x	
occupation / type of employment			x	x	x		x	x	csm	x	x
old loans / nr of other loans	x	x	csm			csm		x	csm		
phone			csm		x			x	csm		
principal											x
residential status / housing	csm	x	x	x	x	csm		x	csm		(csm)
sector of employment							csm				
state of birth	x										
wealth	csm										
working in private / public sector									x		
years of employment / time in present job		x	x		x		csm	x	x		
zip code / region	x		csm		x		x		csm		
Behavioral variable											
collateral type / value			csm								
cosigner / guarantor						x		x			
credit type							x				x
interest / interest rate											csm
length of relationship		x	csm		x		csm				
loan size / credit limit	csm			x		csm	csm	x	x		x
loan to value ratio										csm	
maturity / duration of the loan			csm					x	csm		csm
monthly payments		x									
nr of payments											x
own resources / savings			csm				csm	x			
payment performance											x
purpose of loan											
score / points		x					csm				

2.4.1 The weakness of previous literature

In previous literature the number of variables has usually been between 10 and 20 variables or less. Excluding the study of Kočenda and Vojtek (2009) several studies do not concentrate on selecting the most reliable CSM between variables but settle for one model. The excellence of this thesis is to have high number of explanatory variables and evaluating the importance of variables with the help of several models.

There are certain socio-demographical variables that would be easy to include in the model and are proven to have significant results but are not used as a predictive variable in the previous studies. Such are income, ownership of a real estate and time since last moving. Previous studies have mostly concentrated on having a reliable technique in determining default but have not tested those with up to date, broad sample with essential variables. This study employs a unique set of variables such as housing type, military service and time of applying a loan, which have not been used in previous studies before.

Most of the studies have employed data from U.S or Asia. There is not much evidence on European not to mention Nordic determinants on default (note Jacobson & Roszbach, 2003). Especially the socio-demographical variables may give varying results depending on nationality. For example income is not purely comparable between nationalities. Housing type, education and number of household vary depending on culture and thus are not necessarily comparable between continents or even between countries.

Many previous academic studies have been lacking the credibility and practicability due to the small size of the sample used in model estimation. This thesis, instead, is able to provide applicable results with a sample of 14 595 observations.

3 Data description and summary statistics

In this chapter I will present the data for the empirical study and analysis. After describing the dataset I will introduce the variables and give some statistical information about them.

3.1 Data

This study uses a unique dataset from one of Finland's largest and well-known consumer credit companies who has over 150 000 customers. The lender has specialized in providing small- and medium-sized loans to retail customers. The collected data includes several socio-demographic variables such as education, marital status, size of household etc. I also have information on the customership or here: behavioral variables including for example the scores based on which the customer has been evaluated and the length of relationship between the customer and the lender.

The initial sample consisted of 103 037[23] applications received between May 27th, 2008 and September 1st, 2009. Out of these 14 595 were accepted and 88 442 rejected. From the accepted 29%[24] turned out to default and 71% performed well. Each customer is allowed to have only one loan at the same time. So there is no need to aggregate several loans for one individual, as is often the case for scoring companies.

From the empirical analysis I have excluded observation of customers who applied for a loan but were rejected due to small credit scores evaluated by the company. The dataset itself would have been larger but the amount of defaulted loans would have remained the same. The true creditworthiness status of the rejected applicants is unknown and their characteristics might differ from those who were granted the loan. The exclusion might cause a potential selection bias but is common in the literature and according to Banasik et al. (2003) has only a minimal effect on results.

[23] Same customer is allowed to apply for a loan three times during a 90-day period. After the third rejected application this customer is blocked from getting an approval. Therefore, the sample may include maximum of three applications for the same applicant.
[24] The amount of defaulted customers was by December 15th, 2009.

The information for socio-demographical variables is given by the customer at the time of filling loan application online. Along with the terms and conditions the company has placed, the customer is obligated to give true and fair information. The behavioral variables are observed at the time of application and customership.

The following sections describe the variables.

3.1.1 Variables

Table 2 lists variables and their definitions. The variables are divided based on their socio-demographical or behavioral characteristics in the analysis. I employ 30 explanatory variables out of which 23 are socio-economical and 7 behavioral describing the relationship between the customer and the lender. The same data has been used for the company's own scoring model and customer evaluation and thus all of the categories are taken as given. The classification is a common practice (Kočenda and Vojtek, 2009) in modeling CSM and especially when using logistic regression as empirical analysis classes to determine if either one has more explanatory power. The categories for variables can be found in Table 14 (Appendix).

Table 2: Variable definition

This table presents the definitions for each variable in the analysis. Variables are divided into two categories: socio-demographical and behavioral. The column Definition describes whether the variable is dummy i.e. is possible to have two binomial values or categorical i.e. has three or more values. The categories for each variable can be found in Table 14.

Variable	Definition	Socio-demographical	Behavioral
AGE	categorical, age of applicant	x	
CITY	dummy, takes value 1 if applicant lives in one of the 5 largest cities	x	
COTTAGE	dummy, takes value 1 if applicant owns a cottage	x	
CREDIT	dummy, takes value 1 if applicant has one or more credit cards	x	
DEFAULT	dummy, takes value 1 if applicant has defaulted the loan		x
EDUCATION	categorical, education	x	
EMPLOYMENT	categorical, the type of employment	x	
FREEEMAIL	dummy, takes value 1 if applicant has given a free email address	x	
GENDER	dummy, takes value 1 if applicant is female	x	
HOUSING	categorical, the residential type	x	
HOUSINGTYPE	categorical, housing type	x	
INCOME	categorical, monthly income (in EUR)	x	
LEVEMPL	categorical, level of employment	x	
LOANSIZE	categorical, the amount of loan in euros		
MARITAL	categorical, marital status	x	
MILITARY	dummy, takes value 1 if applicant has completed military service	x	
MONTHLY	categorized, the amount applicant repays per month (in EUR)		x
MOVING	categorical, years since last moving	x	
NATIONALITY	dummy, takes value 1 if applicant isn't a finnish citizen	x	
NATIVE	categorical, native of applicant	x	
NRADULTS	categorical, nr of adults in household	x	
NRCHILDREN	categorical, nr of children in household	x	
PAYBACK	categorized, the time of payback (in months)		x
PHONE	dummy, takes value 1 if applicant has called to have credit		x
POSTAL	categorical, in which postal area the applicant lives in	x	
PREVLOAN	dummy, takes value 1 if has had loan earlier from the company		x
REPAYMENTBEH	dummy, takes value 1 if applicant has had payment problems		x
SCORE	categorical, the scores applicant has received given by the lender	x	
SIZEHOUSEH	categorical, nr of persons in household	x	
TIME	categorical, the time of day applying a loan		x
TIMEFIN	categorical, nr of years the applicant has lived in Finland	x	

3.1.2 Variable definition and descriptive statistics

To illustrate the properties of the whole dataset I include Table 3 that presents the statistics when both the defaulted and non-defaulted observations are included.

Table 3: Descriptive statistics for whole sample

This table presents the descriptive statistics for the whole sample. The definitions for categories can be found in Table 14 (Appendix).

Variable	Nr of observations	Mean	Stdev	Min	Max
AGE	14595	2,38	0,752	1	5
CITY	14595	0,27	0,445	0	1
COTTAGE	14595	0,17	0,379	0	1
CREDIT	14595	0,66	0,474	0	1
DEFAULT	14595	0,29	0,452	0	1
EDUCATION	14595	1,27	0,995	0	3
EMPLOYMENT	14595	0,78	0,416	0	1
FREEEMAIL	14595	0,82	0,387	0	1
GENDER	14595	0,44	0,496	0	1
HOUSING	14595	0,49	0,725	0	3
HOUSINGTYPE	14595	0,89	0,923	0	3
INCOME	8229	2,36	1,119	0	4
LEVEMPL	14595	3,36	1,917	0	10
LOANSIZE	14595	1,57	1,179	0	3
MARITAL	14595	1,01	1,108	0	5
MILITARY	14595	0,44	0,497	0	1
MONTHLY	14595	1,57	1,05	0	3
MOVING	14595	4,23	2,727	0	8
NATIONALITY	14595	0,07	0,486	0	4
NATIVE	14595	0,06	0,293	0	2
NRADULTS	14595	0,4	0,523	0	2
NRCHILDREN	14595	0,42	0,843	0	3
PAYBACK	14595	3,93	2,196	0	6
PHONE	14595	0,07	0,257	0	1
POSTAL	14595	3,19	3,072	0	9
PREVLOAN	14595	0,31	0,461	0	1
REPAYMENTBEH	14595	0,16	0,368	0	1
SCORE	14595	1,71	1,498	0	4
SIZEHOUSEH	14595	0,82	1,184	0	4
TIME	14595	2,24	0,753	0	4
TIMEFIN	14595	0,05	0,378	0	4

The dataset is divided into two parts, the loans that turned out to be good (at the time of the study) and the ones that turned out to be bad. On December 15th, 2009 4 191 of those who obtained a loan had defaulted while 10 404 borrowers still fulfilled their minimum repayment obligations at that time. The magnitudes of estimated means and corresponding standard

errors in Table 4 imply that a formal test for differences in means between the two groups of defaulted and non-defaulted loans will not yield significant test statistics for any variable. To formalize the distributions of the explanatory variables a logistic regression is conducted and the results are interpreted in Chapter 5.

Table 4: Descriptive statistics

This table shows the descriptive statistics for both defaulted and non-defaulted customers.

	Defaulted loans (N=4191)				**Non-defaulted loans (N=10404)**			
Variable	Mean	Stdev	Min	Max	Mean	Stdev	Min	Max
AGE	2,21	0,76	1	5	2,44	0,74	1	5
CITY	0,30	0,46	0	1	0,26	0,44	0	1
COTTAGE	0,15	0,36	0	1	0,18	0,39	0	1
CREDIT	0,60	0,49	0	1	0,68	0,47	0	1
DEFAULT	1,00	0,00	0	1	0,00	0,00	0	1
EDUCATION	1,24	0,95	0	3	1,29	1,01	0	3
EMPLOYMENT	0,79	0,41	0	1	0,77	0,42	0	1
FREEEMAIL	0,83	0,38	0	1	0,81	0,39	0	1
GENDER	0,34	0,47	0	1	0,47	0,50	0	1
HOUSING	0,58	0,72	0	3	0,45	0,73	0	3
HOUSINGTYPE	1,05	0,92	0	3	0,83	0,91	0	3
INCOME	2,27	1,08	0	4	2,42	1,14	0	4
LEVEMPL	3,41	1,86	0	10	3,35	1,94	0	10
LOANSIZE	1,71	1,16	0	3	1,51	1,18	0	3
MARITAL	1,12	1,08	0	5	0,97	1,12	0	5
MILITARY	0,48	0,50	0	1	0,43	0,49	0	1
MONTHLY	1,55	1,03	0	3	1,58	1,06	0	3
MOVING	3,59	2,67	0	8	4,49	2,71	0	8
NATIONALITY	0,03	0,18	0	1	0,02	0,13	0	1
NATIVE	0,07	0,34	0	2	0,05	0,27	0	2
NRADULTS	0,50	0,54	0	2	0,36	0,51	0	2
NRCHILDREN	0,53	0,90	0	3	0,37	0,81	0	3
PAYBACK	4,23	2,10	0	6	3,81	2,22	0	6
PHONE	0,07	0,25	0	1	0,07	0,26	0	1
POSTAL	3,09	3,05	0	9	3,23	3,08	0	9
PREVLOAN	0,48	0,50	0	1	0,24	0,43	0	1
REPAYMENTBEH	0,00	0,00	0	0	0,23	0,42	0	0
SCORE	1,33	1,45	0	4	1,86	1,49	0	4
SIZEHOUSEH	1,03	1,23	0	4	0,74	1,15	0	4
TIME	2,24	0,77	0	4	2,25	0,75	0	4
TIMEFIN	0,06	0,43	0	4	0,04	0,35	0	4

To illustrate the descriptive characteristics more the Appendix (Table 14) provides cross-tabulated data on variables against default.

Next, more precise definitions are given to the variables.

3.1.2.1 Response variable

The response or dependent variable is called DEFAULT, which describes the customer's repayment ability. The company's definition of default is identical to the Basel II framework[25]: the borrower is in default if he is more than 90 days overdue with any payment[26] connected with the loan. After this time period the loan is forwarded to a debt-collection agency and the customer is prevented for having another credit from the company. Suomen Asiakastieto who registers defaults will also get the information after the municipal court[27] has given its decision of default. Depending of the type of the default the tag in customer's default register will remain from two to five years. This will affect the customer's credit standing in the future. For example getting a bank loan or credit card without guarantors is difficult or even impossible and even renting an apartment or applying a new job will become difficult[28].

Response variable is a binary or dummy variable and thus can have two values: 0 and 1. Customers who have performed well will receive value 0 and customers who have defaulted their payments will receive value 1. Table 3 gives the mean and standard error of the variables used in the study. The mean value of DEFAULT is 0,29, which implies that 71% of customers on an average repay their loans as scheduled while 29% default on their payments.

[25] See www.basel-ii-association.com for details.
[26] Any payment includes principal or interest.
[27] Käräjäoikeus in Finnish.
[28] The Finnish Law of Credit Reference (luottotietolaki) regulates issues of default and credit references. The grounds for default are also regulated.

3.1.2.2 Explanatory variables

This section presents the explanatory or predictor variables in the study. Explanatory variables are further divided to socio-demographical and behavioral categories.

Each explanatory variable has 14 595 observations[29]. The division of amounts between defaulted and non-defaulted among different categories can be found in Table 14 (Appendix).

3.1.2.2.1 Socio-demographical variables

The socio-demographical[30] variables characterize the customer and his features at the time of the application.

AGE defines the age of an applicant in years and is described as a categorical variable ranging from 0 (under 20 years) to 5 (over 70 years) with mean 2,38 (std. error = 0,75) showing the average age of a customer to be between 26-45 years. The first category includes customers under 20 years old but has no observations, as the company's policy is not to grant loan for under 20-year-olds at the moment. It is often assumed that older borrowers are more risk averse and will therefore be less likely to default. Dunn and Kim (1999), Arminger et al. (1997) as well as Agarwal et al. (2009) can confirm this empirically. They found that probability to perform well is greater for older people. Similarly, I expect the risk of default to decline in the later stage of life.

[29] INCOME in an exception and has only 8 231 observations due to companies policy to not to require this information in the beginning.
[30] Socio-demographic refers to different groups of people within the society.

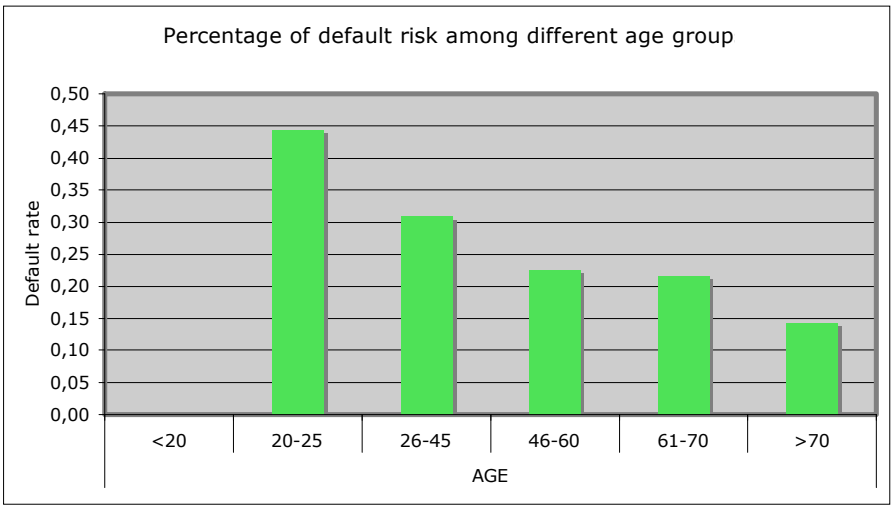

Figure 3: Percentage of default risk among different age groups

Figure 3 presents the percentage of default risk among different age groups. It can be seen that 20-25-year-olds have the most defaults. Almost 45% of customers in this group will default on their payment. Age group 26-45 has also defaulted more than an average customer (29%). This is consistent with Statistics Finland[31] (2010): 25-49 years old have the most payment troubles.

Autio et al. (2009) analyzed the use of instant small loans among young people (18-29 years) in Finland and found that the young customers, who borrow money pay bills overdue, have weaker financial position and recognize overall flaws in their money management. Often the reason for applying an instant loan is to cover a rent that is already overdue. Young people are also seen to take new loans to pay off previous ones. These facts might reflect also on the behavior of customers in consumer credit market.

The variable CITY divides customers into two groups: those who live in one of the five largest cities[32] and those who live elsewhere. The division is based on to the postal code applicant has given in the application. Rosbach (2004) has employed big city variable and found that people living in one of the three metropolitan areas in Sweden seem to default more and thus have significantly smaller chance of being granted a loan.

[31] Tilastokeskus in Finnish.
[32] Five largest cities are Helsinki, Espoo, Tampere, Turku and Oulu. Information is based on the amount of population.

COTTAGE is a dummy variable having a value 1 if an applicant owns a cottage. This is a sign of owning a real estate and thus having more financial wealth. Those with high wealth are 19 percent less likely to default their debt (Agarwal et al., 2009) compared to those with low wealth.

The ownership of a credit card is defined with variable CREDIT. Agarwal et al. (2009) found that borrowers with higher amount of other debt are significantly less likely to default on their credit card debt. This might be consequence of several credit institutions of monitoring the applicant's financial standing and repayment behavior.

Regarding education I expect that better educated people have more stable, higher-income employment and thus default less. This characteristic is represented with EDUCATION that has four subcategories. Steenackers and Goovaerts (1989) show that customers with high-educated professions were less likely to default on their loans. Figure 4 presents the percentage of default risk among different education categories. It can be seen that no major differences in default rates exist but those who have went to college default less (25%) compared to others.

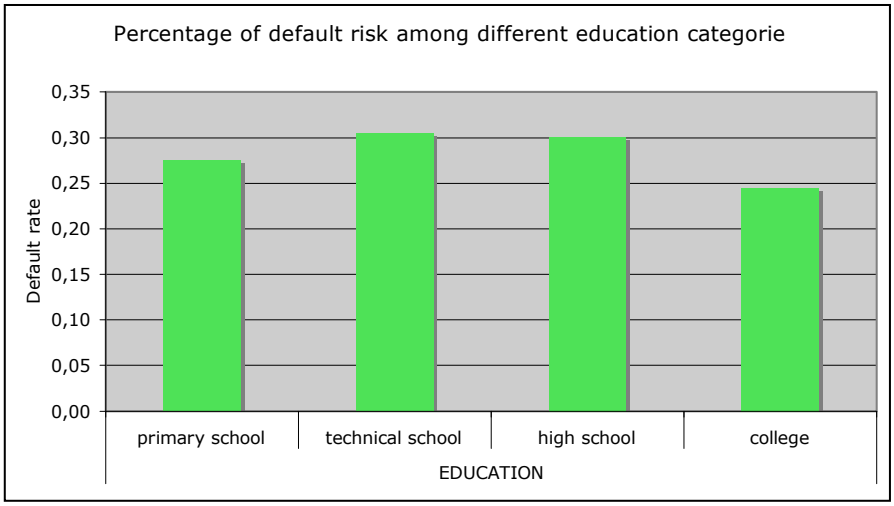

Figure 4: Percentage of default risk among different education categories

EMPLOYMENT describes the type of employment. Variable has four subcategories but has observations only in two categories. In Finland, the type of employment may not be a suitable proxy today as many employments begins with fixed-term agreement and often will not even turn to permanent agreement. Only 22 percentages of customers had fixed-term agreement.

FREEEMAIL is to define whether having a free email address can affect on default. Having a free email address instead of a purchased one or an email provided by employer might indicate the customer is not afford to purchase one or is not employed.

GENDER in addition to age is one of the most used socio-demographical variables to differentiate the predictive power between men and women. There is clear evidence that women default less frequently on loans (Arminger et al., 1997) possibly because they are more risk averse. Figure 5 describes the default rate differences between men and women. It can be seen that 34% of male customers defaulted while only 23% of female customers had payment troubles.

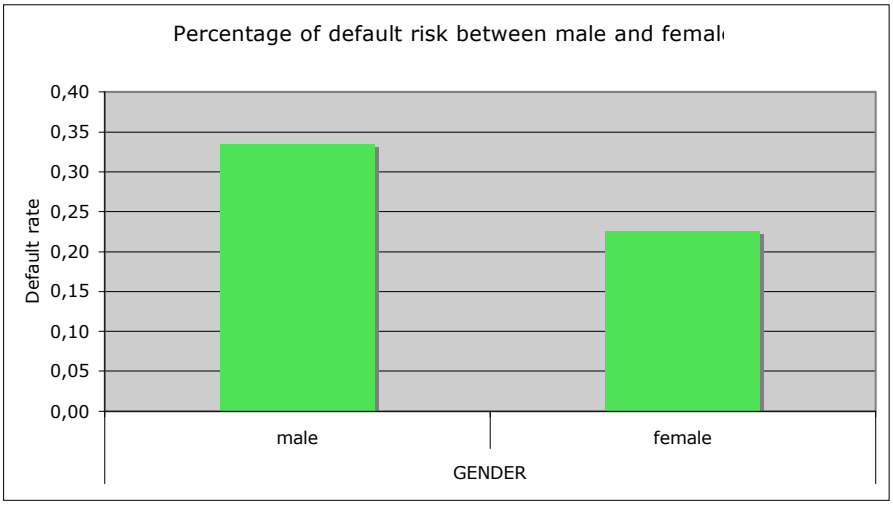

Figure 5: Percentage of default risk between male and female

HOUSING describes the residential type of applicant. As shown by Steenackers and Goovaerts (1989) and Agarwal et al. (2009) residential status can indicate financial wealth in particular in the case of home ownership. Agarwal et al. show evidence that an individual who owns a home is 17 percent less likely to default and 25 percent less likely to file for bankruptcy. I expect the risk of default to be lower for a debtor who owns a home compared to those who rent, have employment relationship apartment or partial ownership of an apartment.

HOUSINGTYPE defines if applicant lives in a house, a row house, an apartment or another type of accommodation. Housing type might indicate a level of financial wealth and thus have an effect on default. HOUSINGTYPE does not include the information of ownership but I expect people with more wealth to live in a house as renting a house rather than an apartment is often more expensive.

INCOME presents the borrower's monthly income in Euros and is categorized in five subcategories. INCOME has only 8 229 observations due to company's policy not to require income in its application form until 2009. According to Jacobson and Roszbach (2003) and Agarwal et al. (2009) among others, income has significant predictive power and those with high income and high wealth are less likely to default on their debt. Note, that in my study, the average (median) monthly income is under 2 500 Euros. In contrast, the average monthly income in Finland is about 2 876 Euros[33]. Thus the bank's borrowers – including the defaulted ones – have an income that lies clearly below the national average.

LEVELEMPL is a variable to define in which of the 11 level of education an applicant is. Occupation is one of the commonly incorporated variables since it is highly correlated with income (Dinh & Kleimeier, 2007).

LOANSIZE is the amount of credit the applicant is granted. The customer may have applied for larger amount but has been denied the loan. He is able to try lower amount for maximum of three times. Several studies use loan size as a predictor variable but the overall results are ambiguous and thus no clear expectations can be formed. Jacobson & Roszbach (2003) show that loan size has no significant influence on default risk. In the study of Kocenda and Vojtek (2009) small loans appear to be more risky if variable 'own resources' is included. However, if this information is not used, the regression identifies that the larger loans as more risky.

MARITAL is to investigate whether different marital status can predict default as it is often seen as sign of responsibility, reliability or maturity of borrowers. It is very common variable in default literature and for example Agarwal et al. (2009) suggest that a borrower who is married is 24 percent less likely to default on his credit card debt and 32 percent less likely to file for bankruptcy. This is consistent with the statistics of this study (see Figure 6). A

[33] See www.stat.fi for details.

customer who is married defaults in 24% of the cases while single customers tend to default most often, in 36% of the cases. Similar to Arminger et al. (1997) other categories than single are not significantly different from the category married. Hence, I expect the risk of default to increase for single customers.

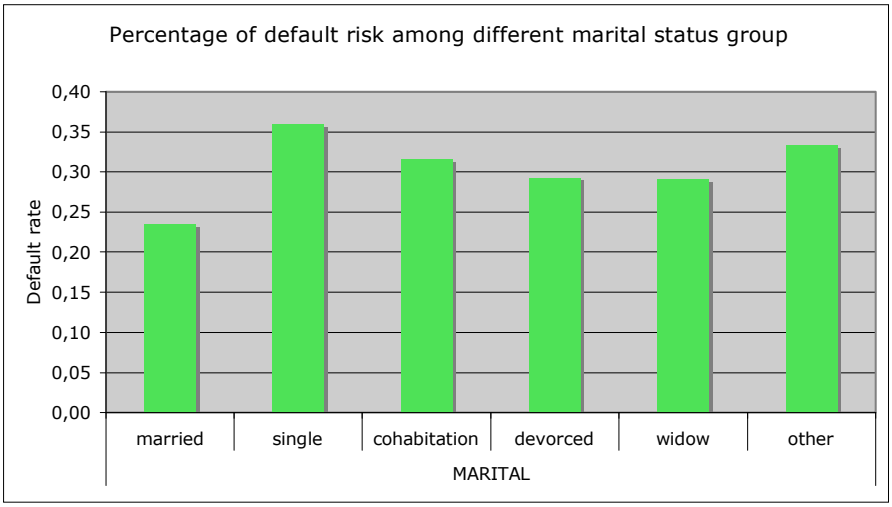

Figure 6: Percentage of default risk among different marital status groups

The accomplishment of military service is defined with MILITARY. As only men tend to fulfill military service compared to women, this variable is highly correlated with GENDER and I except those who have not accomplished military service to default less.

Moving has shown to predict default (Agarwal et al., 2009): the risk of personal bankruptcy and default is higher for an individual who migrates out of his state of birth. This paper employs MOVING to investigate whether years since last moving has expected effect on default behavior. Steenackers and Goovaerts (1989) give evidence that the less time since last moving the more likely the customer is to default.

The Finnish nationality is not a requirement for applying a consumer credit whereas holding a Finnish social security number is. NATIONALITY is divided in two: those who have the Finnish nationality and those who have another origin. Nationality as predictor variable is employed also by Steenackers and Coovaerts (1989) but does not seem to have significant results on default. Figure 7 show that customers who are not Finnish citizens seem to default in 44% of the cases while the default rate for Finnish is 28%.

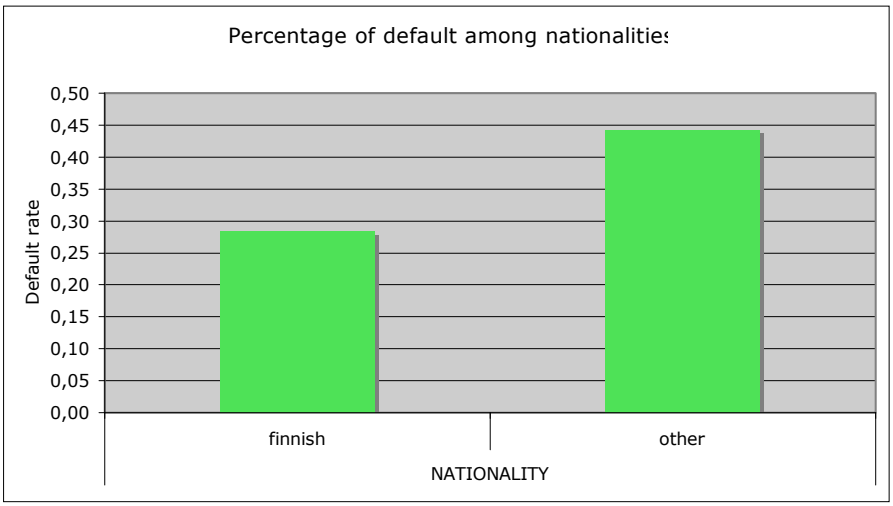

Figure 7: Percentage of default among nationalities

Customers who have Finnish as their native language default as much as an average customer, 29% of the cases (see Figure 8). Those who have Swedish as their mother language tend to have a default rate of 24%. NATIVE might correlate with NATIONALITY as those with mother language other than Finnish or Swedish tend to default in 46% of the cases.

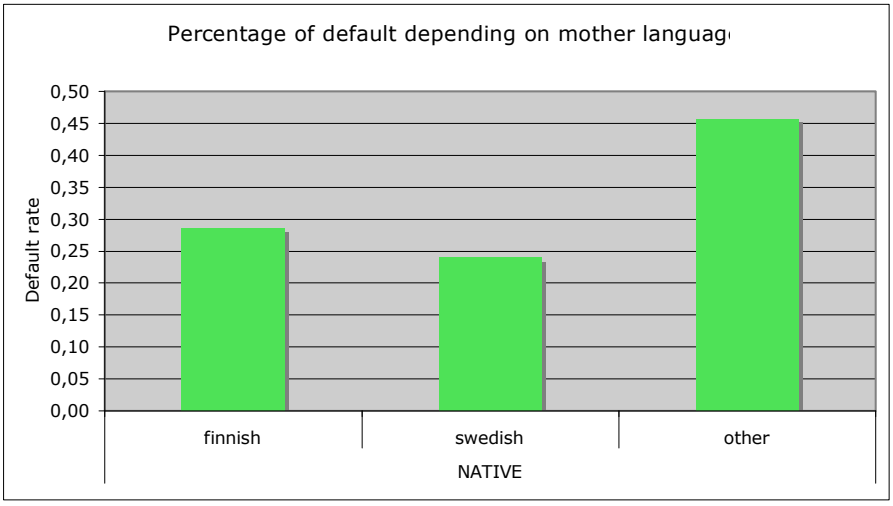

Figure 8: Percentage of default depending on mother language

NRADULTS defines the number of grown-ups in household. Similar to variable MARITAL this is seen as sign of responsibility and support from the spouse in financially uncertain

times. Interestingly, the default rate seems to increase with the number of adults in the household (see Table 14 in Appendix).

NRCHILDREN represents the number of children that the borrower has to support. As the number of children increases, so does the pressure on the borrower's income due to higher expenses such as food and day care fees. For example, moving from zero to one child increases the default percentage from 26% to 39%. Dunn and Kim (1999) found that default is somewhat less likely for married cardholders, but its likelihood increases with number of children.

The geographical area of customer living in is defined with POSTAL. According to Steenackers and Goovaerts (1989) the geographical region is significant predictor of default. In most of the studies region is to find people with similar wealth as they tend to live in the same location and might thus indicate borrower's level of financial wealth. In this study the categorization is based on zip codes and thus may not be a suitable criterion as the categories are as much as 10 and have both cities and country sides within them.

SCORE is the amount of points the customer has received at the time of application. Agarwal et al. (2009) found that borrowers who have lower FICO risk score are more likely to default on their credit card debt, which is consistent also with findings of Gross and Souleles (2001). Consistent with how the SCORE is determined, the fewer scores the customer has the more often he defaults (Figure 9). SCORE is based on the previous customers' socio-demographical characteristics and as supposed the characteristics related to default are somewhat similar.

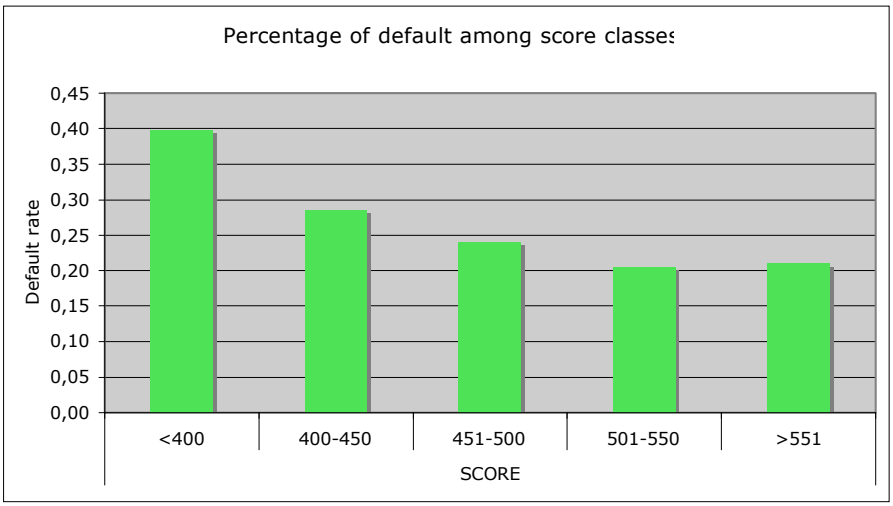

Figure 9: Percentage of default among score classes

SIZEHOUSEHOLD describes the number of people in household. This expresses the variables of NRADULTS and NRCHILDREN but is included in the analysis as combined variable to reflect also the information of single custody. Figure 10 describe that a household size of three increases the risk of default significantly.

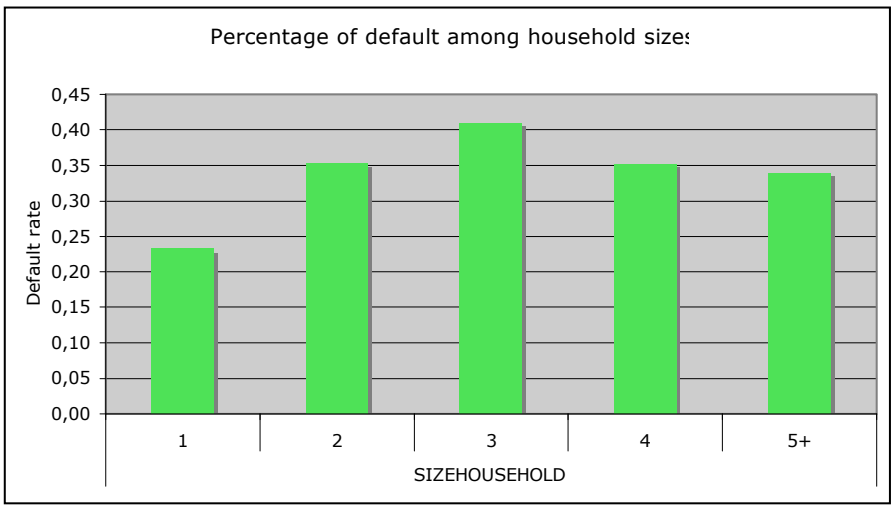

Figure 10: Percentage of default among household sizes

TIMEFIN describes the time in months spent in Finland. The default percentage is extremely high, 71%, for those who have lived in Finland for only 1-3 months (Figure 11). This might be consequence for travelers to return home out of money and as unemployed. Consistent with findings of Agarwal et al. (2009) people who migrate from their birth state are more

likely to default and file for bankruptcy. The statistics of this paper show that those who have always lived in Finland default less, 29%, than those who have moved from elsewhere.

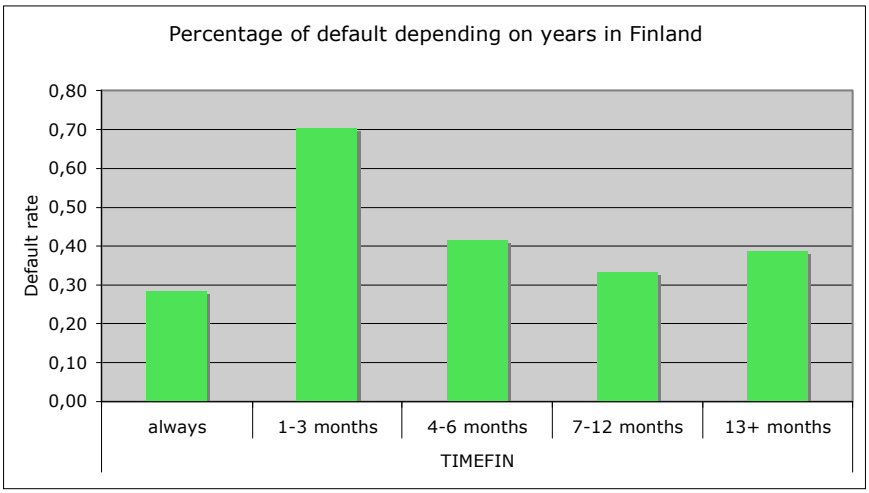

Figure 11: Percentage of default depending on years in Finland

3.1.2.2.2 Behavioral variables

The behavioral variables characterize the relationship between the customer and the bank. In some of the studies (Kočenda and Vojtek, 2009) behavioral variable defines the behavior on customer's own current account. Bank-related variables are for example the amount of resources, date of account opening and whether a collateral has been placed. However, the organization I am utilizing does not operate as a bank and thus client deposits cannot be made and no account-related information of the customer is usually available. Though I use behavioral variables they primarily describe the behavior of the customer related to the credit taking and payment matters. Nevertheless, there are two variables that describe the history of customers' current behavior on customership:

Variable PREVLOAN describes whether the customer has had consumer credit or any other loan from the company earlier. Steenackers and Goovaerts (1989) find the number of previous credits to be significant determinant of default. Dinh and Kleimeier (2007) document the default to be least frequent for repeat borrowers. PREVLOAN can be indication of the relationship with the bank. In Kocenda and Vojtek's (2009) work the length of the

relationship between client and the bank is the most important behavioral characteristic indicating that the longer the history with the bank the less likely it is for them to default.

Company has its own register for customers who have had some payment troubles during the customership but have repaid until the 90 days has passed. A variable to describe this is called REPAYMENTBEH. However, for any of the defaulted customer there is not default information. In order to avoid this variable from interfering with the results and increase the number of degrees of freedom I will exclude this variable from the empirical analysis.

PAYBACK measures the maturity of loans in months. Usually in consumer credit markets the loan duration is proposed by the borrower and thus reflects the borrower's intention, risk aversion, or self-assessment of repayment ability. Dinh and Kleimeier (2007) found loan duration to have significant effect on default prediction. The limit for loan duration is four years. Özdemir's work (2004) show evidence that the longer the maturity the higher the default risk.

Customer is able to define the amount in Euros he is willing to repay per month. This is described with variable MONTHLY. MONTHLY can be seen correlated with PAYBACK: The longer the payback period customer is willing to have the smaller the monthly repayment amount. I expect customers with smaller monthly amount to default more often compared to those with larger payback entries.

Customers are able to apply loan by phone, which is a characteristics defined with PHONE. Only 7 percentages of granted customers used phone as an application channel.

Customers are able to apply loan regardless of the time of the day. TIME reflects the time the application has been placed in. Customers who have applied for a loan between 11pm and 5am tend to default more than those who apply for a loan during office hours (Figure 12). This might be component of the Finnish law to deny creditors to transferring any money between 12am and 7am. This regulation is rationalized with weakened discretion of a borrower in certain situations.

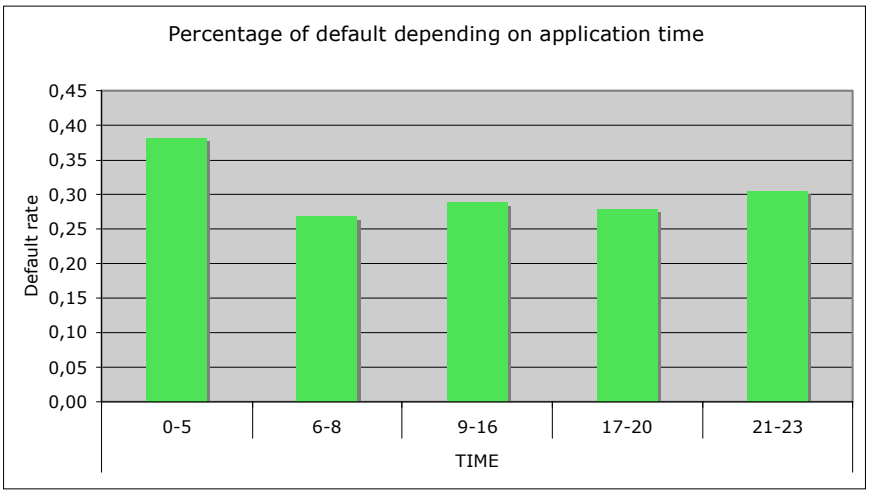

Figure 12: Percentage of default depending on application time

MONTHLY, PAYBACK, PHONE and TIME could be considered as socio-demographical variables but are related to credit and to the company offering the loan are thus represent behavioral variables.

4 Methodology and analysis

This section provides justification and describes the methods I am going to use. Firstly, I rationalize the choice of logistic regression from the broad spectrum of different techniques. Similar to Kočenda and Vojtek (2009) I then define backward stepwise as empirical method information values to determine the most significant variables. At the end of this chapter I present the means of determining the quality of the model.

4.1 The most employed techniques

To study determinants of default the three most common predicting methods in the previous literature have been discriminant analysis (DA - also known as linear discriminant analysis LDA), logistic regression (LR) and linear regression (usually OLS). The following – both parametric and non-parametric - techniques have also been used in CSM: neural networks (NN) and classification trees (CT). The following section briefly summarizes some of the previous literature in the area of these techniques.

Arminger et al. (1997) have compared three different methods concerning credit risk. Those are LR, CT analysis and NN. In the study they used sex, job duration, age, car ownership, telephone ownership and marital status as predictor variables. The large dataset from a major bank in Germany specializing in consumer loans is divided into two subsamples to compare the techniques: the cross validation sample and the test sample. They report that predictive power is about equal for all techniques with the LR as the best technique. First, the performance of each method is analyzed by means of a test sample. The results for the CT analysis are slightly worse than the LR and NN. The results for the performance of the NN in cross validation sample are similar or slightly improved compared to the results of the test sample, whereas the two other technique have a slightly lower performance compared to the test sample.

Desai et al. (1996) compares NN, LDA and LR in building credit scoring models in the credit union environment and defining the predictive power of each model. They used data from

three different credit unions[34] in the Southeast United Stated for the period 1988 through 1991. The variables employed can be found in Table 1. The results are ambiguous: NN provides good estimate if the measure of performance is percentage of bad loans correctly classified. If the measure of performance is percentage of good and bad loans correctly classified, LR models are comparable to the NN approach. In any case LR does better than LDA. No attention to the significance of variables was given in this study. By comparing LDA, LR, NN and neural discriminant model Lee et al. (2002) found similar results: all four models provide on average the same classification rate between default and non-default customers.

Kočenda and Vojtek (2009) estimate determinants of default via parametric and non-parametric techniques, LR and CT. Both methods give reliable results and they state that non-parametric model i.e. CT can also be successful and able to create good models. Although literature indicates that also other techniques like CT and NN prove to have good estimates, there is a lot of evidence (Luo and Lei, 2008 and Yang et al., 2009) that logistic regression is very successful and often the best in determining default predictors and default probability.

According to Hand and Henley (1997) there is no overall best method. What is best will depend on the details of the problem: on the data structure, the characteristics used, the extent to which it is possible to separate the classes. Amarnath has prepared a short summary of the techniques. As well as Hand and Henley (1997) Armarnath's opinion is not to classify these but to consider each case separately. He describes that classification methods such as LR, nearest neighbour and tree-based methods are easy to understand and are thus appealing to the users. Amarnath suggest that neural networks are well suited to situations where we have a poor understanding of the data structure.

Several studies (Chen & Huang, 2003, Lawrence & Arshadi, 1995 and Laitinen & Laitinen, 2000) stand for logistic regression due to its high predictive power as an empirical model. Thomas (2000) as well as Kočenda and Vojtek (2009) have used logistic regression while analyzing credit defaults. It seems to be a method that is very successful when determining low and high-risk loans. Logistic regression is the most common technique for predicting default. There is a lot of critique against logistic regression (as well as LDA) due to the fact

[34] 962 observations for credit union L, 918 observations for credit union M, and 853 observations for credit union N.

that unlike OLS regression, logistic regression does not assume linearity in relationship between the independent variables and the dependent does not require normally distributed variables. However, several studies like Chen and Huang (2003) show that most of the credit scoring datasets are only weakly non-linear and thus give an appropriate estimate. It has been noted (Aldrich and Nelson, 1984, Lo, 1986 and Wilson et al., 2000) that the use of logit or probit estimators is more efficient compared to estimation based on DA. If the performance measure is the percentage of good and bad loans accurately classified, LR is as good as NN (Chen & Huang, 2003). The percentage of bad loans correctly classified in an important performance measure for CSMs since the cost of granting a loan to a defaulter is much larger than that of rejecting a good applicant.

It can be seen that all of these predicting methods have their own special features. By far the dominant methodologies, in terms of JBF[35] publications has been LDA followed by LR (Altman & Saunders, 1997). Martin (1977) uses both LDA and LR to predict bank failures in the 1975-76 period, when 23 banks failed. He found that both models gave similar classifications in terms of identifying failures and non-failures. However, one of the basic underlying assumptions in LDA is the assumption of normally distributed variables, which is violated in this case as most of the variables used in a CSM are categorical variables.

4.2 Logistic regression

Justificated by the facts presented in the previous section and continued in this section I decided to select logistic regression in this study. Logistic regression[36] is a parametric approach and a type of predictive model where the response variable is dichotomous and thus able to have only two exclusive values (usually coded as 0 or 1) – in this case default or non-default. The explanatory variables will have values that are either continuous or categorical. Logistic regression estimates the probability of a certain event occurring by fitting data to a logistic curve. The technique applies maximum likelihood estimation after transforming the dependent into a logit variable. Thus, logistic regression estimates the probability of a certain event – here: default - occurring. Logit analysis uses a set of variables to predict the probability of borrower default, assuming that the probability of default is logistically

[35] JBF stands for Journal of Banking and Finance.
[36] Logistic regression is also called logistic or logit model.

distributed i.e. the cumulative probability of default takes a logistic functional form and is constrained to fall between 0 and 1. The advantage in logistic regression is that the coefficients of explanatory variables are able to describe directly their predictive power and importance in the model.

Unlike ordinary linear regression, logistic regression does not assume that the relationship between a response variable and explanatory variables is a linear one. Even if logistic regression is considered as a generalized linear model, it is used for binomial regression only. Nor does it assume that the response variable or the error terms are distributed normally (see Ohlon, 1980 and Altman and Sabato, 2007).

The goal of the logistic regression is to predict the category of outcome for individual cases and to find the best fitting model to describe the relationship between the response variable and explanatory variables: explanatory variables are to predict changes in the response variable. In this study a model is created that includes only explanatory variables that are useful in predicting default.

The probability p that a loan will default given the predictors is the computed with the logistic distribution:

$$\text{logit}\,(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + x_{i1}\,\beta_1 + \ldots + x_{in}\beta_n$$

where p is the probability that p = 1 (loan will default), β_0 is the regression constant and $\beta_1 \ldots \beta_n$ are the regression coefficients which are to be estimated from the data. x_i are the explanatory (categorical or dummy) variables (predictors).

4.2.1.1 Odds ratio

With odds ratio I am able to define whether the probability of a certain event is the same for two groups. Odds ratio is a relative measure of risk, describing how much more likely it is that someone in one group (defaulted) is exposed to the factor compared to someone in

another group (non-defaulted). It can also be understood as a measure of effect size, describing the strength between two binary data values. Odds ratio is defined as

$$\text{Odds}_i = \left(\frac{defaulted_i}{defaulted}\right)\left(\frac{non-defaulted}{non-defaulted_i}\right)$$

Where *defaulted* and *non-defaulted* are the total number of defaulted and non-defaulted and $defaulted_i$ and $non\text{-}defaulted_i$ are the number of defaulted and non-defaulted observations in the i:th category of a variable. The odds of an event occurring is the probability that the event will occur divided by the probability that the event will not occur. An odds ratio of 1 implies that the event is equally likely in both groups and that the variable is not able to discriminate between defaulted and non-defaulted. An odds ratio greater than one implies that the event is more likely in the first group.

The odds ratios for each category of variables from the dataset can be found in the Appendix (Table 14).

4.2.2 Information value

Odds ratios as well as information value plays an important role in logistic regression. Both factors show the degree of the ability of the variable to discriminate between defaulted and non-defaulted loans.

By the means of odds ratios calculated in previous section I have computed the information values for each category. The total information value for each variable was then computed by summing up the category values. The information value analysis describes the information value of a variable and is defined as

$$\text{IV}_i = \ln\ (\text{Odds}_i)\left(\frac{defaulted_i}{defaulted}\right) - \left(\frac{non-defaulted_i}{non-defaulted}\right)$$

It tells us what is the predictive power of each variable. The higher the information values the higher the predictive power of the variable in the certain category. I am employing the

information value to exclude those variables from the analysis that have the lowest information values and thus small predictive power. Logistic regression gives the best results if the number of variables is not too large. Kočenda and Vojtek (2009) consider 20 variables to be the highest number of variables to employ. With information value evaluation I am able to select variables with some predictive power to the analysis. According to Kočenda and Vojtek information value above 0,2 is taken as a sign of the strong predictability of a variable in banking practice.

Information values for the categories of variables can be found in the Appendix (Table 14). Variables that have the most predictive power are INCOME, PREVLOAN, SCORE, MOVING and AGE[37]. Unlike findings of Kočenda and Vojtek (2009) and Anderson (2007) no clear division between socio-demographic and behavioral importance can be done. However, behavioral variables PREVLOAN and SCORE as two of the three most predictive variables have high predictive power as behavioral variables.

Furthermore, even if the information values are not high compared to banking practice and to the results of Kočenda and Vojtek (2009), they behave logically. SCORE is obviously one of the most predictive variables as it already includes the information of other variables. Consistent with findings of Kočenda and Vojtek, and Özdemir (2004) the socio-demographic variables have on average significantly lower information values than those who characterize the relationship between the lender and the customer.

In my analysis I decided to employ variables with information value higher than 0,01 and not for example 0,02 or 0,05 so that I could include more socio-demographic variables into the analysis. Thus PHONE, EMPLOYMENT, FREEEMAIL, TIME, MONTHLY, COTTAGE and CITY were removed from the model as insignificant variables.

4.3 Forward and backward stepwise selection

The optimal set of explanatory variables is usually obtained with a forward or backward selection. The idea behind backward selection is to compare the likelihood of a model when

[37] As the dataset consisted only of categorical variables the results might change substantially if the categories were reconsidered.

one variable at a time is included or excluded. Kočenda and Vojtek (2009) as well as Hand and Henley (1997) and Steenackers and Goovaerts (1989) suggest using stepwise selection to select characteristics to use in CSM.

Stepwise regression is a method where some of the variables are eliminated from the full model to achieve better suitability. Forward stepwise method sequentially adds variables to maximize the model's predictive accuracy. The fit of the model is tested after addition or elimination of specific variables to ensure the model still fits the data. At each step, the variable that leads to the greatest improvement in predictive accuracy – in terms of the highest score statistic conditional upon a significance level of less than 5% can be found. When no more variables can be added to the model or eliminated from the model, the analysis is complete.

Regarding the consumer credit market this method may be relevant due to the costs related to data collection. The application form may be considered to be too broad and time consuming and it could be improved by selecting only the most critical variables. This could increase the competitiveness against competitors.

The forward stepwise analysis begins with having first a model with a constant only, which is followed by adding variables one by one. In this thesis I decided to include variables whose information value is above 0,1. As conducted forward stepwise selection I ensure the backward stepwise selection gives the same results.

When having several variables it is critical to extract the results with more than one model. The selection of a model is done based on the Akaike Information Criterion (AIC) to employ forward-backward stepwise model selection. AIC is discussed in section 4.4.1.

4.4 Quality of the model

Due to the ordinal -instead of cardinal- nature of explanatory variables, I am not able to take correlation into account. Instead, the quality of the model and goodness-of-fit of a logistic regression is often measured with the three tests introduced next.

4.4.1 Akaike Information Criterion (AIC)

The AIC is a way of selecting a model from a set of models. The chosen model is the one that minimizes the Kullback-Leibler distance between the model and the truth. It's based on information theory, but a heuristic way to think about it is as a criterion that seeks a model that has a good fit to the truth. I start with the simplest model, with a regression on a constant only, which is a common procedure. From the first model each variable is left out one by one. After each step the model is tested with information criterion. Whether the information criterion improves it is safe to include more variables. The procedure is continued in a way the AIC does not worsen critically. AIC is defined as

$$AIC = - 2 (\ln (L)) + 2 K$$

where L is the maximized likelihood function for the estimated model and K is the number of free parameters in the statistical model.

In all my models I employed only variables with information above 0,01. If I would include more variables the AIC would increase more than 200%. I first estimate model 1, which is the output of forward stepwise technique, restricted the p-value to be between 0,05 and 0,1 for variables included in the stepwise procedure. The score for each customer can be calculated by summing the respective coefficient values, where the coefficient has a value of 0 for reference category. In addition to the first model, I construct two other models to define how the information criterion changes due to high number of categories in several variables. This leads to high number of degrees of freedom. Insignificant variables or meaningless coefficients are also reasons to construct additional models.

4.4.2 Log-likelihood ratio (LR) test

LR test is often used as a substitute for a standard F-test. The F-test is usually employed in the cases of OLS regressions but cannot be used in this study due to the fact that the response variable is not normally distributed. The LR test is performed by subtracting the residual deviances of constrained and unconstrained models. The likelihood ratio test uses the ratio of

the maximized value of the likelihood function for the full or alternative model over the maximized value of the likelihood function for the simpler or null model. The formula for the LR test statistic is

$$D = -2 (\ln L(m1)) - (\ln (L(m2)) = -2 \ln \left(\frac{L(m1)}{L(m2)} \right)$$

where L(m1) denotes the likelihood of the null model (here: Model 1) and L(m2) the likelihood for alternative model (here: Models 2 and 3, respectively).

4.4.3 Pearson Chi-Square test

In logistic regression chi-square test is the most appropriate in testing goodness-of-fit. Pearson Chi-Square test allows me to test the independence and goodness-of-fit of two categorical variables and are based upon a chi-square distribution. Chi-square is a statistical test commonly used to compare observed data with data we would expect to obtain according to a specific hypothesis.

Chi-Square is calculated by finding the difference between each observed and theoretical or expected frequency, squaring them, dividing by the theoretical frequency, and taking the sum of the results:

$$X^2 = \text{sum} \left(\frac{(O-E)^2}{E} \right)$$

where:

O = an observed frequency
E = an expected (theoretical) frequency, asserted by the null hypothesis.

5 Empirical results

This chapter gives justification for the selection of variables into the three models. After creating the models the results will be interpreted and the models will be tested with three tests of goodness-of-fit.

5.1 Model selection

The first model was selected as the ideal model using the forward and backward technique. For the first model I included all the explanatory variables that have information value above 0,01 so that not too many variables would be excluded in the beginning[38]. The estimates for the Model 1 are presented in Table 11 (Appendix), which also contains the list of variables used. Applying the forward stepwise method, 12 variables are included in the Model 1. The amount of variables is surely small enough to produce reliable results.

The total information values can be found in Table 14. It also contains the chosen variables in each of the models.

Model 1 has few weaknesses: firstly, some of the variables have insignificant coefficients and high p-values. Secondly, the initial model has a high number of degrees of freedom[39] concerning these specific variables. Due to these drawbacks, Model 2 is constructed.

There is no reason to increase information value above 0,01, to 0,02 or 0,05 since having 0,01 as a cut-off value will already drop 7 variables. AIC would decrease 10%[40] or 20%[41] when having 0,05 or 0,02, respectively, as cut-off values for information value selection. Decreases are desirable but at this point we do not know which variables are significant and how will they behave if we exclude some of the strongest predictors outside the model. In addition LEVEMPL gives irrational coefficients when it comes to group laid off. The variable is also

[38] To have 0,02 or 0,05 as the cut-off value, stepwise selection would exclude too many variables outside the model and include only 6 or 9 variables instead of current 12 variables.
[39] Model 1 has 46 degrees of freedom as the amount decreases to 21 in Model 2.
[40] From 10 270 to 9 222.
[41] From 10 270 to 8 309.

highly insignificant. Even if INCOME doesn't seem to give significant results I decided to employ it to the Model 2 since it has high information value and a explicit odds ratio. Model 2 is better in the sense that it has a lower number of variables and thus may give more reliable results. Therefore, I keep 0,01 as the cut-off value and exclude the weakest variables LEVEMPL, PAYBACK and POSTAL from Model 2 to decrease the number of degrees of freedom. The coefficients of the Model 2 can be found in Table 12 (Appendix).

Finally, I estimate Model 3. The justification for the third model is driven by the fact that INCOME, PREVLOAN and SCORE are very strong default predictors compared to other variables. Therefore it is important to analyze the properties of other variables and define what is the capability of the model without strongest predictors. By excluding the strongest behavioral variables PREVLOAN and SCORE I can analyze the socio-demographical variables more carefully. Further, the SCORE is connected to every other variable in the model hence it is initially built based on other variables and thus may be highly correlated with those. Therefore it is interesting to see whether it is possible to discriminate successfully without the knowledge of how much scores the customer has received. The justification to have INCOME included in the model is that it is a socio-demographical variable and dropping it from the model would increase AIC substantially[42]. I employ all the other variables that have information value higher than 0,01 and ones that were included in the Model 1.

The coefficients of the Model 3 are presented in Table 13 and suggest that Model 3 is able to discriminate among customers without knowledge of the scores the customer had and whether he has had a loan from the company previously.

It turns out that excluding PREVLOAN and SCORE the AIC decreases against Model 1 but increases against Model 2[43].

[42] Dropping INCREASE, PREVLOAN and SCORE would let AIC to increase to 15 920 while dropping only PREVLOAN and SCORE AIC decreases to 9 633 compared to the Model 1.
[43] To exclude only PREVLOAN from Model 1 the AIC reaches 8 919 while excluding SCORE the AIC increases to 10 250.

Table 5: Information values for variables

Variable	Information Value	Model 1	Model 2	Model 3
INCOME	0,400923	x	x	x
PREVLOAN	0,251429	x	x	
SCORE	0,145415	x	x	
MOVING	0,113234			x
AGE	0,106136			x
SIZEHOUSEH	0,092787			x
HOUSING	0,088212	x	x	x
NRADULTS	0,074267			
GENDER	0,072071	x	x	x
HOUSINGTYPE	0,062405			x
MARITAL	0,061146			
REPAYMENTBEH	0,058092			
NRCHILDREN	0,050342			
PAYBACK	0,042453	x		x
LEVEMPL	0,031372	x		x
CREDIT	0,029561	x	x	x
LOANSIZE	0,028570	x	x	x
TIMEFIN	0,013839			
EDUCATION	0,012677	x	x	x
NATIONALITY	0,012487	x	x	
NATIVE	0,011350			
POSTAL	0,011152	x		
MILITARY	0,010219			
CITY	0,008176			
COTTAGE	0,007222			
MONTHLY	0,006397			
TIME	0,004698			
FREEEMAIL	0,001536			
EMPLOYMENT	0,001160			
PHONE	0,000567			

5.2 Variable interpretation

Statistical analysis was performed using SPSS 17. Tables 11, 12 and 13 (Appendix) shows the results of logistic regression with Wald significance[44]. I now turn to assessing and interpreting the results.

Since I model the probability of default, a higher coefficient reflects a higher default probability. The score for each client can be calculated by summing the respective coefficient values, where the coefficient has a value 0 for reference category.

5.2.1 Socio-demographical variables

Similar to Arminger et al. (1997) gender seems to be one of the strongest indicators of default. In all three models GENDER is a significant variable showing that female customers have much less difficulty in paying their debts and seem to default less than man. According to odds ratios the customers who haven't served in the military tend to fail less often to repay the loan on time. This is logical as women present 44% of the population in this specific group. According to Suomen Asiakastieto (2009) the proportion of men and women among those who default has remained the same for over ten years. However, no generalization can be made since this study presents only the determinants in consumer credit market as Suomen Asiakastieto reports the ratio related loan markets in general.

Another strong predictor is CREDIT; customers who have one or more credit cards tend to default less compared to those who do not have one. Customers who have a credit card have been evaluated by a credit institution and hence provide some kind of guarantee for the consumer credit company. Nowadays it might be difficult to obtain a credit card and many credit card providers require the customer to be employed, have a regular income and to not have default information in Suomen Asiakastieto's register. According to Dunn and Kim

[44] A Wald test is used to test the statistical significance of each coefficient (b) in the model. A Wald test calculates a Z statistic, which is $z = \frac{B}{SE}$. This z value is then squared, yielding a Wald statistics with a chi-square distribution.

(1999) the ownership of one credit card compared to not having a card decreases the risk to default but as having two or more credit cards the risk to default increases substantially.

In my analysis INCOME is not as significant as expected. This is consistent with findings of Warren (2002): the median American who files for bankruptcy comes from the middle-class and is not categorized as a poor. Other variables like education, occupation and home ownership matter instead, according to her study. However, in Models 1 and 2 I can draw the conclusion that customers who have the highest income class and earn 2500 euros per month or more tend to default less. Interestingly, it seems that income class <1000 euros tend to default less than those who earn more, anyhow below 2500 euros. This might be a consequence of customers announcing their income to be higher as it is in reality. Therefore, the scoring model takes the dishonesty into account and revises the scores, which damages the granting likelihood of customers who have announced the real income. According to Vasanthi and Raja (2006) income of a customer is the most important socio-demographic characteristic. Evidence from their study shows that the main cause for mortgage default is a fall in household income. Customers who earn more than 2 500 euros per month seem to have a smaller default rate than those with lower level of income. According to Van Order and Zorn (2000) loans in low- and moderate income borrowers default more, but not by a very large factor.

As expected, NATIONALITY show predictive power in Model 1 suggesting that customers who are Finnish citizens tend to default less compared to those who have other citizenship. This is consistent with findings of Steenackers and Goovaerts (1989).

POSTAL is included in Model 1 but does not give reliable indication of default. This is presumable as each category includes both large and small cities. This variable also has a high number of degrees of freedom due to several categories and thus might result in providing insignificant results. However, I can draw the conclusion that customers living in Helsinki metropolitan area (postal code 00000-09999), Turku (20000- 29999), Tampere (30000-39999) and Pohjois-Karjala and Pohjanmaa (80000-89999) tend to default more than customers from other areas.

LEVEMPL doesn't seem to be a significant predictor in the Model 1 or Model 3 but may result from the high number of degrees of freedom. The coefficients imply that every other

category is more likely to pay their loans at a time compared to the reference category, which is agriculture entrepreneur. One exception there is, however, and it is students. This is consistent with findings of Autio et al. (2009): students and other young people recognize flaws in their money management. One can make the assumption that both agriculture entrepreneurs and students have irregular income and a seasonal working time and thus may have troubles paying their loans back. Group laid off has a coefficient of -21,17, which is highly irrational and insignificant and excluding it from the Model 2 is thus justified.

As expected, HOUSING is partly a strong predictor of default in all models. It seems that customers who have a home of their own default more often than those who rent. It is not surprising that a home via employment relationship or partial ownership are not significant categories as they are quite uncommon among the customers and cannot differentiate customers. This finding is consistent with the study of Agarwal et al (2009): an individual who owns a home is significantly less likely to file for bankruptcy compared to those who do not have real estate of their own.

Though the variable EDUCATION is not a significant predictor, its B coefficients have the expected signs. Consistent with findings of Kočenda and Vojtek (2009) education level is a significant predictor of default. Customers with a higher level of education have much less difficulty paying their loans. People with university degrees seem to be less risky than those with lower education: primary school, technical school or high school. The default rate seems to be highest when it comes to primary school. Also Burrows (1998), Vandell and Thibodean (1985) and Quercia and Stegman (1992) found similar results from the mortgage industry: households who defaulted their mortgage were often in an unskilled manual occupation, or were uneducated persons who are unemployed most of the year. There are also empirical studies that differ from these (see for example Mills and Lubuele, 1994): households who are skilled had performed equally well compared to uneducated ones.

When excluding the most important behavioral variables PREVLOAN and SCORE, new socio-demographical variables emerges into the analysis. Those are MOVING, AGE, SIZEHOUSEH and HOUSINGTYPE.

Consistent with early literature MOVING is a relatively important variable showing that the longer the time since last moving the less likely the customer is to default. Agarwal et al.

(2009) found that a borrower who migrates from his state of birth is 17 percent more likely to default, while a borrower who continues to live in his state of birth is 14 percent less likely to default.

In Model 3 AGE suggest that age group 61-70 year-olds are more likely to pay their credit while 20-25-year-olds tend to default more often. This is consistent with the statistics of Suomen Asiakastieto (2010): the age group of 25-49 year-olds had the most payment troubles. Most of these were due to consumer credit. The majority of notes from instant loans were from age group 18-24. The oldest group also seems to have more payment troubles but is insignificant. Several studies (Dunn & Kim, 1999, Jacobson & Roszbach, 2003 and Agarwal, 2009) confirm my finding of age being a determinant of default. The results suggest that customers of age 46-60 (and 26-45) default less than age group 20-25. Significant changes in odds ratio (see table 14) show that AGE is successful in discriminating between defaulted and non-defaulted especially in age group 20-25.

SIZEHOUSEHOLD suggest that households with two or four members default the least. Interesting would be to study whether the households of three members include two adults and one child or one adult and two children and thus default more often. Variables NRADULTS, NRCHILDREN etc don't seem to have significant predictive power or significance in any of the models, which suggest that a combination of these three variables should be employed to find out how the results differ.

As expected the type of housing seems to matter, according to HOUSINGTYPE. Those living in a house tend to default less than those living in a row house or an apartment. Houses are typically more expensive both to own and to rent and thus indicated of having more wealth.

5.2.2 Behavioral variables

Similar to findings of Kočenda and Vojtek (2009), LOANSIZE show that small loans appear to be more risky in all three models.

As predicted, SCORE behaves logically: the higher the SCORE, the more likely it is for customers to perform well compared to those who have lower SCOREs. As predicted, the

SCORE is a relatively important variable. Customers who are given the highest scores tend to default less than those with fewer scores.

PREVLOAN is also a strong predictor and highly significant (at 90% confidence level) showing that those who have had a loan from the company before tend to fail their payments compared to new customers. From the company's point of view it is critical to monitor the payment behavior of its current customers in order to prevent them from defaulting later. This is incoherent with the results of Kočenda and Vojtek (2009) who show that the longer the relationship between the customer and the lender the more likely it is for the customer to pay back. One explanation is obviously the fact that the company I'm studying doesn't operate as a bank and thus can't follow the customers' account behavior. According to Autio et al.:s (2009) study those young people who take consumer credit once are likely to do it again. Inconsistent with findings of Dinh and Kleimeier (2007) default is more frequent for repeat borrowers. However, my analysis does take stance on the number of previous loans per one customer.

PAYBACK is partly significant suggesting that customers who have chosen to pay their loan pack in 18 to 48 months seem to default more than those who have chosen to repay within 12 months. A conclusion is that those who perform well with their loan need a short-time financing to survive sudden costs but are solvent in long-term.

5.3 Quality quantification

Next I will compare the quality of the three models using Akaike Information Criterior test, Log Likelihood ratio test and Pearson Chi Square test. I have also included the classification tables at the end of this section.

5.3.1 Akaike Information Criterior (AIC)

Table 6: AIC test

	Akaike Information Criterior (AIC)
Model 1	10 270
Model 2	6 370
Model 3	9 633

As discussed earlier, AIC decreased substantially to Model 2 when I excluded the variables with insignificancy and with high number of degrees of freedom. When excluding PREVLOAN and SCORE from Model 3 AIC reaches to 9 633 again. A conclusion can be drawn that it is not recommendable to exclude the most significant variables due to AIC but is needed, however, to increase the importance of some socio-demographical variables. As can be seen from Table 6, Model 2 is best model to discriminate between good and bad customers according to AIC.

5.3.2 Log Likelihood Ratio Test

Table 7: LR test

	Log Likelihood	DEV
Model 1	-5 072,30	10144,60
Model 2	-5 112,29	10224,58
Model 3	-5 212,99	10425,98

The LR test was performed by subtracting the residual deviances of constrained (both Models 2 and 3) against unconstrained models (Model 1) using the log-likelihood ratio test (LR test). When comparing Model 1 with Model 2 the test statistics is LR = 79,98 with 25 degrees of freedom, and statistics comparing Model 1 with Model 3 is LR = 281,38 with 4 degrees of freedom. Both ratios are highly statistically significant. According to LR test, the power of all the models is approximately the same and thus all the models can be used as a CSM.

5.3.3 Pearson Chi-Square Test

Table 8: Pearson Chi-Square test

	Pearson Chi-Square	df	Sig.
Model 1	918,61	46	0,000
Model 2	838,61	21	0,000
Model 3	637,22	50	0,000

According to Pearson Chi-Square Test all the Models present have efficient results. The significance is 0,000 and the magnitude of Chi-Square is high for all, being highest for Model 1.

5.3.4 Classification Tables

Table 9: Classification tables

			Predicted		
			Non-default	Default	Percentage Correct
Model 1	**Observed**	Non-default	4001	957	80,7%
		Default	1920	1353	41,3%
		Overall Percentage			65,0%

			Predicted		
			Non-default	Default	Percentage Correct
Model 2	**Observed**	Non-default	3956	1002	79,8%
		Default	1945	1328	40,6%
		Overall Percentage			64,2%

			Predicted		
			Non-default	Default	Percentage Correct
Model 3	**Observed**	Non-default	4094	864	82,6%
		Default	2157	1116	34,1%
		Overall Percentage			63,3%

From the classification tables can be seen that goodness-of-fit is adequately good for all three models. The performance of correct predictions by category suggest that good credit risks (non-default) are more likely discovered whereas the performance in the bad credit risk (default) is significantly lower suggesting the Model 1 to give most reliable results. Hand and Henley (1997) got bad risk rate to vary between 43,09% and 43,77% differing between five methods. The overall performance is about equal for all models being slightly lower for

Model 2 and Model 3. I can draw a conclusion that all three models can be employed as CSM due to the fact that they provide similar percentages and better fit[45] than a random model.

5.3.5 Comparing models

Table 10: Model comparison

Test	Model 1	Model 2	Model 3
AIC		x	
LR		x	x
Chi Square	x		
Classification tables	x		

According to model comparison there are no substantial differences between models and thus all three models could be employed to have a reliable CSM. However, as PREVLOAN and SCORE indicate high predictive power it might not always be recommendable to exclude those from the model. The risk management department and management of a lender are to decide which model is the most suitable for the purposes of that particular credit institution.

[45] Assuming a random fit provides 50 percentages fit when there are two binary alternatives.

6 Conclusions

This is the final part of my thesis, which will finish the thesis by summarizing the results and concluding the study by giving theoretical and managerial implications. In the end of the chapter I will also give suggestions for further research in the area of consumer credit and default predicting.

6.1 Summary of the Research

This paper has empirically investigated determinants of consumer credit default with a new set of survey data taken from May 2008 through September 2009. This data set contains a representative sample of more than 14 500 individual consumer credit accounts that have not previously been available in Finland. Consumer credit default is examined in a logistic regression analysis where the number of defaulted payments is fitted to key behavioral aspects of company and a variety of socio-demographical variables.

The data set included unique and sensitive information such as income, age, education and marital status reported by a well-known Finnish consumer credit company. In total I disposed of 31 variables. Table 2 contains definitions for the variables that have been selected for the estimation of the empirical model in Chapter 3. Of the 31 variables, 15 were not used in any of the final models. Most were disgarded because they lacked a univariate relation with response variable.

To improve the first analysis I form two additional models and control some of the weakest as well as the strongest variables. I also test the efficiency of these methods, report the most important determinants of default behavior and compare these models in terms of the power in discriminating between “good” and “bad” customers.

I obtain interesting key sets of results. First of all, the uncontrolled model show evidence that both socio-demographical and behavioral variables have predictive power. I find the most significant determinants: the ownership of a credit card, level of education, gender, housing status, income and nationality. Also the behavioral variables, size of a loan, previous loan and

score, seem to have notable effect on default. For the Model 2 I controlled the three most insignificant variables and found that it is possible to reach significantly lower AIC with fewer variables. In Model 2 the significant predictor variables remain the same. As the behavioral variables, previous loan and score are highly strong, I control these and include the rest of the variables in the third model. It can be seen that while controlling the two strongest variables, other socio-demographical variables emerge to the model. I find that also the time since last moving, age, size of household and housing type have predictive power in the analysis.

One of the contributions of this thesis is that in terms of a logistic regression model I identified a specification that does not contain the two most important financial variables (PREVLOAN and SCORE) but still performs only marginally worse than the initial model.

This paper contributes to the growing literature in several ways. I find that both socio-economical and behavioral variables have predictive power and can thus be considered as determinants of default. The predictive variables show similar evidence as previous studies. Updegrave (1987) as well as Steenackers and Goovaerts (1989) and Dunn and Kim (1999) and Agarwal et al. (2009) show similar results regarding income and age. Housing ownership has seen to be predictor of default (Steenackers & Goovaerts, 1989 and Agarwal et al. (2009). Relationship with the lender has often seen a reliable variable according to Rock (1984). He also found that income and housing ownership could be considered as predictive variable, which is consistent to this study. Similar to this study, Kočenda and Vojtek (2009) have shown education, amount of loan and number of scores to be relatively important when studying default. Arminger (1997) show evidence on gender being one of the most important predictors of default.

Kocenda and Vojtek (2009) as well as Özdemir (2004) suggest that financial or behavioral rather than socio-demographical variables have more influence on customer's payback performance. This is consistent with findings of this thesis. However, only two of the behavioral variables show significant predictive power and thus also the need for socio-demographical variable is justified. Neither type of variable group alone is adequate.

6.2 Theoretical and managerial implications of the research findings

Combined with a dramatic growth of consumer credit and increased regulartory attention to risk management, the development of a reliable CSM is essential.

This study assumes that customers who haven't defaulted during the first year will perform well also during the rest of the loan period (following three years). However, the default rate regarding for only the first year is already 29%, which means that it could reach higher rate during the next years. Sullivan et al. (2001) show evidence of the two major causes of the increase in default and bankruptcy filings in the U.S.: increases in credit card and mortgage debt; an unexpected adverse events (such as unemployment, divorce, health problems, or medical debts) have reduced the ability of households to repay their debt and eventually compel them to file for bankruptcy. To have a reliable CSM is critical, as the competition between financial institutions has come to a tightened stage. According to Allen et al. (2004) banks that use CSMs appear to be more productive at lower costs. Companies are seeking better strategies with the help of improved credit scoring models. As managerial implications the scoring system could be reconstructed and updated regularly. The cut-off value could be raised hence the data set of this study show that as much as 40% of the customers in the lowest score class defaulted.

Through optimizing the lending activity and minimizing the costs of default the lenders would be able to provide smaller interest, which would in turn improve the general performance compared to competitors. As stated in the introduction chapter, the morality of consumer credit has been out in the air. The real annual interest rate can reach up to hundreds of percentages. By means of more accurate CSM the default rates could be brought down and this would in turn make it possible for lenders to decrease the high interest rates related to consumer loans. The Federal Reserve Board (2007) in the U.S published a study that noted that CSMs have increased the availability of credit and reduced the cost of credit. Brill (1998) document that creating and improving CSM has following benefits: cost reduction in credit analysis, faster credit evaluation, closer monitoring of existing accounts and improvement in cash flow and collections. Chen and Huang (2003) state that with sizable loan portfolios, even

a slight improvement in credit scoring accuracy can reduce the creditors' risk and translate considerably into future savings.

In order to have several variables in the scoring model it requires the online application to also be broad. From the customers' perspective this may be time-consuming and frustrating. As shown in this study it is not necessary to use all the 30 explanatory variables to find the probability of default of a customer. To be able to create a model with fewer variables not only the costs would decrease from paying to the CSM vendor but also the application would become more customer-friendly.

Many of the previous studies have employed the same variables as do this thesis. However, there are some additional variables that could be also investigated and applied in this thesis and in the company in question. Especially financial or behavioral variables could be added. The purpose of the loan is studied by Dinh and Kleimeier (2007) and Lieli and White (2008). Loan purpose might indicate the consuming behavior and financial situation of a borrower. The length of relationship with the credit institution is difficult to observe as the company does not operate currently as a bank and thus does not have savings accounts for the borrower but could already be calculated from the time of applying the first loan. In the context of relationship banking, it can be assumed that the longer the relationship, the more the company knows about the customer and the lower the default risk becomes. Peltoniemi (2004) studied the role of relationship banking between small business firms and both bank and large financial institution in Finland. He found that a longer relationship tends to lower the cost of the credit and a long-lasting bank-firm relationship is beneficial, especially to high-risk firms. Number and duration of outstanding loans (including car loan, mortgage and home equity credit) – both within the company and from other financial institutions – might reflect the indebtedness of a customer. On the other hand, this would also describe the trustworthiness of customer in other banks: if borrower is able to have loan from a bank and thus might need to place collateral, he is likely to have credit standing and is thus less likely to default. Agarwal et al. (2009) found that borrowers with higher other debt are significantly less likely to default. The amount of resources the client has would be an interesting variable to study but like length of relationship requires a bank account.

Some socio-demographical variables are worth noticing also. Rock (1984) suggests using debt-income ratio as predictor variable. Another variables that could be employed are

residential and work duration, city of birth or moving from the city of birth, Arminger et al. (1997) found car ownership and employment duration to affect on default. Time with employer matters because it might reflect the borrower's job satisfaction, the more stable his employment will be and the higher his ability to repay his loan. Time at present address might be a proxy for the borrower's maturity, stability or risk aversion but also a signal that a borrower's financial wealth has improved.

Early literature shows evidence of attitudes affecting on default rates (see Banasik & Crook, 2007 and Chiang et al., 2002 and Roberts & Sepulveda, 1999 and Hayhoe, 1999). Consumers' attitude on credit and default is difficult to measure but could be highly significant regarding the default rates.

In this thesis I do not have any specific information of the timing of default; hence I have only the division of "non-default" and "default". However, to investigate the timing of default is highly significant regarding the performance of the companies (see Roszbach, 2004). According to Boyes et al. (1989) it would be necessary to study population survival times more hence improved estimates of time-to-default will increase the ability to measure expected earnings.

As it is obvious for some customers, especially those with low credit scores, to default more, it could be profitable to differentiate the interest rates between customer segments. People with higher credit scores would pay lower interest rates. In the light of these results consumer credit companies are able to target the marketing to those customer groups that are less likely to default. Bertrand et al. (2010) show that advertising contents significantly affects demand in the consumer credit market[46]. Bofondi and Lotti (2006) show supportive evidence: banks that are able to exploit scale economies and have larger market shares operating in more concentrated markets turn out to be early adopters in finding the most reliable CSMs. Today the loan market is highly competitive and especially in banks customers often have the same

[46] Bertrand et al. (2009) used a large-scale direct-mail field experiment to study the effect of advertising content on real decisions of applying for consumer credit in South Africa. An advertisement was sent to 53 000 former clients of a company with variation in advertising content. Results suggest for example that showing fewer example loans, not suggesting a particular use for the loan, or including a photo of an attractive female increase loan demand by about as much as 25% reduction in the interest rate.

margin even if some are riskier than others. With default predictive models lenders would be able to price the customers based on their risky socio-demographic characteristics.

In response to increased number of defaults and bankruptcies companies should move away from only analyzing the credit-risk of individual loans and securities towards developing measures of credit concentration risk such as the measurement of portfolio risk of fixed income securities (Altman & Saunders, 1997). An example of measuring risk differently provides a work of Musto and Souleles (2005) by measuring covariance risk of individual consumers to determine the determinants of default. Altman and Saunders (1997) show that their model for portfolio risk measurement is promising of estimating the optimal composition of loan portfolio.

6.3 Limitations and suggestions for further research

As data on defaulted and non-defaulted loans are collected from a portfolio that already passed a credit scoring procedure in the credit company, a CSM that is based only on these data gives biased results if it is used for the selection of new loans. Due to the selection bias it would be important to also take the customers who were rejected into the analysis. This could be conducted by classifying the rejected loans into “good” and “bad” rejected loans. Of course no guarantee will be given whether a rejected loan would have been good or bad.

Credit companies often use scoring models that tend to classify customers to be either ”good” or ”bad”. The problem here is that the scoring system doesn’t take into account the state of the economy. In following papers it could be interesting to study how default behavior changes when there are hits in the economy.

The dataset of this paper included categorical variables which were taken as given. In following studies it might be worthwhile to reconsider the categories or also include continuous variables if they are available and when it is appropriate. If the variables would be continuous also the correlation between variables would be determined. This might also have an impact on total information values possibly making them higher and causing odds ratios between different categories to be more discriminated. In this paper some of the insignificant variables had more than five categories, which leads them to have high number of degrees of

freedom[47] and the results to be often insignificant. Thus the number of categories could also be lowered. Data used in empirical analysis is usually categorical (see Hand & Henley, 1997 and Crook et al., 1983). However, the alternative approach of coding categorical variables into numeric form and using continuous data models is becoming more common. For example, one strategy is to use logarithms of likelihood ratios. Categories as such may not be informative enough. Instead variables could be combined in order to differentiate between customer types. For example education and age could be combined to compare if an educated 20-25-year-old is more likely to perform well than an unskilled person of the same age group. For example Dunn and Kim (1999) found that married customers' probability to default is lower if they do not have children but the likelihood to default increases with number of children.

In this study I modeled the probability of default. A higher score reflected a higher default probability. In general, in CSMs the higher the score the more likely the customer is to perform well. To be able to compare these models to the model the company uses I should scale the coefficients to match current categories.

There are not many academic studies that look to predict payment behavior, despite its importance in trade credit management decision-making. The approach in empirical studies of default has generally been to determine if socio-demographical and behavioral information can provide signals to predict default. Credit companies could predict default as Wilson et al. (2000) and investigate their payment behavior and not just their characteristics.

[47] Such a high number of degrees of freedom is implied by the fact that each class of categorized variable adds one degree of freedom.

References

Literature

Agarwal, S., Chomsisengphet, S. & Liu, C. (2009). Consumer Bankruptcy and Default: The Role of Individual Social Capital. Working Paper. Available at SSRN: http://ssrn.com/abstract=1408757.

Allen, L., DeLong, G. & Saunders, A. (2004). Issues in the Credit Risk Modeling of Retail Markets. Journal of Banking & Finance, Vol. 27, Issue 4, p. 727-752.

Allonen, H. (2010). Yritysten maksukyvyttömyyden ennustaminen kriisitaloudessa. *Master's Thesis*, Helsinki School of Economics.

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *Journal of Finance*, Vol. 23, Issue 4, p. 589-609.

Altman, E. I., Eom, Y. H. & Kim, D. W. (2007). Failure Prediction: Evidence from Korea. *Journal of International Financial Management & Accounting*, Vol. 6, Issue 3, p. 230-249.

Altman, E. I., & Saunders, A. (1997). Credit Risk Measurement: Developments Over the Last 20 Years. *Journal of Banking and Finance*, Vol. 21, Issue 11-12, p. 1721-1742.

Arminger, G., Enache, D. & Bonne, T. (1997). Analyzing Credit Risk Data: A Comparison of Logistic Discrimination, Classification Tree Analysis, and Feedforward Network. *Computational Statistics*, Vol. 12, Issue 2, p. 293-310.

Autio, M., Wilska, T-A., Kaartinen, R. & Lähteenmaa, J. (2009). The Use of Small Instant Loans Among Young Adults – a Gateway to a Consumer Insolvency. *International Journal of Consumer Studies*, Vol. 33, Issue 4, p. 407-415.

Banasik, J., Crook, J. & Thomas, L. (2003). Sample selection bias in credit scoring models. *Journal of the Operational Research Society*, Vol. 54, Issue 8, p. 822-832.

Bertrand, M., Karlan, D., Mullainathan, S., Shafir, E. & Zinman, J. (2010). What's Advertising Content Worth? Evidence from a consumer Credit Marketing Field Experiment. *The Quarterly Journal of Economics*, Vol. 125, Issue 1, p. 263-305.

Board of Governors of the Federal Reserve System (2007). Report to tge Congress on Credit Scoring and Its Effects on the Availability and Affordability of Credit. Submitted to the Congress pursuant to section 215 of the Fair and Accurate Credit Transactions Act of 2003.

Bofondi, M. & Lotti, F. (2006). Innovation in the Retail Banking Industry: the Diffusion of Credit Scoring. *Review of Industrial Organization.* Vol. 28, Issue 1, p. 343-358.

Boyes, W. J., Hoffman, D. L. & Low, S. A. (2002). An econometric analysis of the bank credit scoring problem. *Journal of Econometrics*, Vol. 40, Issue 1, p. 3-14.

Brown, S., Taylor, K. & Price S. W. (2005). Debt and distress: evaluating the psychological cost of credit. *Journal of Economic Psychology*, Vol. 26, Issue 5, p. 642-663.

Chen, M. C. & Huang, S. H. (2003). Credit Scoring and Rejected Instances Reassigning Through Evolutionary Computation Techniques. *Expert Systems with Applications*, Vol. 24, Issue 4, p. 433-441.

Claessens, S., Krahnen, J. & Lang, W. W. (2005) The Basel II Reform and Retail Credit Markets. *Journal of Financial Services Research*, Vol. 28, Issue 1-3, p. 5-13.

Crook, J. N., Hamilton, R. & Thomas, L. C. (1983). A Comparison of a Credit Scoring Model with a Credit Performance Model. *The Service Industries Journal*, Volume 12, Issue 4, p. 558-579.

Desai, V. S., Crook, J. N. & Overstreet, G. A. (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research*, Vol. 95, Issue 1, p. 23-37.

Dinh, T. H. T. & Kleimeier, S. (2007). A Credit Scoring Model for Vietnam's Retail Banking Market. *International Review of Financial Analysis*, Vol. 16, Issue 5, p. 571-495.

Dunn, L. F. & Kim, T. (1999). An Empirical Investigation of Credit Card Default. Working Paper, Ohio State University, Department of Economics, 99-13.

Gross, D. & Souleles, N. (2001). Liquidity Constraints and Interest Rates Matter for Consumer Behavior? Evidence from Credit Card Data. *The Quarterly Journal of Economics*, Vol. 117, Issue 1, p. 149-185.

Hand, D. J. & Henley, W. E. (1997). Statistical Classification Methods in Consumer Credit Scoring: a Review. *Journal of the Royal Statistical Society: Series A*. Vol. 160, Issue 3, p. 523-541.

Jacobson, T. & Roszbach. K. (2003). Bank lending policy, credit scoring and value-at-risk. *Journal of Banking & Finance*, Vol. 27, Issue 4, p. 615-633.

Jaffee, D. M. & Russell, T. (1976). Imperfect Information, Uncertainty, and Credit Rationing. *The Quarterly Journal of Economics*, Vol. 90, Issue 4, p. 651-66.

Kočenda, E. & Vojtek, M.. 2009. Default Predictors and Credit Scoring Models for Retail Banking. CESifo Working Paper, No. 2862.

Laitinen, E. & Laitinen, T. (2000). Bankruptcy prediction: Application of the Taylor's expansion in logistic regression. *International Review of Financial Analysis*, Vol. 9, p. 327-349.

Laitinen, T. & Kankaanpää, M. (1999). Comparative analysis of failure prediction methods: the Finnish case'. *European Accounting Review*, Vol. 8, Issue 1, p. 67-92.

Lawrence, E. C. & Arshadi, N. (1995). A Multinomial Logit Analysis of Problem Loan Resolution Choices in Banking. *Journal of Money, Credit & Banking*, Vol. 27.

Lee, T. S., Chiu, C. C., Lu, C. J. & Chen, I. F. (2002). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications*, Vol. 23, Issue 3, p. 245-254.

Lieli, R. P. & White, H. (2008). The Construction of Empirical Credit Scoring Models Based on Maximization Principles.

Luo, J-h. & Lei, H-y. (2008). Empirical study of corporation credit default proababolty based on Logit model.

Martin, D. (1977). Early Warning of Bank Failure: A Logit Regression approach. Journal of Banking and Finance, Vol. X, Issue X, p. 249-276.

Mester, L. (1997). What's the point of credit scoring? Federal Reserve Bank of Philadelphia Business Review, September/October, p. 3-16.

Musto, D. K. & Souleles, N. S. (2006). A Portfolio View of Consumer Credit. *Journal of Monetary Economics,* Vol. 53, Issue 1, p. 59-84.

Neophytou, E. & Charitou, A. (2000). Predicting Corporate Failure: Empirical Ecidence for the UK. Working paper, University of Southampton, Department of Accounting and Management Science, No. 01-173.

Peltoniemi, J. (2004). The Value of Relationship Banking. Empirical Evidence on Small Business Financing in Finnish Credit Markets. University of Oulu.

Rock, A.（1984). Sure Ways to Score with Lender, *The Accounting Review*, Vol. 81, p. 36-49.

Roszbach, K. (2004). Bank Lending Policy, Credit Scoring and the Survival of Loans. *Review of Economics and Statistics*, Vol. 86, Issue 4, p. 946-958.

Steenackers, A. & Goovaerts, M. J. (1989). A Credit Scoring Model for Personal Loans. *Insurance: Mathematics and Economics*, Vol. 8, Issue 1, p. 31-34.

Stiglitz, J. E. & Weiss, A. M. (1981). Credit Rationing in Markets with Imperfect Information. *American Economic Review*, Vol. 71, Issue 3, p. 393-410.

Straka, J. W. (2000). A Shift in the Mortgage Landscape: The 1990s Move to Automated Credit Evaluations. *Journal of Housing Research*, Vol. 11, Issue 2, p. 207-232.

Sullivan, T. A., Thorne, D. & Warren, E. (2001). Young, Old, and In Between: Who Files for Bankruptcy? *Norton Bankruptcy Law Adviser*, p. 1-11.

Thomas, L. C. (2000). A Survey of Credit and Behavioural Scoring: Forecasting Financial Risk of Lending to Consumers. *International Journal of Forecasting*, Vol. 16, Issue 2, p. 149-172.

Tsai, M. C., Lin, S. P., Cheng, C. C. & Lin, Y. P. (2009). The consumer loan default predicting model – An application of DEA-DA and neural network. *Expert Systems with Applications,* Vol. 36, Issue 9, p- 11682-11690.

Updegrave, (1987). How lender size you up. *Money*, p. 23-40.

Van Order, R. & Zorn, P. M. (2000). Income, Location and Default: Some Implications for Community Lending. *Real Estate Economics*, Vol. 28, Issue 3, p. 385-404.

Vasanthi, P. and Raja, P. (2006). Risk Management Model: an Empirical Assessment of the Risk of Default. *International Research Journal of Finance and Economics*, Vol. 1, Issue 1.

Warren, E. (2002). Financial Collapse and Class Status: Who Goes Bankrupt? *Osgoode Hall Law Review*, Vol. 41, Issue 1, p. 114.

Wilson, N., Summers, B. & Hope, R. (2000). Using Payment Behaviour Data for Credit Risk Modelling. *International Journal of the Economics of Business*, Vol. 7, Issue 3, p. 333-346.

Yang, Y., Nie, G. & Zhang, L. (2009). Retail Exposures Credit Scoring Models for Chinese Commercial Banks. Computer Science, Vol. 5545, Issue 1, p. 633-642.

Zorn, P. & Lea, M. (1989). Mortgage borrower repayment behaviour: a microeconomic analysis with Canadian adjustable rate mortgage data. *Real Estate Economics*, Vol. 17, Issue 1, p. 118-136.

Özdemir, Ö. & Boran, L. (2004). An Empirical Investigation on Consumer Credit Default Risk. Turkish Economic Association Working Paper 2004 / 20.

Other references

Amarnath, K. N. Statistical Methods in Consumer Credit Scoring. (http://www.hearne.com.au/attachments/Statistical%20methods%20in%20credit%20scoring.pdf)

Federation of Finnish Financial Services (2010). Kulutusluottoselvitys Tammikuu 2010. (http://www.fkl.fi/www/page/fk_www_3994)

Finnish Law of Consumer Protection (2010). (http://www.finlex.fi/fi/laki/ajantasa/1978/19780038)

Finnish Consumer Agency (2010). (www.kuluttajavirasto.fi)

Kulutusluotto (2010). (www.kulutusluotto.org)

Lainatieto.fi (2009). (www.lainatieto.fi/kulutusluotot)

Suomen Asiakastieto (2009). Taantuma lisännyt yritysten ja yksityishenkilöiden maksuhäiriömerkintöjä huomattavasti. (http://www.asiakastieto.fi/asiakastieto/ajankohtaista.jsp?l1=1&T=nu&A=321)

Statistics Finland (2010). Luottokanta 2009, 4. Vuosineljännes. (http://www.stat.fi/til/lkan/2009/04/)

Finnish Population Register Centre (2010). (www.vaestorekisterikeskus.fi)

Appendixes

Table 11: Coefficients for Model 1

Model 1					
Variable	**Value**	**Coefficient**	**Std. Error**	**p-value**	**Wald**
CREDIT	no				
	yes	-,238	,052	,000	20,488
EDUCATION	primary school				
	technical school	,002	,065	,980	,001
	high school	-,024	,084	,777	,080
	college	-,314	,087	,000	13,059
GENDER	man			,000	19,697
	woman	-,232	,053	,000	19,161
HOUSING	own			,000	32,563
	rented	,294	,054	,000	29,864
	employment relationship	,372	,187	,047	3,958
	partial ownership	,001	,126	,993	,000
INCOME	<1000			,005	14,727
	1000-1500	,091	,152	,549	,360
	1501-2000	,113	,155	,465	,534
	2001-2500	,108	,161	,503	,449
	>2501	-,143	,165	,387	,750
LEVEMPL	agriculture entrepreneuer				
	entrepreneuer	-,089	,124	,474	,512
	upper employee	-,031	,120	,793	,069
	lower employee	-,194	,117	,096	2,779
	employee	-,172	,099	,084	2,993
	student	,080	,197	,685	,164
	pensioner	-,018	,141	,897	,017
	maternity / parental leave	-,641	,263	,015	5,925
	unemployee	-,593	,296	,045	4,021
	laid off	-21,170	22255,647	,999	,000
	other	,204	,211	,333	,938
LOANSIZE	1000-1500			,000	41,846
	2000-2500	,128	,074	,083	3,009
	3000-3500	,360	,083	,000	18,941
	4000	,462	,079	,000	34,468
NATIONALITY	finnish				
	other	,390	,153	,011	6,518
PAYBACK	12			,000	41,333
	18	,173	,129	,181	1,791
	24	,010	,106	,928	,008
	30	,345	,129	,007	7,163
	36	,210	,109	,053	3,748
	42	,353	,139	,011	6,422
	48	,440	,096	,000	21,059
POSTAL	00000-09999			,040	17,586
	10000-19999	-,083	,096	,386	,751
	20000-29999	,068	,082	,403	,699
	30000-39999	,050	,082	,542	,372
	40000-49999	-,207	,101	,040	4,214
	50000-59999	-,056	,124	,650	,207
	60000-69999	-,203	,106	,055	3,672
	70000-79999	-,158	,121	,191	1,708
	80000-89999	,205	,117	,080	3,073

	90000-99999	-,015	,091	,872	,026
PREVLOAN	no				
	yes	,780	,049	,000	252,522
SCORE	<400			,000	120,958
	400-450	-,094	,073	,200	1,641
	451-500	-,340	,077	,000	19,533
	501-550	-,648	,085	,000	57,982
	>551	-,694	,073	,000	91,216
constant		-,755	,205	,000	13,555
AIC = 10 270					

Table 12: Coefficients for Model 2

Model 2					
Variable	**Value**	**Coefficient**	**Std. Error**	**p-value**	**Wald**
CREDIT	no				
	yes	-,227	,052	,000	19,281
EDUCATION	primary school			,000	18,402
	technical school	-,013	,064	,838	,042
	high school	-,031	,082	,702	,146
	college	-,295	,082	,000	13,100
GENDER	man				
	woman	-,238	,052	,000	20,791
HOUSING	own			,000	35,000
	rented	,297	,052	,000	32,239
	employment relationship	,382	,185	,039	4,250
	partial ownership	,011	,124	,931	,008
INCOME	<1000			,001	17,587
	1000-1500	,091	,142	,519	,416
	1501-2000	,098	,139	,481	,497
	2001-2500	,090	,144	,533	,389
	>2501	-,169	,147	,251	1,319
LOANSIZE	1000-1500			,000	119,751
	2000-2500	,206	,069	,003	8,796
	3000-3500	,509	,075	,000	46,205
	4000	,676	,067	,000	101,811
NATIONALITY	finnish				
	other	,379	,151	,012	6,336
PREVLOAN	no				
	yes	,773	,049	,000	252,510
SCORE	<400			,000	126,313
	400-450	-,071	,072	,327	,962
	451-500	-,324	,076	,000	18,262
	501-550	-,631	,083	,000	57,429
	>551	-,671	,069	,000	93,222
constant		-,730	,153	,000	22,728
AIC = 6 370					

Table 13: Coefficients for Model 3

Model 3					
Variable	**Value**	**Coefficient**	**Std. Error**	**p-value**	**Wald**
AGE	<20				
	20-25			,000	42,783
	26-45	-,083	,076	,276	1,185

	46-60	-,440	,084	,000	27,187
	61-70	-,406	,160	,011	6,441
	>70	,811	1,432	,571	,321
CREDIT	no				
	yes	-,228	,052	,000	19,283
EDUCATION	primary school			,000	28,794
	technical school	-,052	,065	,419	,654
	high school	-,055	,084	,515	,424
	college	-,414	,088	,000	22,212
GENDER	man				
	woman	-,368	,051	,000	51,724
HOUSING	own			,000	28,298
	rented	,299	,060	,000	24,646
	employment relationship	,420	,183	,022	5,250
	partial ownership	-,012	,125	,921	,010
HOUSINGTYPE	house			,028	9,132
	rowhouse	,122	,070	,083	3,009
	apartment	,184	,067	,006	7,668
	other	,297	,168	,077	3,134
INCOME	<1000			,000	31,792
	1000-1500	,054	,150	,719	,130
	1501-2000	,046	,153	,762	,092
	2001-2500	-,018	,159	,908	,013
	>2501	-,337	,163	,038	4,294
LEVEMPL	agriculture entrepreneuer			,022	20,830
	entrepreneuer	-,041	,122	,739	,111
	upper employee	-,128	,118	,279	1,171
	lower employee	-,211	,115	,066	3,376
	employee	-,046	,097	,634	,227
	student	-,015	,196	,940	,006
	pensioner	,165	,148	,266	1,238
	maternity / parental leave	-,583	,262	,026	4,960
	unemployee	-,392	,292	,179	1,802
	laid off	-20,844	22756,416	,999	,000
	other	,359	,208	,085	2,975
LOANSIZE	1000-1500			,000	38,682
	2000-2500	,112	,073	,122	2,397
	3000-3500	,342	,081	,000	17,644
	4000	,430	,077	,000	31,022
MOVING	under one year			,000	48,067
	2	-,028	,087	,750	,102
	3	-,098	,090	,280	1,169
	4	-,116	,099	,244	1,357
	5	-,079	,100	,429	,626
	6	-,329	,111	,003	8,821
	from 7 to 10	-,374	,094	,000	15,853
	from 11 to 15	-,409	,106	,000	14,816
	>15	-,495	,100	,000	24,602
PAYBACK	12			,000	41,013
	18	,190	,127	,135	2,238
	24	,002	,104	,983	,000
	30	,348	,127	,006	7,531
	36	,218	,107	,042	4,148
	42	,298	,137	,030	4,737
	48	,432	,095	,000	20,830
SIZEHOUSEHOLD	1			,000	22,077
	2	-,204	,061	,001	11,092

	3	,045	,079	,565	,331
	4	-,236	,079	,003	8,903
	5+	-,171	,104	,102	2,670
constant		-,160	,207	,438	,603

AIC = 9 633

Table 14: Information Values for variables

Variable	Value	Non-default	Default	Total	%non-default	%default	odds	information value
AGE	<20	0	0	0				
	20-25	882	702	1584	0,08478	0,16750	1,97584	0,05634
	26-45	4665	2087	6752	0,44839	0,49797	1,11059	0,00520
	46-60	4234	1233	5467	0,40696	0,29420	0,72293	0,03658
	61-70	605	166	771	0,05815	0,03961	0,68114	0,00712
	>70	18	3	21	0,00173	0,00072	0,41374	0,00090
	Total	10404	4191	14595	1,00000	1,00000	Total	0,10614

Variable	Value	Non-default	Default	Total	%non-default	%default	odds	information value
CITY	in small city	7696	2930	10626	0,73972	0,69912	0,94512	0,00229
	in one of the 5 largest cities	2708	1261	3969	0,26028	0,30088	1,15598	0,00588
	Total	10404	4191	14595	1,00000	1,00000	Total	0,11431

Variable	Value	Non-default	Default	Total	%non-default	%default	odds	information value
COTTAGE	no	8498	3556	12054	0,81680	0,84848	1,03879	0,00121
	yes	1906	635	2541	0,18320	0,15152	0,82705	0,00602
	Total	10404	4191	14595	1,00000	1,00000	Total	0,12153

Variable	Value	Non-default	Default	Total	%non-default	%default	odds	information value
CREDIT	no	3297	1673	4970	0,31690	0,39919	1,25968	0,01900
	yes	7107	2518	9625	0,68310	0,60081	0,87953	0,01056
	Total	10404	4191	14595	1,00000	1,00000	Total	0,15110

Variable	Value	Non-default	Default	Total	%non-default	%default	odds	information value
EDUCATION	primary school	2325	883	3208	0,22347	0,21069	0,94280	0,00075
	technical school	4693	2057	6750	0,45108	0,49081	1,08809	0,00335

	high school	1445	622	2067	0,13889	0,14841	1,06858	0,00063
	college	1941	629	2570	0,18656	0,15008	0,80447	0,00794
	Total	10404	4191	14595	1,00000	1,00000	Total	0,01268

Variable	**Value**	**Non-default**	**Default**	**Total**	**%non-default**	**%default**	**odds**	**information value**
EMPLOYMENT	permanent	2356	890	3246	0,22645	0,21236	0,93777	0,00091
	fixed-term	8048	3301	11349	0,77355	0,78764	1,01822	0,00025
	part-time				0,00000	0,00000	0,00000	0,00000
	unemployee				0,00000	0,00000	0,00000	0,00000
	Total	10404	4191	14595	1,00000	1,00000	Total	0,00116

Variable	**Value**	**Non-default**	**Default**	**Total**	**%non-default**	**%default**	**odds**	**information value**
FREEEMAIL	official email address	1954	724	2678	0,18781	0,17275	0,91981	0,00126
	free email address	8450	3467	11917	0,81219	0,82725	1,01854	0,00028
	Total	10404	4191	14595	1,00000	1,00000	Total	0,00154

Variable	**Value**	**Non-default**	**Default**	**Total**	**%non-default**	**%default**	**odds**	**information value**
GENDER	man	5471	2755	8226	0,52586	0,65736	1,25008	0,02935
	woman	4933	1436	6369	0,47414	0,34264	0,72265	0,04272
	Total	10404	4191	14595	1,00000	1,00000	Total	0,07207

Variable	**Value**	**Non-default**	**Default**	**Total**	**%non-default**	**%default**	**odds**	**information value**
HOUSING	own	6769	2143	8912	0,65062	0,51133	0,78592	0,03355
	rented	3047	1806	4853	0,29287	0,43092	1,47139	0,05332
	employment relationship	156	81	237	0,01499	0,01933	1,28897	0,00110
	partial ownership	432	161	593	0,04152	0,03842	0,92518	0,00024
	Total	10404	4191	14595	1,00000	1,00000	Total	0,08821

Variable	Value	Non-default	Default	Total	%non-default	%default	odds	information value
HOUSINGTYPE	house	5259	1609	6868	0,50548	0,38392	0,75951	0,03344
	rowhouse	1845	848	2693	0,17734	0,20234	1,14099	0,00330
	apartment	3147	1652	4799	0,30248	0,39418	1,30315	0,02428
	other	153	82	235	0,01471	0,01957	1,33047	0,00139
	Total	10404	4191	14595	1,00000	1,00000	Total	0,06240

Variable	Value	Non-default	Default	Total	%non-default	%default	odds	information value
INCOME	<1000	5612	1022	6634	0,53941	0,24386	0,45208	0,23464
	1000-1500	965	705	1670	0,09275	0,16822	1,81361	0,04493
	1501-2000	1622	1211	2833	0,15590	0,28895	1,85343	0,08210
	2001-2500	1031	699	1730	0,09910	0,16679	1,68307	0,03524
	>2501	1139	516	1655	0,10948	0,12312	1,12463	0,00160
	Total	10404	4191	14595	1,00000	1,00000	Total	0,39850

Variable	Value	Non-default	Default	Total	%non-default	%default	odds	information value
LEVEMPL	agriculture entrepreneuer	905	329	1234	0,08699	0,07850	0,90246	0,00087
	entrepreneuer	870	342	1212	0,08362	0,08160	0,97586	0,00005
	upper employee	1613	549	2162	0,15504	0,13099	0,84493	0,00405
	lower employee	1340	459	1799	0,12880	0,10952	0,85034	0,00313
	employee	4102	1965	6067	0,39427	0,46886	1,18919	0,01292
	student	243	116	359	0,02336	0,02768	1,18504	0,00073
	pensioner	897	279	1176	0,08622	0,06657	0,77214	0,00508
	maternity / parental leave	106	33	139	0,01019	0,00787	0,77284	0,00060
	unemployee	148	42	190	0,01423	0,01002	0,70448	0,00147
	laid off	15	1	16	0,00144	0,00024	0,16550	0,00216
	other	165	76	241	0,01586	0,01813	1,14344	0,00030
	Total	10404	4191	14595	1,00000	1,00000	Total	0,03137

Variable	Value	Non-default	Default	Total	%non-default	%default	odds	information value
LOANSIZE	1000-1500	2809	891	3700	0,26999	0,21260	0,78742	0,01372
	2000-2500	2612	962	3574	0,25106	0,22954	0,91429	0,00193
	3000-3500	1849	821	2670	0,17772	0,19590	1,10227	0,00177
	4000	3134	1517	4651	0,30123	0,36197	1,20163	0,01116
	Total	10404	4191	14595	1,00000	1,00000	Total	0,02857

Variable	Value	Non-default	Default	Total	%non-default	%default	odds	information value
MARITAL	married	5011	1543	6554	0,48164	0,36817	0,76441	0,03048
	single	2106	1181	3287	0,20242	0,28179	1,39211	0,02626
	cohabitation	2137	988	3125	0,20540	0,23574	1,14772	0,00418
	devorced	944	392	1336	0,09073	0,09353	1,03085	0,00009
	widow	180	74	254	0,01730	0,01766	1,02057	0,00001
	other	26	13	39	0,00250	0,00310	1,24123	0,00013
	Total	10404	4191	14595	1,00000	1,00000	Total	0,06115

Variable	Value	Non-default	Default	Total	%non-default	%default	odds	information value
MILITARY	no	5950	2186	8136	0,57190	0,52159	0,91204	0,00463
	yes	4454	2005	6459	0,42810	0,47841	1,11750	0,00559
	Total	10404	4191	14595	1,00000	1,00000	Total	0,01022

Variable	Value	Non-default	Default	Total	%non-default	%default	odds	information value
MONTHLY	0-60	2434	964	3398	0,23395	0,23002	0,98319	0,00007
	61-80	1585	676	2261	0,15235	0,16130	1,05877	0,00051
	81-95	4237	1817	6054	0,40725	0,43355	1,06458	0,00165
	>96	2106	731	2837	0,20242	0,17442	0,86167	0,00417
	Total	10404	4191	14595	1,00000	1,00000	Total	0,00639

Variable	Value	Non-default	Default	Total	%non-default	%default	odds	information value
MOVING	under one year	989	609	1598	0,09506	0,14531	1,52863	0,02133
	2	1036	634	1670	0,09958	0,15128	1,51919	0,02162
	3	1065	546	1611	0,10236	0,13028	1,27270	0,00673
	4	900	394	1294	0,08651	0,09401	1,08677	0,00062
	5	898	417	1315	0,08631	0,09950	1,15277	0,00187
	6	880	309	1189	0,08458	0,07373	0,87168	0,00149
	from 7 to 10	1595	498	2093	0,15331	0,11883	0,77509	0,00878
	from 11 to 15	1070	324	1394	0,10285	0,07731	0,75170	0,00729
	>15	1971	460	2431	0,18945	0,10976	0,57937	0,04349
	Total	10404	4191	14595	1,00000	1,00000	Total	0,11323

Variable	Value	Non-default	Default	Total	%non-default	%default	odds	information value
NATIONALITY	finnish	10231	4054	14285	0,98337	0,96731	0,98367	0,00026
	other	173	137	310	0,01663	0,03269	1,96588	0,01086
	Total	10404	4191	14595	1,00000	1,00000	Total	0,01112

Variable	Value	Non-default	Default	Total	%non-default	%default	odds	information value
NATIVE	finnish	10000	3997	13997	0,96117	0,95371	0,99224	0,00006
	swedish	277	87	364	0,02662	0,02076	0,77969	0,00146
	other	127	107	234	0,01221	0,02553	2,09152	0,00983
	Total	10404	4191	14595	1,00000	1,00000	Total	0,01135

Variable	Value	Non-default	Default	Total	%non-default	%default	odds	information value
NRADULTS	1	6810	2183	8993	0,65456	0,52088	0,79577	0,03054
	2	3438	1921	5359	0,33045	0,45836	1,38709	0,04185
	3+	156	87	243	0,01499	0,02076	1,38445	0,00188

	Total	10404	4191	14595	1,00000	1,00000	Total	0,07427

Variable	Value	Non-default	Default	Total	%non-default	%default	odds	information value
NRCHILDREN	0	8296	2945	11241	0,79739	0,70270	0,88125	0,01197
	1	763	495	1258	0,07334	0,11811	1,61051	0,02134
	2	908	531	1439	0,08727	0,12670	1,45175	0,01470
	3+	437	220	657	0,04200	0,05249	1,24975	0,00234
	Total	10404	4191	14595	1,00000	1,00000	Total	0,05034

Variable	Value	Non-default	Default	Total	%non-default	%default	odds	information value
PAYBACK	12	1356	376	1732	0,13033	0,08972	0,68835	0,01517
	18	567	204	771	0,05450	0,04868	0,89316	0,00066
	24	1646	516	2162	0,15821	0,12312	0,77822	0,00880
	30	582	244	826	0,05594	0,05822	1,04076	0,00009
	36	1477	555	2032	0,14196	0,13243	0,93281	0,00066
	42	453	221	674	0,04354	0,05273	1,21109	0,00176
	48	4281	2072	6353	0,41148	0,49439	1,20151	0,01522
	Total	10404	4191	14595	1,00000	1,00000	Total	0,04236

Variable	Value	Non-default	Default	Total	%non-default	%default	odds	information value
PHONE	hasn't called	9643	3909	13552	0,92686	0,93271	1,00632	0,00004
	has called	761	281	1042	0,07314	0,06705	0,91665	0,00053
	Total	10404	4191	14595	1,00000	1,00000	Total	0,00057

Variable	Value	Non-default	Default	Total	%non-default	%default	odds	information value
POSTAL	00000-09999	3141	1321	4462	0,30190	0,31520	1,04404	0,00057
	10000-19999	812	311	1123	0,07805	0,07421	0,95080	0,00019
	20000-29999	1227	535	1762	0,11794	0,12765	1,08241	0,00077

	30000-39999	1162	534	1696	0,11169	0,12742	1,14082	0,00207
	40000-49999	823	269	1092	0,07910	0,06419	0,81140	0,00312
	50000-59999	414	180	594	0,03979	0,04295	1,07933	0,00024
	60000-69999	762	260	1022	0,07324	0,06204	0,84703	0,00186
	70000-79999	567	184	751	0,05450	0,04390	0,80560	0,00229
	80000-89999	530	215	745	0,05094	0,05130	1,00704	0,00000
	90000-99999	966	382	1348	0,09285	0,09115	0,98168	0,00003
	Total	10404	4191	14595	1,00000	1,00000	Total	0,01115

Variable	**Value**	**Non-default**	**Default**	**Total**	**%non-default**	**%default**	**odds**	**information value**
PREVLOAN	no	7916	2196	10112	0,76086	0,52398	0,68867	0,08836
	yes	2488	1995	4483	0,23914	0,47602	1,99056	0,16307
	Total	10404	4191	14595	1,00000	1,00000	Total	0,25143

Variable	**Value**	**Non-default**	**Default**	**Total**	**%non-default**	**%default**	**odds**	**information value**
REPAYMENTB	no delays	8049	4191	12240	0,77364	1,00000	1,29258	0,05809
	delays	2355	0	2355	0,22636	0,00000	0,00000	0,00000
	Total	10404	4191	14595	1,00000	1,00000	Total	0,05809

Variable	**Value**	**Non-default**	**Default**	**Total**	**%non-default**	**%default**	**odds**	**information value**
SCORE	<400	2683	1776	4459	0,25788	0,42377	1,64326	0,08239
	400-450	2113	843	2956	0,20309	0,20115	0,99040	0,00002
	451-500	1818	575	2393	0,17474	0,13720	0,78516	0,00908
	501-550	1575	407	1982	0,15138	0,09711	0,64150	0,02409
	>551	2215	590	2805	0,21290	0,14078	0,66124	0,02983
	Total	10404	4191	14595	1,00000	1,00000	Total	0,14542

Variable	**Value**	**Non-default**	**Default**	**Total**	**%non-default**	**%default**	**odds**	**information value**

SIZEHOUSEHOLD	1	6497	1970	8467	0,62447	0,47005	0,75272	0,04386
	2	1890	1032	2922	0,18166	0,24624	1,35550	0,01964
	3	705	489	1194	0,06776	0,11668	1,72188	0,02658
	4	887	482	1369	0,08526	0,11501	1,34898	0,00891
	5+	425	218	643	0,04085	0,05202	1,27336	0,00270
	Total	10404	4191	14595	1,00000	1,00000	Total	0,10169

Variable	**Value**	**Non-default**	**Default**	**Total**	**%non-default**	**%default**	**odds**	**information value**
TIME	0-5	141	87	228	0,01355	0,02076	1,53173	0,00307
	6-8	842	308	1150	0,08093	0,07349	0,90807	0,00072
	9-16	6413	2599	9012	0,61640	0,62014	1,00607	0,00002
	17-20	2343	905	3248	0,22520	0,21594	0,95887	0,00039
	21-23	665	292	957	0,06392	0,06967	1,09004	0,00050
	Total	10404	4191	14595	1,00000	1,00000	Total	0,00470

Variable	**Value**	**Non-default**	**Default**	**Total**	**%non-default**	**%default**	**odds**	**information value**
TIMEFIN	always	10265	4084	14349	0,98664	0,97447	0,98766	0,00015
	1-3 months	13	31	44	0,00125	0,00740	5,91972	0,01093
	4-6 months	32	23	55	0,00308	0,00549	1,78427	0,00140
	7-12 months	48	24	72	0,00461	0,00573	1,24123	0,00024
	13+ months	46	29	75	0,00442	0,00692	1,56503	0,00112
	Total	10404	4191	14595	1,00000	1,00000	Total	0,01384