Norwegian School of Economics

Bergen, Fall 2018

Putting the brakes on consumer loans

How lenders can reduce default on consumer loans in Norway

Bjørnar Bergsvik and Otto Kierulf

Supervisor: Steffen Juranek and Øivind Anti Nilsen

Master Thesis, Master of Science in Economics and Business Administration, Financial Economics

NORWEGIAN SCHOOL OF ECONOMICS

Preface

This thesis concludes many years of education at the Norwegian School of Economics. Although the process has been both long and challenging, we were left with feelings of purpose and joy after we finished. We hope that our work can provide value for others out there.

We would like to thank a number of persons for their contributions to the final product and enabling us to succeed with this thesis.

First, we would like to thank our supervisors Steffen Juranek and Øivind Anti Nilsen, for important feedback and guidance that undoubtedly improved the quality of the thesis. We especially appreciate your patience and understanding during the entire process.

Next, we would like to thank the anonymous bank for providing us with the dataset for consumer loans. This thesis would not have been possible without this basis for analysis. A special thank you to our main contact at the bank for quick and thorough explanations of the

A thank you to Peer Timo Andersen-Ulven from Bank Norwegian, for providing useful information on both the practices of Bank Norwegian and the market, and Vegard Daltveit, who shared insightful information from his own research on consumer loans. We also want to recognise the help from Kjetil Gromholt and Øystein Fjalestad at Eika kredittbank, Audun Bø and Lasse Hammer. Your tip and comments guided us through the early stages of the process.

Finally, we would like to thank our friends and families for all your love and support.

Bjørnar Bergsvik

Otto Kierulf

Bergen, 20.12.2018

Summary

In this thesis, we provide a method for lenders to reduce defaults on consumer loans in the Norwegian market. Using a dataset of 10 836 loans provided by a Norwegian consumer bank, we compare the characteristics of the loan takers and the loans and build two models for predicting probability of default.

By studying the characteristics of the loan takers and the loans, we want to see if appropriate to use for predicting defaults. We do this by comparing the defaulting loans with the non-defaulting, to see if any differences exist. Looking at both demographic and financial characteristics, we also find that certain groups of the population are more likely to default. Evidence suggest that this is particularly true for people under the age of 40 and people with an annual income less than 450 000 NOK. In contrast, our analysis shows that being married, having a master's degree or being a private owner reduce the risk of default. We also see that a higher interest rate or more principal free months increase the risk of the loan. We conclude that the characteristics can be used for predictive purposes.

The purpose of the predictive models is to assist lenders in reducing defaults on future consumer loans. Based on stepwise backward selection, Mallow's Cp and machine learning with Monte Carlo cross-validation, two logistic regression models are constructed. These models return predicted probability of default for a loan, using characteristics of the loan taker and the loan. The lender can utilize the models to ensure that no approved loans exceed the lender's risk preference, by adjusting the attributes of the loan according to a desired threshold for probability of default. For instance, a probability threshold of 15 percent correctly predicted more than half of the defaulted loans.

In order to understand why the lender should try to reduce defaults, we also investigate why defaults occur in the market and what the financial consequences are. We conclude that lenders should reduce defaults to reduce costs, and that this can be done by adjusting the loans more properly to the loan takers. A general description of the market for consumer loans in Norway is also presented in order to give the reader a better understanding of the subject.

Contents

1. Introduction

When a lender provides a loan to a borrower, the profitability of the lender is dependent on the borrower's ability to repay the loan and avoid default. If large amounts of the loans default, the lender will likely go bankrupt and the economy suffer. This is partly what caused the financial crisis of 2007-08, where trillions of dollars were lost as high-risk borrowers defaulted on their mortgages when the housing prices plummeted (Goodhart, 2008, p.337-338). It is therefore crucial for lenders to identify risky borrowers and keep defaults at a sustainable level.

In Norway, the default rate on consumer loans[1] has almost doubled over the last four years; from to 4.1 percent in 2014 to 7.3 percent in the third quarter of 2018 (Finanstilsynet, 2014 & 2018a). In their semi-annual report in June 2018, Finanstilsynet stated that defaults on consumer loans can lead to large expenses for the consumers, and reduce the profitability and reputation of the banks (Finanstilsynet, 2018b, p.4). It is important for the Norwegian lenders to better assess risk of default, and stop the growing default rate. Asymmetric information between the borrower and the lender complicates this, as the borrower knows more about the probability of repayment than the lender, and several screening devices is needed for the lender to assess the risk of loans (Stiglitz and Weiss, 1981).

In this thesis, we propose one such screening device for estimating the probability of default on consumer loans. Focusing on the Norwegian market, we compare the characteristics of the defaulting loans with the non-defaulting loans to see if sufficiently significant to be used for predictive purposes. The data is provided by a Norwegian consumer bank that wish to be anonymous[2], and includes characteristics of both the borrowers and the loans. We then move on to construct two models for predicting probability of default for consumer loans, one with and one without inputs from the lender. A lender can use any of these models to estimate the risk of the loan. A probability threshold can then be implemented to decide which loans should be approved and not, in order to reject loans too risky for the lender.

[1] A definition of consumer loan, in addition to other useful definitions for this thesis, is presened in Appendix A

[2] The bank that provided the dataset will simply be referred to as «the bank» in the remainder of the thesis.

1.1 The purpose of the thesis

The thesis tries is to answer the following question:

1. *How can lenders reduce default on consumer loans in Norway?*

We hope that lenders will use this thesis to reduce the number of default on consumer loans in Norway. We think that the default rates on consumer loans in Norway are too high, and that the lenders have a responsibility to reject or adjust more loan applications.

1.2 Motivation

The background for this thesis started with a curiosity to learn more about the market for consumer loans in Norway, and why it has changed so quickly. Ten years ago, neither of us had heard of consumer loans. Now, it is difficult to last 24 hours without being exposed to an advertisement, or reading about it in the news.

While many of our colleagues and friends swear that consumer loans only serves the lenders, we recognise that it provides value to consumers that lack collateral and can afford the increased costs later on. However, we find the rising default rates worrisome. As we explain in the thesis, defaults incurs costs on both the borrower, the lender, and potentially the society, and should be minimised at all times. We wanted to present a method specifically to be used for lenders of consumer loans in Norway, as we did not find this anywhere else.

We presume that all of the lenders are already using similar models, however there is always something new that can be added to either the method or the way of thinking. This thesis was motivated by a hope of adding value to new or existing models on consumer loans, and help reduce the default rates on consumer loans in Norway.

1.3 Reliability

We consider the primary data collected from the bank to be very reliable, as the observations represent actual consumer loans approved by the bank. We were able to decide on both the time horizon and the variables of the dataset, and after thorough cleaning of the data as discussed in Chapter 4, only the observations that were complete and considered representative of the general population were used.

SIFO is an institute for consumer research overseen by Oslomet, a Norwegian state university. Seeing that SIFO conducts research solely on consumer-related topics, and has been doing so since 1970 with funding provided by the Norwegian government, we consider their expertise and surveys to be reliable.

Statistisk Sentralbyrå (SSB) is subordinate of the Norwegian Ministry of Finance and acts as the Norwegian office for official government statistics. It is structured to be politically independent and unbiased, with a purpose of publishing Norwegian statistics on a regular basis available for everyone. We consider SSB's data to be very reliable.

The data used to analyse the market for consumer loans is collected and reported by Finanstilsynet, an independent government branch with mandate to supervise financial institutions in Norway. Since Finanstilsynet is subject to Norwegian Law and follow strict guidelines, we consider the data to be highly reliable.

To complement the analysis based on data from Finanstilsynet, additional data is collected from Norges Bank. Norges Bank is a separate legal entity owned by the state, responsible for managing monetary policy and ensuring financial stability. We consider this data to be reliable for the same reasons as stated in the paragraph above.

1.4 Structure

Chapter one introduces the thesis and how the paper is organized.

Chapter two presents a general overview of the Norwegian market for consumer loans.

Chapter three analyses reasons for defaults and its effect on lenders and borrowers.

Chapter four describes and evaluates the dataset received from the bank.

Chapter five studies the characteristics of the borrowers and the loans in the dataset, by comparing defaulted loans with non-defaulted loans.

Chapter six builds and presents two predictive models for probability of default on consumer loans.

Chapter seven concludes the thesis and discusses further research.

2. The Norwegian market for consumer loans

Total household debt in Norwegian is roughly 3 300 billion NOK (Christensen, 2018). Approximately three percent of this is composed of consumer loans (Finanstilsynet, 2018e, p.37). While the share is low, Finanstilsynet have expressed their concerns for the recent increased growth in this market. (Finanstilsynet, 2017a) As of now, the growth poses little threat to the overall financial stability in Norway, however it can lead an increasing number of households being vulnerable to default and economic distress. (Hagen, Turtveit, Vatne, 2017) This chapter gives an overview of the lenders and borrowers in the market and presents possible reasons for growth.

2.1 Definition of consumer loan

Consumer loan is an unsecured loan provided by either a bank or financial institution. It is different from a traditional mortgage, meaning that the lender does not require any collateral for the debt issued (DnB, 2018). Therefore, the interest rates on consumer loans are usually much higher than a secured loan. redit cards are also considered a consumer loan.

2.2 Lenders

2.2.1 Market shares

Through an extensive study, Finanstilsynet has been monitoring a selection of lenders providing consumer loans in Norway. Each year they release several reports updating their study of this market. Since these reports present the most accurate information available, the following analysis of the market is based mostly on these reports. Currently, the selection is composed of 30 financial institutions that offer consumer credit. Below is a list of 28[3] lenders, provided by senior advisor Jo Singstad at Finanstilsynet.

[3] The list was received 02.11.2018 and reflects the report released by Finanstilsynet in June 2018.(Finanstilsynet, 2018b

Avida Finans	Eika Kredittbank	Instabank	Santander Consumer Bank
Bank Norwegian	Eksress bank	Komplett Bank	Sbanken
BB Bank	Enter Card	Monobank	SEB Kort
Danske Bank	Eurocard	MyBank	Sparebank 1 Kredittkort
Diners Club Norge	Folkefinans	Nordea Bank	Sparebanken Vest
DNB	Gjensidige Bank	Nordea Finans Norge	Svea Finans
Easybank	Ikano Bank	Resurs Bank	Ya Bank

Table 2.2.1: Providers of consumer loans in Norway. (Source: Singstad, personal communication, 02 November 2018)

Even though all lenders in this selection offer consumer loans, only some have consumer loans as their main area of business. For some of the other lenders, including the three largest banks in Norway[4], consumer loans make up only a small fraction of the assets. The leading providers of consumer loans are Bank Norwegian AS, Santander Consumer Bank, yA Bank and Komplett Bank (Hagen et al., 2017). Below is an overview of their respective market shares in 2016 estimated in a report published by Norges Bank.

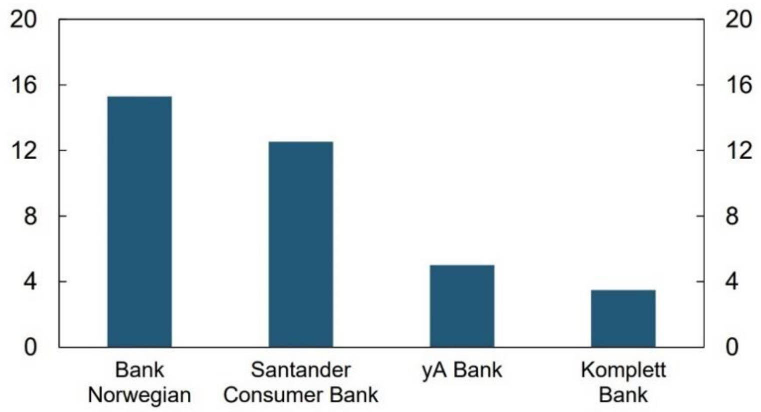


Figure 2.2.1: Market share of consumer banks (Source: Hagen et al., 2017)

2.2.2 Historical growth

The market for consumer loans have experienced an extraordinary growth. A report from Finanstilsynet has estimated the current total loan amount to 111 billion NOK, compared to 44.5 billion in 2009 (Finanstilsynet, 2010 & Finanstilsynet, 2018a). This corresponds to an

[4] According to Finans Norge, based on gross lending: DnB, Nordea Bank and Danske Bank (Finans Norge, 2017)

annual average growth rate of 8.1 percent. In contrast, the growth of overall household debt was flat throughout the same period, with an annual average growth rate of six percent. The difference is illustrated in Figure 2.2.2.

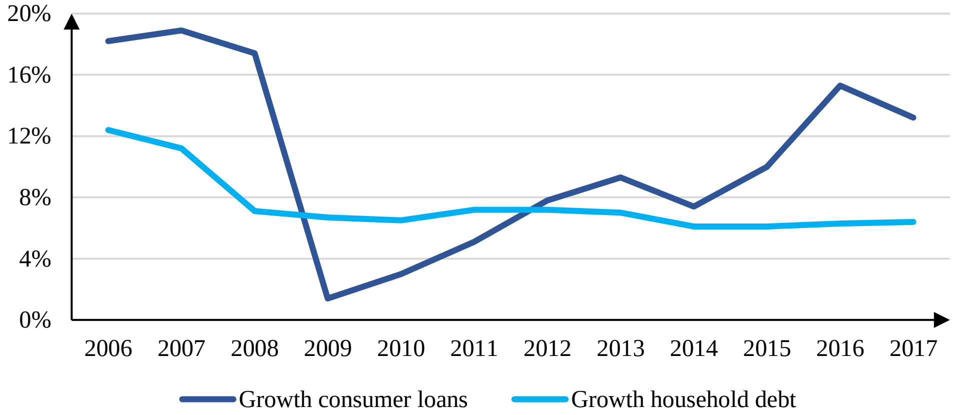
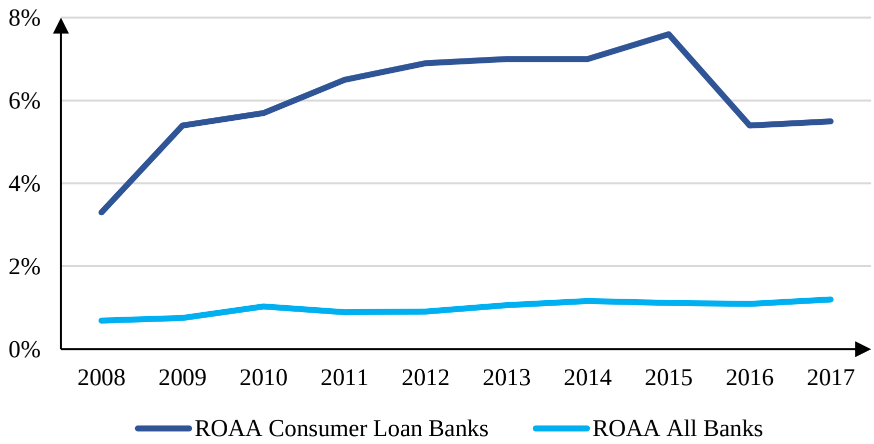


Figure 2.2.2: Growth rate consumer loans vs household debt (Source: Finanstilsynet, 2018c)

2.2.3 Reasons for growth

Supplying consumer loans has turned out to be a highly profitable business in Norway. Below is a visualisation of the historical development of return on average assets (ROAA)[5], differentiated by banks focusing on consumer loans and all banks on aggregate.



[5] ROAA is a common indicator to measure performance in the financial industry. It is derived by dividing net income by average of total assets. (Investopedia, 2018)

Figure 2.2.3: ROAA of banks focusing on consumer loans vs all banks (Source: Finanstilsynet, 2008-2018)

Figure 2.2.3 shows that banks focusing on consumer loans have achieved an average ROAA of approximately 6.3 percent. In comparison, the ROAA of all banks combined averaged to one percent for the same period. These differences illustrate how profitable this industry has been over the last years.

To explain the high profits of the banks focusing on consumer loans, the two largest profit-drivers are examined; interest rates and financing costs.

In order to compensate for the increased risk associated with unsecured debt, the interest rates of consumer loans are much higher than secured loans. Figure 2.2.4 shows that the historical net interest rate[6] for banks specialising on consumer loans has been steadily high the last nine years.

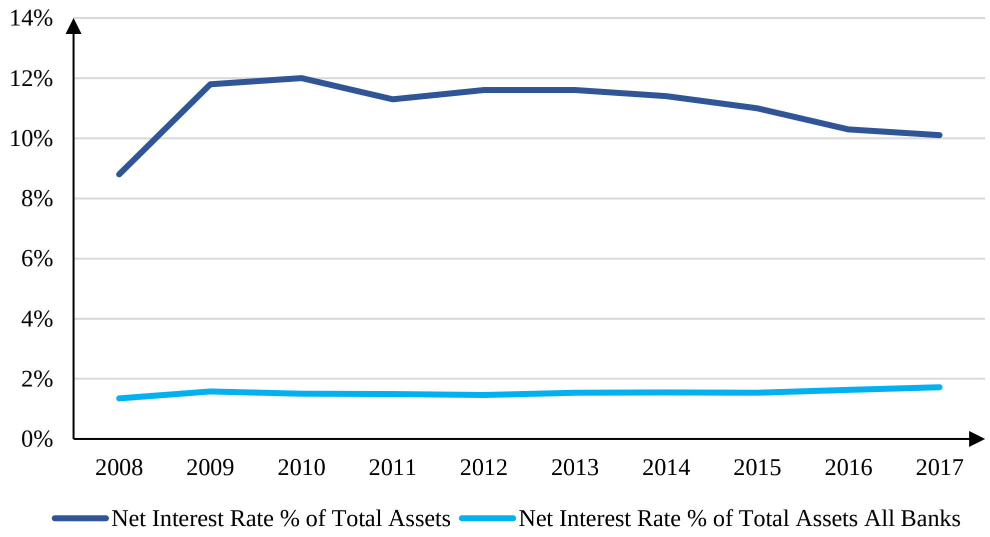


Figure 2.2.4: Net interest rate of total assets

Low costs of financing enable the banks focusing on consumer loans to have higher interest margins than other banks (Hagen et al., 2017). In most cases, the outstanding loans are fully financed by deposits from individuals, and no additional debt is needed. For example, Bank

[6] Net interest rate is the difference between interest earned on lending activities and interest paid on deposits and other interest-bearing liabilities.

Norwegian AS had 33.6 billion NOK in deposits from customers in 2017, while outstanding loans to clients amounted to 32.4 billion NOK (Bank Norwegian, 2017). By offering slightly higher deposit rates than the other banks, the banks focusing on consumer loans are an attractive alternative for depositors. Consequently, the financing costs are higher, but it enables the banks to attract enough depositors to avoid more expensive debt. Figure 2.2.6 shows a comparison of historical deposit rates in Norway.

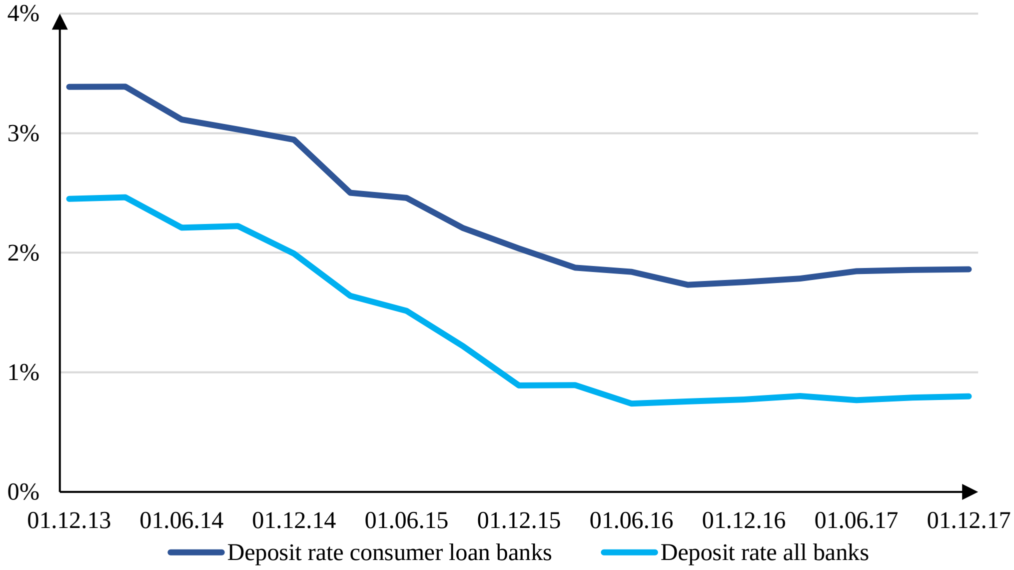


Figure 2.2.6: Deposit rates 2013-2017

Even though the consumer loans are risky, the deposits are risk-free for the savers because of the Norwegian Banks' Guarantee Fund arrangement. All banks headquartered in Norway are required to be members[7] (Finansforetaksloven, 2015, §19-1 – §19-4), and the fund guarantees a deposit of up to two million NOK for private savers in all its member banks (Bankenes Sikringsfond, 2018). This means that a person can have accounts in several banks, all guaranteed by the same fund. Banks pay a yearly fee of 0.01 percent of average guaranteed deposits and 0.005 percent of average risk-weighted assets[8] in their portfolio. As a result, banks with a large amount of unsecured debt pay a higher fee because of the increased fee from the risk-weighted assets.

[7] Foreign financial institutions that are members of the EEA and accepting deposits from Norwegian residents have a right to become a member.

[8] Risk-weighted assets is a measurement of the total exposure a bank faces in terms of credit risk, market risk and operational risk (Finanstilsynet, 2017c).

2.3 Borrowers

Lenders offering consumer loans are highly dependent on someone willing to buy their products, otherwise the market would not exist. Therefore, when explaining reasons for growth, it is also necessary to look at the demand side of the market; the borrowers.

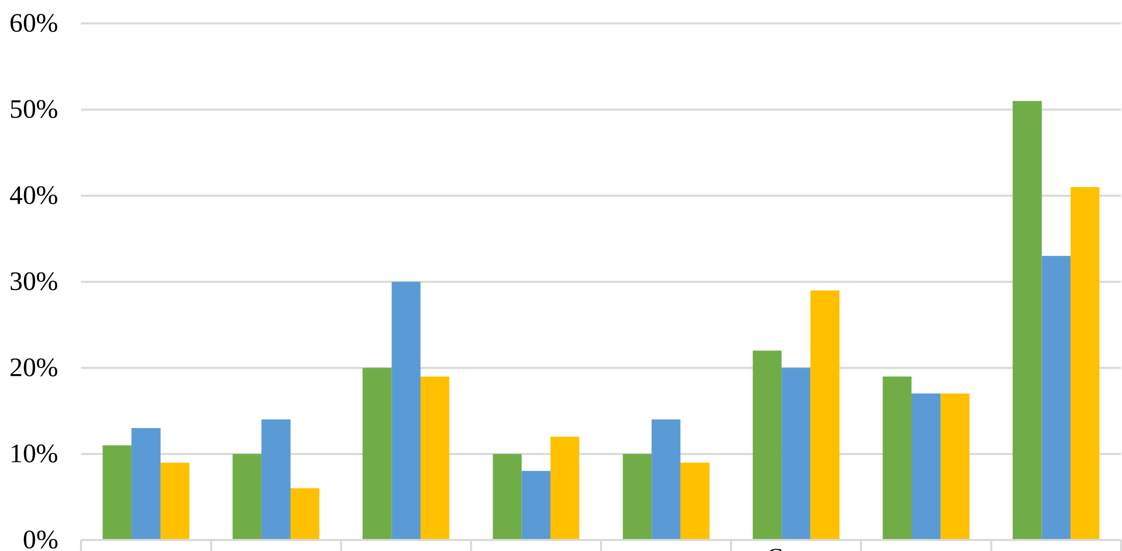
2.3.1 Decision to borrow

Elements from theories of consumer choice can be used to understand the decision to apply for a consumer loan. A thorough analysis of this theory and how it relates to borrowing can be read in a study by Lillebø and Hansen from 2016 (Hansen & Lillebø, 2016, p.14-18). The main point is that if a consumer's income does not cover their preferred consumption levels, they will need additional credit to fulfil their consumption needs. However, this will be at the expense of future consumption, and the utility the borrower gets from receiving the loan today must be higher than the losses incurred later. Concerning consumer loans, this means that if the necessity for a loan is large enough, the borrower will accept higher costs in the future.

Borrowing to finance consumption can also be related to what behavioural economists call the "present bias". People with a strong present bias have tendencies to place a higher emphasis on immediate payoff rather than later in time, even if the total value is lower (Bachmann, De Giorgi and Hens, 2018, p. 21). People with a strong present bias have a greater need of getting rewarded today, which can lead to economically irrational decisions. The study "Present-Biased Preferences and Credit Card Borrowing" from the American Economic Journal suggest that people with stronger present biases are more likely to borrow through credit cards (Meier & Sprenger, 2010, p.208).

2.3.2 Applications of consumer loans

OsloMet conducted a survey of loan takers, asking how they spent their consumer loans (SIFO, 2017, p.26). A complete overview of the results is presented in Figure 2.3.1.



	Other	Purchase of house/cabin	Refinancing debt	Purchase of car	Renovating house/cabin	Cover unexpected costs	Travelling	Consumer goods
2014	11%	10%	20%	10%	10%	22%	19%	51%
2015	13%	14%	30%	8%	14%	20%	17%	33%
2017	9%	6%	19%	12%	9%	29%	17%	41%

Figure 2.3.1: Survey of reasons for consumer loans (Source: SIFO, 2017, p.26)

The most common applications of the loans in 2017 were Consumer goods (41 percent), Cover unexpected costs (29 percent), Refinancing debt (19 percent) and Travelling (17 percent)[9]. Comparing the results across the three years, only small differences can be found. The only significant change was Refinancing debt, which varies from 19 to 30 percent.

2.3.3 Drivers for growth in borrowing

Hansen and Lillebø present an extensive empirical study of growth factors in the Norwegian consumer loan market in the period 2002-2016 (Hansen & Lillebø, 2016, p.74-98). According to their analysis, the key drivers for growth have been effective marketing strategies, increased availability and the Norwegian Banks' Guarantee Fund arrangement. Additionally, growth in GDP and unemployment correlates with growth in consumer loans based on their model. Two of the marketing strategies highlighted as possible drivers (Hansen & Lillebø, 2016, p. 93) will be presented in the next paragraphs.

[9] Respondents could give more than one answer, meaning that some respondents may have used their loan on several of the alternatives in the survey.

The first marketing strategy involves loyalty programs through a bank's credit card products. By offering bonuses for each payment made with the credit card, banks encourage consumers to a continuous use of credit cards instead of traditional debit cards. An article in Bergens Tidende from 2016 suggests that this strategy has been successful, claiming that Norwegians have become more inclined to finance daily consumption with credit (Mikalsen, 2016). A report published by Norges Bank's in 2017 also provides evidence of this, showing a quadrupling in the transaction volume with credit cards from 2005 to 2015 (Hagen et al., 2017). The total growth in consumer loans could be a result of the more favourable view on credit in Norway.

Promotions of rapid turnaround on loan applications is the second marketing strategy that may have led to an increased growth (Hansen & Lillebø, 2016, p. 93). Feeding on the present bias presented earlier, a rapid turnaround may tempt additional consumers to apply for consumer loans and increase growth[10].

2.4 Reactions from the government

2.4.1 Guidelines

In June 2017, Finanstilsynet issued a press release expressing concerns regarding the high growth of unsecured debt in Norwegian households. (Finanstilsynet, 2017a). In order to protect borrowers and make the banks more sustainable, specific guidelines for the consumer loan practices in Norway were introduced. The next paragraphs will give a short overview of the guidelines.[11]

Before granting a loan, the lender must be able to document that a credit assessment of the borrower has been conducted. The assessment should include the borrower's gross income, other debt obligations and relevant expenses that can affect the credit worthiness of the borrower. In addition to this, borrowers should be able to withstand a five percent increase in total interest costs. If a potential borrower has a total debt that is five times larger than their gross income, a loan should not be granted. Durability of consumer loans should not exceed

[10] This effect might disappear in the future, as the second marketing strategy became illegal in April 2017 through an updated regulation on marketing of credit (Regjeringen, 2017)

[11] The guidelines apply to both domestic and foreign financial institutions operating in Norway.

five years and ought to have a requirement of periodic installment payments (Finanstilsynet, 2017b).

These guidelines were effective from the fourth quarter of 2017. After a follow-up through questionnaires and inspections of the financial institutions, Finanstilsynet found that the implementation of the guidelines was, in many cases, not satisfactory (Finanstilsynet, 2018d). Based on the questionnaires, Finanstilsynet estimated that 35.9 percent of the granted loan applications did not meet the requirements set forth by the guidelines. In particular, they found many deviations from the requirements regarding durability and periodic installments. As a response to this, Finanstilsynet proposed in august 2018 that the guidelines should be made into official regulations under Norwegian law. The main argument for this change was to ensure that Finanstilsynet could better enforce the regulations and impose sanctions on those who fail to comply. An official decision on these regulations is expected to be made in 2019.

3. Aspects of default

This chapter will look at various aspects of default. The beginning of the chapter will present a formal definition of default, before moving on to possible reasons for why a person would default on a consumer loan and the effects of default. The goal of this chapter is to provide a basis for why defaults happen, and further explain why it is important for lenders to estimate probability of defaults.

3.1 Definition of default in IFRS 9

According to IFRS 9[12], a bank must consider a loan to be defaulted when a loan payment is 90 days past due (IASB, 2014, p 416). Since this standard became effective by law in Norway in 2014 (Kapitalforskriften, 2014, §5-11), all of the banks in Norway has adopted this practice when dealing with consumer loans. The remainder of the thesis will therefore adhere to the 90-days definition of default.

3.2 Possible reasons for default

In September 2015, SIFO performed a study where they asked people about reasons for why they were late on their loan payments. Figure 3.2.1 displays the result from this study[13].

[12] IFRS 9 is the standard in the IFRS that addresses the accounting practice for financial instruments, such as consumer loans.

[13] The people in the study were asked to name the main reasons for delinquency on any loans, not just consumer loans. It is also important to note that delinquency is not the same as default.

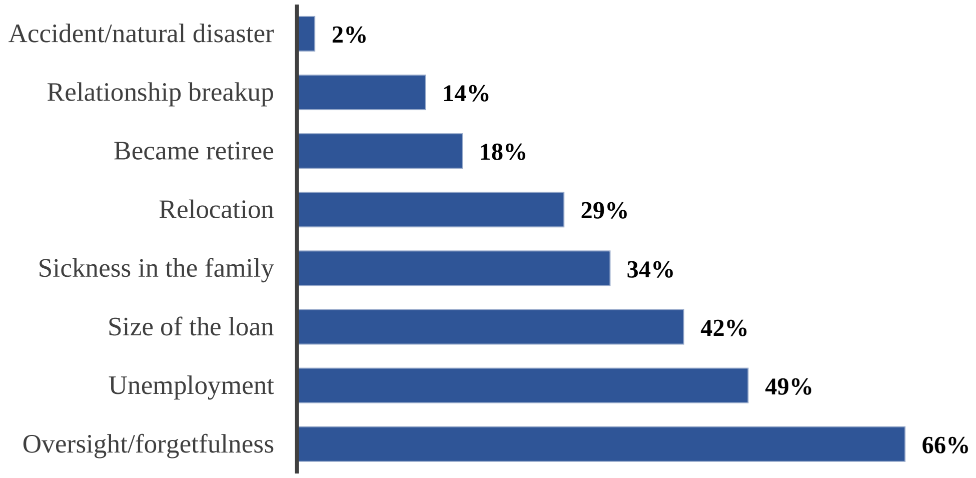


Figure 3.2.1: Reasons for delinquency on loans. (Source: Lavik & Borgeraas, 2015)

3.2.1 Oversight

The most reported reason for delinquency on loans was oversight or forgetfulness. This is something the contact at the bank confirmed, as they partly suspect many of their customers check their mailbox irregularly or forget to report change of address[14]. Although oversight is a reportedly common cause for loan delinquency, it is a less probable reason for default. The following two reasons explains why.

Oversight is likely a less dominant reason for default on consumer loans than on other loans, as many borrowers use a consumer loan to refinance their other loans (SIFO, 2017, p.26). Through refinancing, the borrower compiles several loans into one larger loan, making it easier to keep track of the debt. Furthermore, people are more inclined to maintain a larger loan as the cost of default will be more substantial.

Another reason for why oversight is a smaller issue for defaults on consumer loans is the follow-up process used by some of the Norwegian lenders. Table 3.2.1 demonstrates the process at Bank Norwegian, as described by their CRO Peer Timo Andersen-Ulven. The number of days is days after a payment is due.

[14] In Norway, a warning of debt collection must be sent by physical mail (Forbrukerrådet, 2018). If a person does not report change of address, it is likely that the warning will reach the person after the date of maturity.

1st month	**14 days**	**21 days**	**26 days**
	1st invoice	SMS-reminder	Block further credit
2nd month	**30 days**	**45 days**	**52 days**
	2nd invoice	Termination warning[15]	SMS-reminder
3rd month	**60 days**	**75 days**	**90 days**
	3rd invoice	SMS-reminder	Default

Table 3.2.1: Follow-up process of Bank Norwegian. (Source: Andersen-Ulven, personal communication, 16 November 2018)

As seen above, the follow-up process involves numerous invoices and SMS-reminders with the goal of reducing forgetfulness. Andersen-Ulven explained that SMS-reminders are a particularly effective tool, adding that 96 percent of Bank Norwegian's customers has paid after the first SMS-reminder.

3.2.2 Unemployment

In macroeconomics, unemployment is widely used as a factor in analysis due to how severely it reduces individuals' purchasing power. This also applies to an individual's capability of repaying a consumer loan. What follows is a comparison between unemployment rates in Norway and default rates on consumer loans during the last 15 years.

[15] An official warning to the debtor that the credit agreement will be terminated and the principal amount will be charged before maturity.

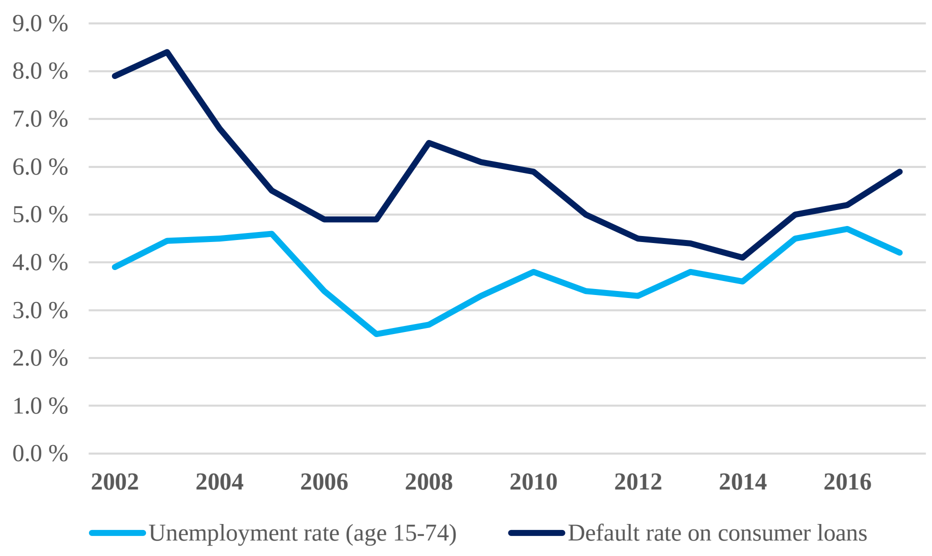


Figure 3.2.2: Unemployment rate vs default rate in Norway. Sources: (SSB, 2018) (Finanstilsynet, 2007), (Finanstilsynet, 2014), (Finanstilsynet, 2016-2017)

As illustrated by Figure 3.2.2, the relationship between unemployment and default on consumer is shifting. The rates have a high, negative correlation during the years 2003-2005 and 2008-2010, while it is strongly positive during 2005-2008 and 2013-2016. This suggests that unemployment has low causal effect on default rates[16]. Although the relationship is small, the data suggests that unemployment has a small upward effect on the default rate. Unfortunately, this is something that is nearly impossible for a lender to predict when evaluating a loan application.

3.2.3 Size of the loan

More than 40 percent of the participants in SIFO's study answered that the size of the loan, ergo the size of the loan payments, was one of the reasons for delinquency. If lenders provide loans that exceeds the borrowers' borrowing capacities, the number of defaults will increase.

In December 2016, E24 published an article concerning the ease of obtaining consumer loans (Vedeler, 2016). E24's journalist sent the same loan application for a consumer loan of 500 000 NOK to 15 different lenders. The table below is extracted from E24's article and portrays the diversity in the lenders' counteroffers to the loan application.

[16] A correlation coefficient of 0.26 between the unemployment rate and the default rate confirms this.

Bank	*Applied amount*	*Approved amount*	*Effective rate*
Bank Norwegian	500 000	500 000	18,53 %
Monobank	500 000	-	N/A
Instabank	500 000	200 000	22,36 %
yA Bank	500 000	200 900	16,44 %
Collector	500 000	40 000	25,71 %
DnB	500 000	110 000	19,74 %
Average	**500 000**	**175 150**	**20,56%**

Table 3.2.2: Offerings of loan from different lenders. (Source: Vedeler, 2016)

Table 3.2.2 shows how the offered loan amount ranges from zero to 500 000 NOK for the same loan application. For instance, the approved amount from Bank Norwegian was more than double the amount from any of the other lenders, signalling it is likely to be above the borrowing capacity of the borrower. Monobank, on the other hand, rejected the loan application altogether.

Table 3.2.2 also demonstrates how the effective rate varies from 16.44 percent to 25.71 percent, something a borrower may underestimate when applying from different lenders. Even though the individual is capable of repaying the principal amount within maturity, the effective rate can increase the total cost of a consumer loan by more than half. This can be illustrated with a simple example. The total cost of a loan is calculated using the following formula, with principal amount P, effective monthly interest rate r and total number of payments n

$$\textit{Total cost} = P * \frac{r(1+r)^n}{(1+r)^n-1} * n - P$$

Applying this formula to a loan of 150 000, with an effective yearly rate of 19 percent and a maturity of five years, will result in a total cost of 83 465, excluding any signing fees[17]. This is 55.64 percent of the principal amount, and even though all consumer banks are obligated to include an example with the effective rate and total cost when promoting a loan, customers may disregard it or fail to apply it to their own application.

[17] Many consumer banks charge a signing fee of 800 - 1 000 NOK. The debtor can often choose to deduct this fee from the received loan amount.

The information in this subchapter suggests that people may default on their consumer loans because the costs of the loan are too high, and that lenders can affect defaults by adjusting the requirements of the loan.

3.3 Effects of default

In order to gain an improved understanding of the importance of estimating defaults on consumer loans, this subchapter will analyse the effect of loan default on both the borrower and the lender.

3.3.1 Effect on the borrower

After a borrower is delinquent on the loan, the lender will proceed by sending additional invoices and warnings, as illustrated earlier in Table 3.2.1. This results in increasing costs for the borrower, as each invoice will include an additional fee. On top of this is an interest fee of 8.5 percent (Regjeringen, 2018). When the loan defaults and the responsibility of payment collection is transferred to a debt collection company, they will charge additional, higher fees for their services[18], which will further increase the cost for the borrower. If the borrower still has not repaid the loan and corresponding charges, the lender may file a lawsuit against the borrower. If the court rules in favour of the lender, the borrower's income will be docked and the person might be forced to sell off assets in order to repay the loan (Tvangsfullbyrdelsesloven, 1993, §7-2).

In addition to the economic burden of default, borrowers also receive a payment remark after defaulting on their debt[19]. In Norway, Brønnøysundregistrene collects payment remarks from different lenders and registers them in Løsøreregisteret, only accessible for approved agents such as credit agencies (Brønnøysundregistrene, 2017). When a person applies for a loan, the lender can request the person's history of payment remarks from one of these approved agents for a fee. If a payment remark is found, the lender usually rejects the loan application. Loan

[18] For more information on the costs and process of debt collection in Norway, see Finansportalens' "Verdt å vite om inkasso og betalingsanmerkninger" (Finansportalen, 2018)

[19] A payment remark remains on record until the debtor has repaid the loan, or four years have passed since the date of the default without additional actions from the lender or the debt collection company (Foreldeseslovenen, 2018b).

default therefore reduces the borrower's financing ability for several years, in addition to the immediate costs.

3.3.2 Effect on lenders

Loan defaults are also a concern for the lenders. While it is common for many banks to sell defaulted loans to a debt collection company[20] (Ekeseth, 2018, Null TDN Finans, 2015, Trumpy & Christensen, 2018), the debt collection company will always pay below the principal amount, as the loans are highly risky. This usually incurs a loss for the lender, upwards 30 percent for the most risky loans according to Andersen-Ulven (Andersen-Ulven, personal communication, 16 November 2018). Although the lender receives income when selling a defaulted loan, stricter rules for loss-recognition were introduced January 2018 in IFRS 9 (International Accounting Standards Board, 2014). Whereas lenders only needed to recognise a loss after the loan had defaulted before 2018, they now have to perform continuous risk estimation on their outstanding loans, and reduce the reported value for loans that are considered more risky than before[21]. The result of this is that defaulting loans will lower reported profitability of banks at an earlier stage.

The lender also runs the risk of going bankrupt if large amounts of the loans default. When the income from outstanding loans is reduced, the lender will need to take on additional debt to avoid problems with liquidity[22]. Consequently, a credit agency will usually lower the credit rating of the lender, making it more difficult to raise additional debt from the market (Kisgen, 2006, p.1039-1040). If this downward trend continues, the lender will reach a point where it is unable to cover its expenses. This results in bankruptcy[23].

[20] The loans are sold either on a regular basis, called a forward flow agreement, or as a one-time portfolio.

[21] For readers interested in reading an overview of the new impairment model, see PwC's publication «Moving from incurred to expected credit losses for impairment of financial assets is a game changer» (PwC, 2014).

[22] Liquidity refers to a company's ability to cover short-term expenses, such as salaries, taxes and withdrawn deposits.

[23] For information on costs of bankruptcy, see «The costs of bankruptcy – a review» by Ben Branch (Branch, 2002)

4. Data description

In this chapter, the data collected from the bank will be presented and evaluated. The data will then be used in Chapter 5 and Chapter 6 for further analysis.

4.1 Descritption of the data

The original data provided by the bank includes 17 784 approved consumer loans in Norway. The observations consist of 29 variables, both numerical and categorical, which are inputs from the borrowers and the bank. The data is cross-sectional[24], and ranges over three years. To secure anonymity, the bank supplied no data that can identify the borrowers, such as name or location.

4.1.1 Inputs from borrowers

Out of the 29 variables, the applicants for consumer loans provided 19 either through an online form on the bank's website or to an agent[25]. The inputs can be divided into two main groups: demographic and financial information. The demographic information relates to characteristics such as age, gender, education, employment, civil status, and number of children. The financial information refers to income, debt and expenses.

4.1.2 Inputs from bank

The remaining 10 variables in the data was inputted by the bank. Three variables regarding the borrower's wealth was collected from the national tax database, while seven variables involves the characteristics of the loans, such as loan amount, interest rate and principal free months.

[24] Cross-sectional data means that the observations are collected at a single point in time, when the loan was approved. Opposite of time-series data (Biørn, 2013).

[25] In Norway, a range of independent and bank agents connects debtors to lenders in exchange for commission.

4.2 Cleaning the data

Before analysing and working with a dataset, it is important to examine and adjust for missing data and other factors that might distort any statistical results (Hair, Black, Babin & Anderson, p.31, 2014). The cleaning process of the data involved removing outliers, adjusting variables and excluding recent observations. Appendix B describes each step of the cleaning process. Appendix C shows how each step affected the dataset, in terms of remaining observations and variables.

4.3 Descriptive statistics

Descriptive statistics for the financial variables included in the dataset are presented in Tables 4.3.1 to 4.3.3. The tables differentiate between defaulted loans and non-defaulted loans, showing the differences in means for each variable. The t-statistic tells if the difference between the means is significantly different from zero.

Table 4.3.1 presents descriptive statistics for the demographic variables in the dataset. Since each variable is binary, the means are presented as proportions of defaulted loans and non-defaulted loans.

	Defaulted	**Non-defaulted**			
Variable	Mean (1385)	Mean (9451)	Std. Error	T-statistic	
Gender					
Male	0.7242	0.6774	0.0129	-3.62	***
Female	0.2758	0.3226	0.0129	3.62	***
Co-Signer	0.0462	0.1382	0.0067	13.80	***
Education					
Master	0.0939	0.1552	0.0087	7.07	***
Bachelor	0.3162	0.3218	0.0134	0.41	
Unfinished Educ.	0.0838	0.0836	0.0080	-0.02	
High School	0.4534	0.3951	0.0143	-4.08	***
Elementary	0.0527	0.0443	0.0064	-1.31	
Employment					
Permanent	0.8693	0.8412	0.0098	-2.87	***
Permanent Public	0.0144	0.0177	0.0035	0.93	
Temporary	0.0051	0.0059	0.0021	0.42	
Self Employed	0.0267	0.0374	0.0048	2.24	**
Retired	0.0116	0.0243	0.0033	3.90	***

Disabled	0.0708	0.0715	0.0074	0.10	
Other	0.0014	0.0013	0.0011	-0.16	
Student	0.0000	0.0002	0.0001	1.41	
WAA	0.0000	0.0003	0.0002	1.73	*
Civil Status					
Married	0.2578	0.3954	0.0128	10.76	***
Domestic Partner	0.2318	0.2226	0.0121	-0.75	
Divorced	0.0599	0.0487	0.0068	-1.67	*
Widowed	0.0094	0.0109	0.0028	0.54	
Unmarried	0.4332	0.3132	0.0141	-8.48	***
Separated	0.0079	0.0092	0.0026	0.49	
Living Arrangement					
Condominium	0.0816	0.0801	0.0079	-0.19	
Cooperative Housing	0.0007	0.0021	0.0009	1.62	
Private Owner	0.4736	0.6001	0.0143	8.82	***
Tenant	0.4440	0.3176	0.0142	-8.91	***

Significance levels: *** p<0.01, ** p<0.05, * p<0.1

Table 4.3.1: Categorical inputs from borrowers

Variable	**Defaulted** Mean (1385)	**Non-defaulted** Mean (9451)	Std. Error	T-statistic	
Amounts in NOK 1000 unless specified with ’					
Income/expenses					
Gross Income	509.63	550.51	5.52	7.40	***
Income Rent	9.88	13.89	0.97	4.13	***
Insurance Payment’	4.33	4.00	4.64	-0.07	
Other Income	2.45	3.05	0.37	1.63	
Rent Expenses	26.04	19.73	0.93	-6.77	***
Debt					
Mortgage	727.33	964.22	25.86	9.16	***
Other Loans	145.46	184.13	4.72	8.19	***
Student Loan	27.01	30.69	1.96	1.88	*
Refinancing	80.63	122.97	3.73	11.35	***
Wealth					
Wealth Time -1	67.16	59.98	8.46	-0.85	
Wealth Time -2	51.61	43.69	6.55	-1.21	
Wealth Time -3	43.49	37.33	6.21	-0.99	
Other					
Age’	39.38	43.30	0.31	12.66	***
Children under 18’	0.62	0.62	0.03	-0.11	

Significance levels: *** p<0.01, ** p<0.05, * p<0.1

Table 4.3.2: Numerical inputs from borrower

Variable	Defaulted Mean (1385)	Non-defaulted Mean (9451)	Std. Error	T-statistic	
Loan Attributes					
Loan Amount (*NOK 1000*)	173.90	232.07	4.25	13.7	***
Difference from applied (*NOK 1000*)	-21.00	-14.31	2.35	2.85	***
Duration	120.40	128.44	1.54	5.21	***
Principal Free Months	6.41	4.66	0.37	-4.67	***
Interest Rate	15.77	14.35	0.07	-19.1	***

Significance levels: *** p<0.01, ** p<0.05, * p<0.1

Table 4.3.3: Inputs from bank

4.4 Evaluation of the data

4.4.1 Validity

Although the dataset only contain observations from one consumer bank, the data is considered valid for the Norwegian market. Several of the characteristics, including the distribution of age groups (Finanstilsynet, 2018e) and the interest rate (Hagen et al., 2017), are approximately the same for the dataset and the market. The large amount of observations, and the cleaning process described in the appendices, also increases its representation of the market.

4.4.2 Limitations

The dataset solely involves one-time consumer loans, and not credit card loans. Even though credit card loans contribute to 45 percent of the total consumer debt in Norway (Finanstilsynet, 2018e) and share similar attributes, the applicability to credit card loans is uncertain. This is partly due to the difference in interest rates and approving processes between the two types of loans.

Since the dataset only consists of Norwegian customers, it is uncertain if the data is applicable to other countries. Demographic and cultural factors may result in other characteristics being more dominant in defaults, while varying market conditions can affect the inputs of the lenders. For example, marital status in another country might not have the same financial implications as in Norway. Similarly, different market conditions might raise or lower the interest rates provided by lenders, affecting the number of defaults and estimations of a model.

5. Using characteristics for modelling purposes

The purpose of this chapter is to analyse the characteristics of the borrowers and the loans, and decide if suitable for predicting probability of default.

To examine how the characteristics affect the risk of default, the variables are differentiated by default and non-default. Differences in means between defaults and non-defaults are then used as a comparable measure on the risk of default and a simple two-tailed T-Test is applied to investigate if the differences are significantly different from zero. Of 10 836 borrowers, 1 385 have defaulted on their loans. This results an overall default rate of 12.78 percent, which will be used a benchmark for the analysis.

5.1 Demographics

This subchapter will give a summary of the descriptive statistics of the demographic variables. Most of the variables in this category are dummy-variables with a value of one if a person belongs to the group and zero if otherwise. Therefore, the means are presented as proportions of total defaults and total non-defaults.

The differences in proportions of defaults and non-defaults between the groups can give an indication of how each group performs in servicing their consumer loans, relative to the other groups. If a group has a higher proportion of defaults compared to non-defaults, this can imply that the group is more likely to default.

5.1.1 Age

Defaulting borrowers are on average younger than non-defaulting borrowers, as can be seen from Table 4.3.2. To further analyse how age affect default rates, borrowers are split into five different age groups.

Age Group	*Default*	*Non-default*	*Total*	*Difference*	*T-statistic*		*Default Rate*
18 – 29	21.2 %	13.3 %	14.3 %	7.9 %	-6.85	***	18.9 %
30 – 39	33.9 %	26.4 %	27.3 %	7.6 %	-5.61	***	15.9 %
40 – 49	25.6 %	30.1 %	29.6 %	-4.5 %	3.57	***	11.1 %
50 – 59	14.7 %	21.1 %	20.3 %	-6.4 %	6.1	***	9.3 %
60 -	4.6 %	9.2 %	8.6 %	-4.6 %	7.26	***	6.8 %

Significance levels: *** p<0.01, ** p<0.05, * p<0.1

Table 5.1.1: Age groups

The results shown in Table 5.1.1 suggest a negative linear relationship between default and age, where the probability of default decreases with an increase in age. Defaults are most frequent among the youngest borrowers, with a default rate of almost 19 percent. People from the ages 30 to 39 also have a default rate above the overall average. The oldest group has the smallest default percentage of 6.5 percent.

5.1.2 Education

Education	*Default*	*Non-default*	*Total*	*Difference*	*T-statistic*		*Default Rate*
Master	9.4 %	15.5 %	14.7 %	-6.1 %	7.07	***	8.1 %
Bachelor	31.6 %	32.2 %	32.1 %	-0.6 %	0.41		12.6 %
Unfinished Educ.	8.4 %	8.4 %	8.4 %	0.0 %	-0.02		12.8 %
HighSchool	45.3 %	39.5 %	40.3 %	5.8 %	-4.08	***	14.4 %
Elementary	5.3 %	4.4 %	4.5 %	0.8 %	-1.31		14.8 %

Significance levels: *** p<0.01, ** p<0.05, * p<0.1

Table 5.1.2: Education

Approximately half of the loans are held by clients with a bachelor's degree or higher. This group is responsible for 43 percent of the defaults, making lower education responsible for 57 percent of the defaults. There is no significant difference between defaults and non-defaults on bachelor level, and as a result, it is hard to say if having a bachelor's degree will influence the risk of default. However, the differences in defaults for borrowers with a master's degree are significant and, given the default rate of 8.1 percent, can reduce the risk of default. A study from SSB in 2018 shows that the average salary for a wage earner with a master's degree is 25 percent higher than employees with a bachelor's degree. (Bye, 2018) The correlation[26] between Gross Income and Master in the dataset is 0.25, which further supports this relationship. In contrast, the correlation between Gross Income and Bachelor is merely 0.02.

High School is the other group that has a significant difference in means at the one percent level. The negative difference of 5.95 percent suggests that a person is more likely to default on a consumer loan if the highest form of education obtained is from high school.

[26] A correlation matrix with some selected variables is presented in Appendix D

5.1.3 Employment

Employment	*Default*	*Non-default*	*Total*	*Difference*	*T-statistic*		*Default Rate*
Permanent	86.9 %	84.1 %	84.5 %	2.8 %	-2.87	***	13.2 %
Self Employed	2.7 %	3.7 %	3.6 %	-1.1 %	2.24	**	9.5 %
Retired	1.2 %	2.4 %	2.3 %	-1.3 %	3.90	***	6.5 %

Significance levels: *** p<0.01, ** p<0.05, * p<0.1

Table 5.1.3: Employment (excerpt)

Being self-employed or retired seems to lower the default rate to respectively 9.5 percent and 6.5 percent, while permanent employees are slightly above the overall average default rate. However, as more than 84 percent of the borrowers belong to the permanently employed group, it is probable that the increase in probability is a result of other factors.

5.1.4 Living Arrangement

Living Arrangement	*Default*	*Non-default*	*Total*	*Difference*	*T-statistic*		*Default Rate*
Condominium	8.2 %	8.0 %	8.0 %	0.2 %	-0.19		13.0 %
Cooperative Housing	0.1 %	0.2 %	0.2 %	-0.1 %	1.62		4.7 %
Private Owner	47.4 %	60.0 %	58.4 %	-12.7 %	8.82	***	10.4 %
Tenant	44.4 %	31.8 %	33.4 %	12.6 %	-8.91	***	17.0 %

Significance levels: *** p<0.01, ** p<0.05, * p<0.1

Table 5.1.4: Living Arrangement

Most of the clients are either tenants or private owners of real estate. Test results show a significant difference in defaults in the two categories: Tenants have an average default rate of 17 percent, while private owners have a default rate of 10.4 percent. A reason for the large difference might be the economic advantage of owning real estate in Norway[27].

5.1.5 Civil status

Civil Status	*Default*	*Non-default*	*Total*	*Difference*	*T-statistic*		*Default Rate*
Married	25.8 %	39.5 %	37.8 %	-13.8 %	10.76	***	8.7 %
Domestic Partner	23.2 %	22.3 %	22.4 %	0.9 %	-0.75		13.2 %
Divorced	6.0 %	4.9 %	5.0 %	1.1 %	-1.67	*	15.3 %
Widowed	0.9 %	1.1 %	1.1 %	-0.2 %	0.54		11.2 %
Unmarried	43.3 %	31.3 %	32.9 %	12.0 %	-8.48	***	16.9 %
Separated	0.8 %	0.9 %	0.9 %	-0.1 %	0.49		11.2 %

[27] Owning property financed by debt is highly advantageous in Norway, as private home owners get a 23 percent tax refund for interest rate costs (Skatteetaten, 2018). There is no such arrangement for rent expenses.

Significance levels: *** p<0.01, ** p<0.05, * p<0.1

Table 5.1.5: Civil Status

The difference in means between defaulted and non-defaulted loan takers that are married is high. Significant at the one-percent level, the mean for defaulted loan takers is 39.54 percent, while the mean for the non-defaulted loan takers is 25.78 percent. This suggests that married people have a significantly lower probability of default than unmarried people do.

5.1.6 Gender and co-signer

	Default	*Non-default*	*Total*	*T-statistic*		*Default Rate*
Male	72.4 %	67.7 %	68.3 %	-3.62	***	13.5 %
Female	27.6 %	32.3 %	31.7 %	3.62	***	11.1 %
Co-Signer	4.6 %	13.8 %	12.6 %	13.80	***	4.7 %

Significance levels: *** p<0.01, ** p<0.05, * p<0.1

Table 5.1.6: Gender and co-signer

As seen in Table 5.1.6, men are more likely to default on consumer loans than women. One reason for this can be that men are less averse to risk, as shown in a study on gender differences and risk taking by Byrnes, Miller and Schafer in 1999 (Byrnes et al., 1999, p. 367-383).

Having co-signer significantly reduces the probability of default. This makes sense intuitively, as the risk is spread across two individuals instead of one.

5.2 Financial statistics

5.2.1 Income

Mean income in the dataset is 549 238 and is close to the overall average in Norway [28]. Mean income in the selection of defaulted loans is 509 631 NOK, and 550 505 NOK in the selection of non-defaulted loans. The results from the t-test suggest that there is a significant difference in income between the two groups, but the difference is only eight percent. To analyse the income of the customers further, the population is divided into income groups.

[28] According to SSB, the average income in Norway in 2017 was 531 720 NOK (SSB, 2017)

Income Group (In thousands)	*Default*	*Non-default*	*Total*	*Difference*	*T-statistic*		*Default Rate*
0 – 250	0.9 %	0.5 %	0.6 %	0.4 %	-1.52		20.6 %
250 – 450	47.5 %	37.7 %	38.9 %	9.8 %	-6.86	***	15.6 %
450 – 650	35.2 %	39.8 %	39.2 %	-4.6 %	3.34	***	11.5 %
650 – 850	11.1 %	14.0 %	13.6 %	-2.9 %	3.21	***	10.4 %
850 – 1 000	3.2 %	4.2 %	4.1 %	-1.0 %	1.99	**	10.0 %
1 000 000 –	2.2 %	3.8 %	3.6 %	-1.7 %	3.77	***	7.7 %

Significance levels: *** p<0.01, ** p<0.05, * p<0.1

Table 5.2.1: Income groups

Table 5.2.1 shows that borrowers earning less than 450 000 NOK are more likely to default. The income group 250 - 450 has a default rate of 15.6 percent and a significantly higher proportion of the defaulted loans than non-defaulted loans. The difference in the lowest income group is barely significant at the 90 percent confidence interval. The conclusion is that the default rate declines with increasing income. This is not surprising, as one would expect that higher earners have a better chance of repaying a loan.

5.2.2 Debt

Lenders often require loan applicants to provide information on their debt. The dataset has three kinds of debt: mortgage, student loan and other loans.

Debt (NOK 1000)	*Default*	*Non-default*	*Difference*	*T-statistic*	
Mortgage	727.33	964.22	-236.89	9.16	***
Other Loans	145.46	184.13	-38.67	8.19	***
Student Loan	27.01	30.69	-3.68	1.88	*
Refinancing	80.63	122.97	-42.34	11.35	***

Significance levels: *** p<0.01, ** p<0.05, * p<0.1

Table 5.2.2: Debt

The results from testing the differences in means between defaults and non-defaults indicate that having more debt reduces probability of default. Differences in means for Mortgage and Other Loans are highly significant. Average mortgage debt for non-defaulted customers is 32 percent higher than defaulted.

Mortgage is positively correlated with both Gross income (0.51) and Private Owners (0.59), suggesting that customers with a mortgage has a stronger financial position, ceteris paribus. In most cases, a mortgage is backed up by an underlying asset, which further supports this theory. Finally, it is also worth mentioning that homeowners have successfully passed a credit assessment when applying for a mortgage.

Explaining the differences in means for Other Loans proves much more difficult, as it is not clear what types of loans are included in this variable. Refinancing means that the applicant are borrowing to pay existing debt or to renegotiate terms on an old lending contract. A higher refinancing amount leading to lower probability of default could be the fact that refinancing often lower costs (Nordea, 2018).

Debt ratio is a measure used in the banking industry to evaluate the debt burden of a loan applicant (Finansdepartementet, 2016). Dividing all existing debt, including Mortgage, Student Loan and Other Loans, by Gross Income returns the debt ratio of the borrower before the issuance of the consumer loan. This will control for the income effect seen previously in this subchapter. Then, testing for differences in mean debt ratio between defaults and non-defaults will give a more accurate view of the relationship between debt and default.

Debt ratio	*Default*	*Non-default*	*Total*	*Difference*	*T-statistic*		*Default Rate*
0-1	46.9 %	33.5 %	35.2 %	13.4 %	-9.41	***	17.0 %
1-3	32.2 %	38.3 %	37.5 %	-6.1 %	4.48	***	11.0 %
3-5	18.1 %	23.8 %	23.1 %	-5.7 %	5.06	***	10.0 %
5-	2.7 %	4.4 %	4.2 %	-1.7 %	3.45	***	8.3 %

Significance levels: *** p<0.01, ** p<0.05, * p<0.1

Table 5.2.3: Debt ratios

The test supports the notion of the mean debt ratio being lower for those defaulting on their loans. Groups with a debt ratio lower than three, experience significantly more defaults than groups with a debt ratio more than three. Even when controlling for the income effect, having a higher debt ratio seems to reduce the probability of default.

5.3 Loan attributes

All characteristics discussed so far are taken from variables that are not influenced by the bank. The attention will now shift to the inputs of the bank: interest rate, loan amount, duration and principal free months. The challenge with analysing these variables is that they are already affected by an internal risk assessment made by the bank. For example, two clients with the same loan amount may have different interest rates based on credit scores.

Loan Attributes	*Default*	*Non-default*	*Difference*	*T-statistic*	
Loan Amount (*NOK 1000*)	173.90	232.07	-58.17	13.69	***
Difference from Applied *(NOK 1000)*	-21.00	-14.31	-6.69	2.85	***
Duration	120.40	128.44	-8.04	5.21	***
Principal Free Months	6.41	4.66	1.75	-4.67	***
Interest Rate	15.77	14.35	1.42	-19.13	***

Significance levels: *** p<0.01, ** p<0.05, * p<0.1

Table 5.3.1: Loan Attributes

5.3.1 Loan Amount

Difference in means between defaults and non-defaults is approximately 58 000. With a significant T-statistic, a lower loan amount indicates an increased chance of defaulting. This seems counter-intuitive from an economic perspective. A lower amount should lessen the debt burden and thereby the monthly costs of the loan, consequently lowering the default risk. Instead, it is likely a result of the bank's unobserved risk assessment, providing larger amount to safer customers.

5.3.2 Difference from Applied and Duration

The difference from the applied loan amount is lower for the non-defaulting loans than the defaulting loans, suggesting that the bank has reduced the approved loan amount more for risky borrowers.

Probability of default seems to be higher for loans with longer maturity. An explanation for this is that duration increases the total cost of the loan, as seen in subchapter 3.2.3. But the difference in mean is approximately seven percent, compared to a difference in means of about 25 percent for the loan amount, suggesting the marginal effect on probability of default is low.

5.3.3 Principal Free Months

Since principal free months will increase the cost of the loan by postponing the down-payments of the principal, it would be fair to assume that the probability of default increases with principal free months. As seen in Table 5.3.1, this assumption holds. The mean for defaulted consumers is 34 percent higher than non-default and significant in the t-test, suggesting that more principal free months will increase the probability of default.

5.3.4 Interest Rate

The mean interest rate is 14.4 percent, which is close to the estimated market mean discussed in Chapter 2. The bank sets the interest rate for each loan separately, based on a risk assessment of the client. Comparing the mean interest rate between defaults and non-defaults can provide information about the relationship between default risk and interest rates.

Results from the t-test show that defaulted loans have a higher interest rate, even though the difference is moderate – roughly 1.5 percentage points.

6. Modelling risk of default as a lender

The results from Chapter 5 suggest that characteristics of the borrowers and the loans are appropriate for predicting probability of default. In this chapter, two predictive models based on the dataset will be presented, one including inputs from the lender and one without, which can be used to reduce the number of defaults and the risk of the lenders.

6.1 Purpose of the models

The purpose of the models is to predict the probability of default on a consumer loan, based on a range of variables provided by both the borrower and the lender. After a lender receives a loan application with inputs from the borrower, it can use one of the models to estimate the probability of default on the loan. This way, the lender can offer loans that match the lender's risk preference, and reduce defaults.

The main target of the models is an agent at a bank who approve consumer loans, but the models can also be used by a risk-officer who is interested in estimating the total risk of a bank's outstanding loans, or a policy analyst who would like to evaluate the risk of default in the market and advice on future policies.

6.2 Methodology

This subchapter will explain the reasons for why logistic regression is appropriate for a predictive model on default and provide a definition of the method. The subchapter will start by describing the linear regression method, which provides the foundation for logistic regression.

6.2.1 Moving beyond linear regression

For a predictive model on default, a method is needed that estimates the probability of default for a loan, written as $p(Y=1)$, based on the variables from the borrower and the lender, X_1, $X_2, \cdots, X_n$. Combining these two statements results in the following formulation

$$p(X_1, X_2, \cdots, X_n) = p(Y = 1 | X_1, X_2, \cdots, X_n)$$

Multiple linear regression, a widely used statistical learning method, assumes that a relatively linear relationship exists between the dependent variable Y and the independent variables X_1,

$X_2, \cdots, X_n$, which can be calculated using separate coefficients, defined as $\beta_1, \beta_2, \cdots, \beta_n$, for each independent variable (James et al, p.71, 2017). The linear function used in linear regression is shown below

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \in,$$

where β_0 is the intercept and $\in$ is the error term[29].

Multiple linear regression uses the data to compute coefficients and intercept that minimizes the least square criterion[30]. Applying this to a predictive model for default results in

$$p(Y = 1) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_n X_n + \in,$$

where $\hat{\beta}$ are estimates of the coefficients and the intercept.

While it is possible to use this model to predict default on a loan, one drawback makes its predictive ability very limited: Since the dependent variable is a binary variable, taking on either the value one (default) or zero (non-default), the predictive output of the model should take on values between one and zero as well. Linear regression often fails to do so, as the straight-line fit will produce values that are below zero or above one for extreme observations (James et al, p.131, 2017). This can be illustrated by running a linear regression on the Default variable using Gross Income as the sole predictive variable[31].

[29] A summation of the measurement error, assumed to be independent of the variables and with a mean of zero (James et al, p.16, 2017).

[30] The lowest possible value of $S = \sum_{i=1}^{n} \in_i^2$. (Draper & Smith, 1998, p.22-23)

[31] The data used for the linear regression, and the logistic regression presented on the next page, includes outliers that were originally removed in the cleaning process of Chapter 4, for illustrational purposes.

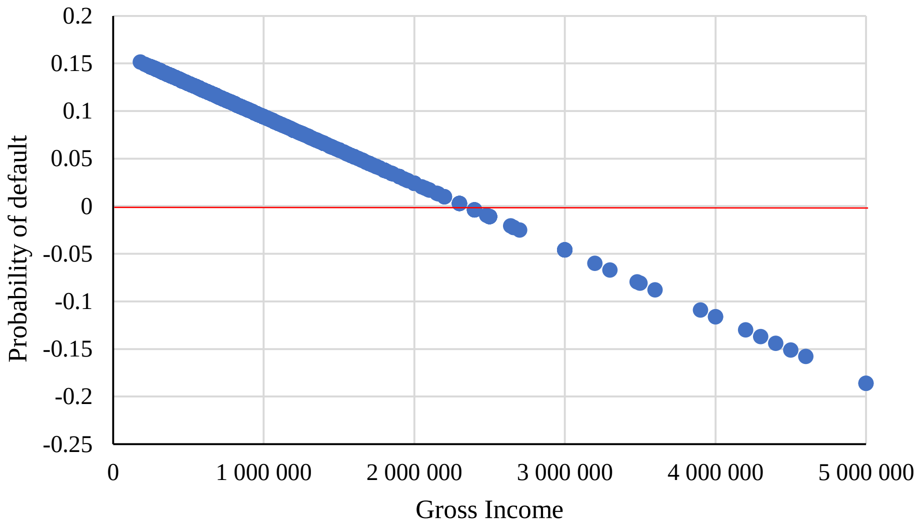


Figure 6.2.1: Probability using linear regression

The probability of default decreases with increasing income, which is intuitive as a higher income will increase the borrower's capability of affording a loan, ceteris paribus. However, observations with Gross Income of more than 2.3 million NOK receive a predicted probability of default below zero, which is both difficult to interpret and distorts the model. Logistic regression fixes this problem.

6.2.2 Logistic Regression

In order to fit the model for outputs between zero and one for all observations, logistic regression uses the logistic function

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n}}$$

to compute the probability of X, where e is a mathematical constant and the base of the natural logarithm (Hosmer et al, p.7, 2013)

By raising the linear function to the power of e, the logistic function ensures that p(X) always will lie between zero and one. This can be confirmed by running a logistic regression on the Default variable once more, instead of a linear regression, using only the Gross Income as input. The output is shown below.

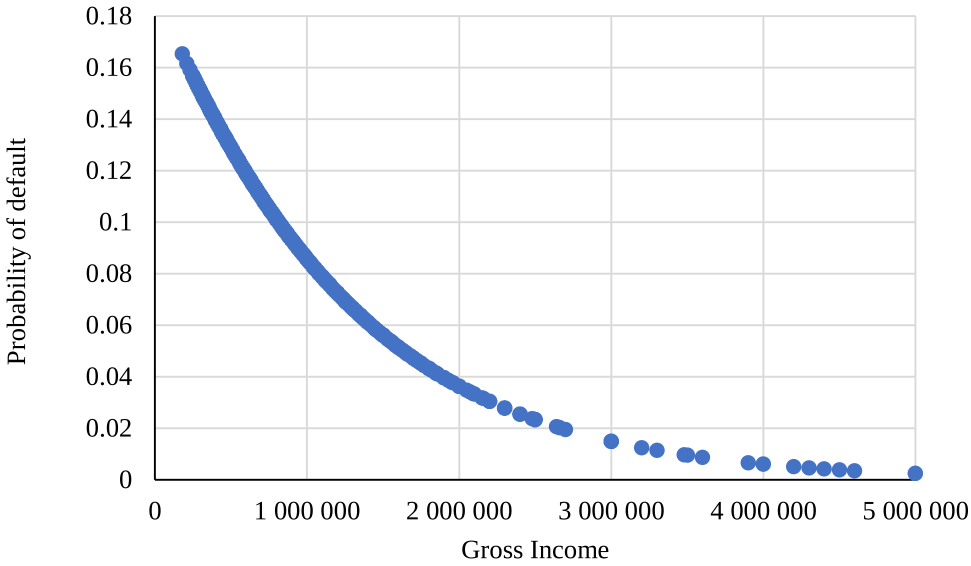


Figure 6.2.2: Probability using linear regression

As Figure 6.2.2 demonstrates, p(X) never drops below zero or above one in the logistic regression. Observations with incomes above 2.3 million NOK receive probabilities of default very close to zero, but never below. It is also worth mentioning that the highest probability of default is 0.16. The reason for this is that the relationship between Default and Gross Income is not strong enough to predict higher probabilities, and more variables are needed.

Moving on with logistic regression for the predictive model, the final formula can be written as

$$p(Y = 1) \; = \frac{e^{\widehat{\beta}_0+\widehat{\beta}_1 X_1+\widehat{\beta}_2 X_2+\cdots+\widehat{\beta}_n X_n}}{1 + e^{\widehat{\beta}_0+\widehat{\beta}_1 X_1+\widehat{\beta}_2 X_2+\cdots+\widehat{\beta}_n X_n}},$$

where $\hat{\beta}$ are computed estimates of the coefficients and the intercept.

While linear regression uses least squares to find the estimates for the coefficients, logistic regression uses an approach called maximum likelihood (Hosmer et al, p.9, 2013). The fitting procedure of this approach is quite extensive and will not be described in depth in this thesis, but in short, it calculates estimates that give the defaulting observations a value as close to one as possible, and the remaining observations a value as close to zero as possible. In order to so, it uses the following likelihood function:

$$\ell(\beta_0, \beta_1, \cdots, \beta_n) = \prod_{i:y_i=1} p(X_i) \prod_{i\prime:y_{i\prime}=0} (1 - p(X_{i\prime}))\,,$$

where ℓ is the likelihood function.

6.3 Subset selection method

When the number of variables in a dataset is large, some variables will likely reduce the accuracy of the model, as they do not have a significant causal effect on the dependent variable (Hair Jr. et al, p.169, 2014). By including these variables, the estimated coefficients for the other predictors will be further away from their true values, and it will be more difficult to make good predictions. In this subchapter, a subset of variables will be determined using stepwise backward selection and Mallow's Cp in order to exclude bad predictors and avoid this problem.

6.3.1 Stepwise backward selection

Stepwise backward selection, proposed by Efroymson in 1960, is a selection method which starts out with running a regression on all of the available variables and then removing one variable at the time to decide the best model for each given number of predictors (Garside, p.196-200, 1965). The best number of predictors is decided afterwards using one or more measures of model fit. At each step, stepwise backward selection removes the variable that is least significant for the model and simultaneously revaluates the remaining predictors and replaces those that fall below a given level of significance, for example five percent.

Performing stepwise backward selection on the dataset results in 43 different models, one for each set of predictors. The next thing to do is decide how many predictors to use by measuring model fit.

6.3.2 Mallow's Cp as measure of model fit

Countless measures of fit have been proposed over the years, including residual sum of squares (RSS), adjusted R-squared, Mallow's C_P and Bayesian information criterion[32]. All of the measures has advantages and disadvantages, and for reasons explained in this subchapter, Mallow's C_P will be used for this model.

[32] Readers interested in learning more about possible measures of fit, logistic regression,or how to build predictive models in R can find more about this in An Introduction to Statistical Learning with applications in R, by James et al.

The RSS is a measure of the disparity between the predicted values $\hat{y}_i$ of a model and the actual values y_i (Draper & Smith, p.29, 1998), calculated by

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Although a lower RSS indicates better predictions, selecting a model solely based on RSS runs the risk of overfitting the model[33]. In fact, the RSS will always decrease as more variables are added to the model, and one will therefore end up with all of the possible variables if using RSS (James et al, p.210, 2013). Mallow's C_P is one of the measures that offers a solution to this problem of overfitting, by penalising the addition of predictors to the model. Mallow's C_P achieve this using the following function

$$C_p = \frac{1}{n}\,(RSS + 2d\hat{\sigma}^2),$$

where *n* is the number of observations, *d* is the number of predictors and $\hat{\sigma}$ is an estimate of the variance error of the model (James et al, p.211, 2013).

As *d* increases by one unit, the RSS of the model needs to decrease by more than $2\hat{\sigma}^2$ in order for the C_p statistic to decrease. This means that the model with the lowest C_p statistic will exclude all variables that do not sufficiently decrease the model's RSS.

Another aspect of the Mallow's C_p that needs to be considered is the trade-off between model variance and bias. While a lower C_p statistic indicates a lower variance, a smaller distance between C_p and the number of predictors *d* suggest a low bias in the model (Draper & Smith, 1998, p.332). When the C_p statistic is significantly lower than *d* the model likely has an overfitting bias, and when *d* is higher than C_p it is a case of underfitting[34]. The optimal C_p statistic is therefore one that is close to the number of predictors, while as low as possible.

Comparing the C_p statistics of the models found by stepwise backward selection to the number of predictors results in Figure 6.3.1 below.

[33] Overfitting relates to the problem of building a model too closely to the given data, and thus failing to make good predictions on new observations (Cawley & Talbot, 2010, p.2084-2086). An overfitted model will falsely use predictors that have no or very little real effect on the dependent variable, or provide poor coefficient estimates.

[34] Opposite of overfitting; instead of including bad predictors, an underfitting model excludes good predictors.

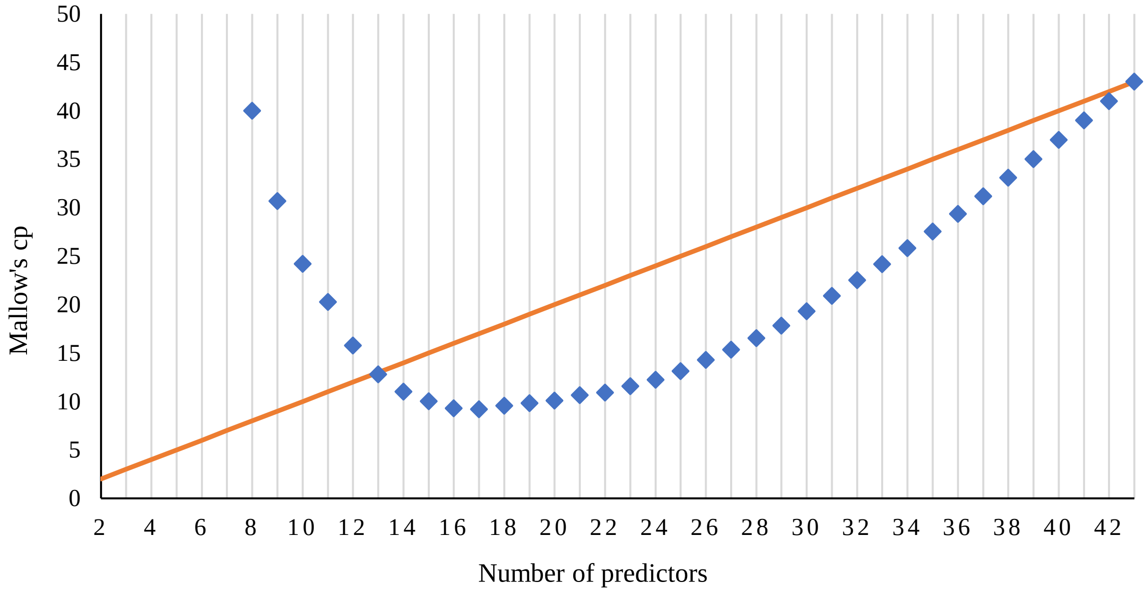


Figure 6.3.1: Mallow's C_p vs Predictors

It is evident that the C_p statistic is very large for models with less than eight predictors. This implies that it is difficult to make good predictions while using only a few variables. Secondly, the models with the lowest C_p statistics contain between 15 and 19 predictors. However, all of these statistics are much lower than the number of predictors, suggesting that the models are overfitted and have included variables that are insignificant on the true output. Based on the criteria that $C_p \approx d$ and that C_p should be as low as possible, the model with 13 predictors seems to be the best fit.

6.3.3 Model with inputs from lender

The 13-predictor model including inputs from the lender found by stepwise backward selection consists of the following variables:

Mortgage	*Loan Amount*	*Domestic Partner*
Principal Free Months	*Interest Rate*	*Married*
Gross Income	*Children_U18*	*Co-Signer*
Student Loan	*Age*	*Man*

Table 6.3.1: Variables in model with lender input

The four variables in the far-right column are dummy variables taking on the value zero or one, and the intercept is the 13th predictor.[35]

A limitation of this model is the inclusion of input from the bank, in the terms of Loan Amount, Interest Rate and Principal Free Months. The bank determines the attributes of the loan based on their own risk assessment of the loan taker, meaning these variables will reflect other variables in the dataset. This problem is known as endogeneity, when one or more of the independent variables are based on other independent variables, and results in biased estimators (Woolridge, 2012, p.86-87). For instance, Loan Amount will likely be higher for a loan taker with favourable characteristics, and might reduce probability of default in the model. This, of course, makes no intuitive sense, as increasing the amount of the loan should increase probability of default, ceteris paribus. In order to provide a less biased alternative, a second model is created using the same dataset but excluding the inputs from the lender[36].

6.3.4 Model without inputs from lender

Optimising stepwise backward regression and Mallow's Cp once more[37], on a dataset without the lender variables, results in a 12-predictor model as shown in Table 6.3.2.

Student Loan	*Age*	*Domestic Partner*
Private Owner	*Man*	*Married*
Gross Income	*Children_U18*	*Co-Signer*
Originally Applied Amount	*Refinancing Amount*	

Table 6.3.2: Variables in model without lender inputs

In addition to the four dummy variables from the first model, Private Owner also takes on the value zero or one. The intercept is the 12th predictor.

[35] The variables are not plotted into the formula for logistic regression to page limitations, but the formula can be found on page 41 and variables equal $X_1$to X_{12}. The coefficients relating to each variable and the intercept will not be presented until next subchapter, after the model has been trained using Monte Carlo cross-validation.

[36] Meaning the variables Loan Amount, Interest Rate, Principal free Months and Duration will be ignored.

[37] Appendix E displays the Mallow's Cp for the second model

6.4 Training and testing the models

Having selected two models, the next step is estimating their coefficients. While it is possible to run logistic regressions on the entire dataset, this will increase the chance of overfitting, as there will be no true way to measure how the models perform on new observations. Instead, the dataset will be split into two parts; one will be used to train the models (training set), and one will be used to measure the accuracy of the models (validation set).

This subchapter will begin by selecting a splitting technique, before presenting the trained models. It will then compare the predictions of the models to the observations in the validation set to measure the prediction accuracy.

6.4.1 Monte Carlo cross-validation

Monte Carlo cross-validation is a powerful approach that splits the dataset into two parts several times (Shao, 1993, p.489). Using a chosen splitting point, the data is split into a larger part, which is treated as a training set, and a smaller part, which serves as a validation set. The training set is then used to produce the coefficients of a model, while the validation set is used to measure the accuracy by comparing the model's predicted output on the observations to the actual output. Using machine learning, this process is repeated for a specified number of times, and the prediction accuracy is averaged over all runs.

Splitting the dataset into the training and the validation set is dependant of two factors: the splitting point and the sampling technique. Robbin and Simon shows in their article from 2011 that a 2/3 to 1/3 split is optimal for datasets with more than 100 observations and many variables (Robbin & Simon, 2011). 2/3 of the dataset will therefore be used as the training set for each run, while 1/3 will be used as validation set. The sampling technique in Monte Carlo cross-validation is simple random sampling, where every observation has an equal chance of being placed in either datasets.

Two limitations of Monte Carlo cross-validation should be highlighted. The first is regarding simple random sampling, and the fact that some observations may only be used for training while others may only be used for testing. This may increase the effect of outliers and bias in the model. However, by running 1 000 data splits, this limitation will be minimised. The second limitation is the overestimation of the test error. Since the Monte Carlo cross-validation does not allow models to use all available observations to estimate the coefficients, the prediction accuracy tend to be lower than if the model was built on the entire dataset. This is important to bear in mind when examining the accuracies of the models later on.

6.4.2 Coefficients and VIFs of the trained models

By running the logistic regression models 1 000 times using Monte Carlo cross-validation with a two-to-one splitting point, the resulting coefficients and variance-inflated factors (VIF) are produced for the two models:

	With lender inputs		**Without lender inputs**	
Predictors	*Coefficient*	*VIF*	*Coefficient*	*VIF*
Loan Amount	-1.05E-06***	1.50		
	(-2.67E-07)			
Mortgage	-9.78E-08*	1.38		
	(4.10E-08)			
Student Loan	-1.78E-06***	1.08	-1.70E-06***	1.08
	(4.82E-07)		(4.79E-07)	
Principal free Months	0.014***	1.04		
	(2.37E-03)			
Interest Rate	0.112***	1.37		
	(0.014)			
Co-signer	-0.804***	1.15	-1.022***	1.12
	(0.150)		(0.148)	
Age	-0.023***	1.19	-0.026***	1.19
	(3.13E-03)		(3.09E-03)	
Married	-0.395***	1.41	-0.398***	1.41
	(0.083)		(0.083)	
Children_U18	0.143***	1.26	0.157***	1.26
	(0.040)		(0.040)	
Gross Income	-5.33E-07**	1.57	-8.93E-07***	1.29
	(2.18E-07)		(1.96E-07)	
Man	0.217***	1.10	0.223***	1.10
	(0.073)		(0.072)	
Domestic Partner	-0.183**	1.21	-0.166*	1.22
	(0.081)		(0.081)	
Originally Applied Amount			-5.91E-07**	1.52
			(2.41E-07)	
Private Owner			-0.279***	1.17
			(0.067)	
Refinancing Amount			-1.23E-06***	1.38
			2.90E-07)	
(Intercept)	-2.129***		0.062	
	(0.303)		(0.162)	

*Significance levels: *** p<0.01 ** p<0.05 * p<0.1*
Standard Errors in pharanteses

Table 6.4.1: Coefficients and VIFs of the models

The VIFs are computed in order to check for multicollinearity in the models. Multicollinearity refers to a situation where collinearity exist between three or more variables, which cannot be observed in a correlation matrix (James et al, p.101, 2013). A higher VIF-statistic indicates a higher degree of collinearity to the other variables[38], and should usually not exceed four (Hair Jr. et al, p.197, 2014). Since none of the VIFs are above four, this suggest the models do not suffer from multicollinearity.

Examining the coefficients, the models seems to be in line with the results from Chapter 5. All of the variables have the same signs as their respective difference in means in Chapter 5, and most of the variables highlighted in Chapter 5 are present in either of the models. However, the sign of the Loan Amount coefficient in the first model is troubling. Although the loan amount is an important factor of the loan, the coefficient produced in the first model does not make intuitive sense - a higher loan amount should increase the probability of default, ceteris paribus. As discussed earlier, this coefficient is likely a result of endogeneity as the bank provided higher loans to safer consumers. The coefficient is highly significant for estimating probability of default, but should not be used for adjusting a loan in order to change the risk, as it would be unwise to increase a loan to reduce risk[39].

Mortgage and Student Loan also need explaining. As discussed in Chapter 5, Mortgage can imply a stronger financial situation and a previous, successful credit assessment, which is why the coefficient is negative. Student Loan, on the other hand, is a bit more difficult to explain. One could argue that a higher student loan is related to a higher income, but the correlation between the two is merely 0.092 in the dataset. Since it was shown as insignificant in Chapter 5, a more plausible explanation is that its included in the models as it captures some information from the missing variables.

[38] The VIF is defined as $\frac{1}{1-R^2}$ where R^2 is the proportion of the variance for the variable explained by the other variables.

[39] Appendix F presents an adjusted model where Loan Amount is excluded, and compares it to the original model. The coefficients are slightly different, and this model can be used for lenders who want to ignore the predictive effect of the loan amount.

For the model without lender input, the coefficients for Originally Applied Amount and Refinancing Amount need further analysis. Based on its coefficient, it seems that people that apply for higher loans are less likely to default. This could be related to a greater cost of defaulting, or that people who apply for higher amount are more certain of their financial situation. Refinancing, as mentioned in Chapter 5, can be positive for avoiding default as it reduces cost for the borrower, and the model seems to support this.

It is also worth mentioning that Duration is the only attribute of the loan that is not included in either model. As discussed in Chapter 5, although it has a significant difference in means, the difference was only seven percent. The low difference might explain why it is not included in the models. Master degree is not represented in the models either, but this is likely a result of the machine learning reducing multicollinearity in the models as the variable was correlated with several of the models variable[40].

Many variations of the presented models can be constructed using other methods and measures of fit. However, the analysis shows that the seven variables present in both of the models, except for Domestic Partner[41], seem to be significant in determining probability of default for consumer loans, regardless of what model is used.

6.4.3 Testing the models

Whether a loan is predicted to default or not depends on the threshold for default. Figure 6.4.1 illustrates the trade-off between captured default and correct predictions for different thresholds using the first model. The trade-off is valid for all thresholds of default, and as the threshold increases, the proportion of correctly classified loans will increase at the cost of predicted defaults.

[40] This be seen in the correlation matrix in Appendix D.

[41] Domestic Partner was only significant at 10 percent in second model, and was not proven significant in Chapter 5.

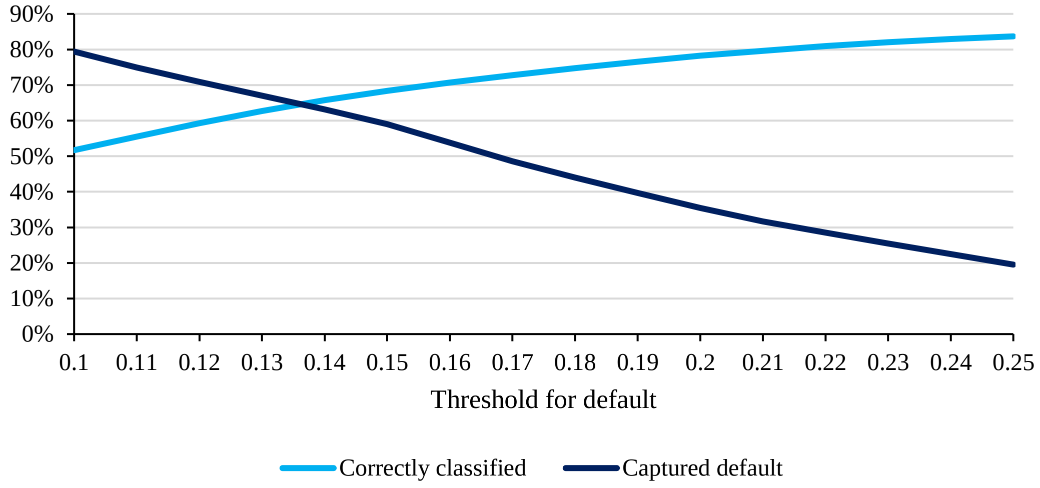


Figure 6.4.1: Trade-off for thresholds of default

Although correctly classified and captured default are the same around the 14 percent threshold, this does not necessarily mean this is the optimal threshold for default. The optimal threshold for default depends on the risk profile of the bank, but Figure 6.4.1 shows that the trade-off is worse for higher thresholds. It is also evident that a threshold of 50 percent will capture minimal of the defaulted observations, which is supported by a study of Ohlson in 1980 (Ohlson, 1980, p.120).

For testing the accuracies of the models, the intercept of the first model will be used. It is highly significant and results in a probability threshold of 10.62 percent[42], which seems appropriate based on Figure 6.4.1. Using this threshold, any observation in the validation set that receives a probability of default higher than 10.62 percent will be classified as default. Table 6.4.3 below shows the prediction accuracies of the models, using 10.62 percent as threshold[43].

[42] $\frac{e^{-2.13}}{1- e^{-2.13}} = 0.1062.$

[43] Appendix G and H displays the model accuracies for a 15 percent and a 20 percent threshold.

	With lender input	Without lender input
Correct NO	1 633/3 214	1 494/3 214
Correct YES	361/471	365/471
False NO	110/471	106/471
False YES	1 581/3 214	1 720/3 214
Correctly classified	**54.11 %**	**50.45 %**
Captured default	**76.65 %**	**77.49 %**

Table 6.4.3: Prediction accuracy of the models

In terms of observations correctly classified, none of the models achieve above 55 percent. This is mostly due to the low threshold, making the models falsely classify almost half of the non-defaulting observations as defaulting. On the other hand, the purpose of the models is to reduce defaults, as the benefit of identifying defaulting loans are larger than falsely predicting default. This means there should be a greater emphasis on the proportion of defaulting observations correctly predicted. Both of the models capture more than 76 percent of the defaulting loans at this threshold, suggesting they can reduce the number of defaults significantly. As expected, the model with inputs from the lender correctly predicts a higher percentage of the observations, by almost four percent, but the second model manages to capture slightly more of the defaulted loans.

6.5 Unobserved variables

This subchapter will present some unobserved variables that might have affected default. The contact at the bank confirmed that they are using additional variables to measure risk than those provided in the dataset, which were omitted for competitive reasons and to ensure borrowers' anonymity.

6.5.1 Payment history

Some of the loan observations includes returning loan takers, for which the bank has information about previous loan payments. If a borrower has proven able to repay a previous loan, the bank will likely assess the risk as lower and provide better loan conditions. This is common for most banks, and enables them to build stronger relationships with good customers.

6.5.2 Location

In order to ensure anonymity, the bank provided no data on the locations of their customers. Since some locations are associated with different demographic groups, for example higher income, the location of the customer might affect the risk of default (Aftenposten, 2018).

6.5.3 Financial stability

If the loan taker holds a depository account with the bank, the bank can use the balance history of the deposit account to analyse the financial stability of the loan taker. An account with a sustained, high balance could imply that the loan taker has a financial buffer. An account with an increasing balance could indicate an increase in income or reduction in expenditures.

7. Conclusion

In conclusion, lenders can reduce defaults on consumer loans by using predictive models based on the characteristics of the borrower and the loan. We have presented two such models, one with and one without inputs from the lender, and shown how a chosen probability threshold for default classifies observations as defaulting or non-defaulting. The lender can set the threshold, preferably between 10 and 20 percent according to our results, and either reject or reduce the cost of all loans that receive a higher probability of default than the threshold.

By implementing our method, lenders can reduce defaults and the associated costs. As described by Kai-Morten Terning, leader for communication and public relations in Norway's largest provider of consumer loans, Bank Norwegian: "We do not wish our customers to experience economic distress. It is not good for a bank, because we lose money, it is not good for the individual, and it is not good for the society either." (NRK, 2018). Defaults impose heavy economic burdens on individual borrowers, while the lost loan amounts reduces the profitability of the lenders. Furthermore, if the default rate for Norwegian consumer loans continues to rise, the consequences will reach a macroeconomic scale, as banks and borrowers alike fall under.

7.1 Further research

For further research, it would be interesting to implement the presented method for models on continuous risk assessment of consumer loans, using time-series data instead of cross-sectional data. This would include collecting and analysing variables relating to the down payments on the loan.

Another research could be related to the national debt registry, which is currently being implemented in Norway and is scheduled to be ready in 2019 (Trumpy, 2018). One could improve the models presented in this thesis by taking advantage of the more transparent debt information in the registry, and perhaps additional variables.

Bibliography

Andersen-Ulven, P.T. (2018, November 16). Bank Norwegian AS. Phone Interview.

Bank Norwegian AS. (2017). Annual Report 2017: Bank Norwegian AS. Retrieved 25.10.18 from https://www.banknorwegian.no/OmOss/InvestorRelations/GetDocument?id=4a91d674-08de-4536-ab29-b813117d3657

Bankenes Sikringsfond. (2016). Medlemskap og avgift. Retrieved 25.10.18 from https://www.bankenessikringsfond.no/medlemskap-og-avgift/category893.html

Biørn, E. (2013). ECON3150/4150 Introductory Econometrics – Lecture no. 1. University of Oslo. Retrieved 18.10.18 from https://www.uio.no/studier/emner/sv/oekonomi/ECON4150/v13/undervisningsmateriale/econ3150_v12_note01.pdf

Branch, B. (2002). The costs of bankruptcy – a review. *International Review of Financial Analysis* p.37-57. Retrieved 15.11.18 from https://www.sciencedirect.com/science/article/pii/S1057521901000680

Brønnøysundsregistrene. (2017, September 21). Retrieved 29.10.18 from https://www.brreg.no/om-oss/oppgavene-vare/alle-registrene-vare/om-losoreregisteret/

Bye, K.S. (2018) Mastergrad gir høyere lønn. *Statistisk sentralbyrå.* Retrieved 29.11.18 from https://www.ssb.no/arbeid-og-lonn/artikler-og-publikasjoner/mastergrad-gir-hoyere-lonn

Byrnes, J.P., Miller, D.C. & Schafer, W.D.. (1999). Gender Differences in Risk Taking: A Meta-Analysis. *Psychological Bulitin, Volume 125, Nr. 3*, p. 367-383, Retrieved: 17.12.18 from https://www.researchgate.net/profile/James_Byrnes2/publication/232541633_Gender_Differences_in_Risk_Taking_A_Meta-Analysis/links/00b49514c47ab0f093000000/Gender-Differences-in-Risk-Taking-A-Meta-Analysis.pdf

Cambridge University Press. (2018). Meaning of «default» in the English Dictionary. Retrieved 10.12.18 from https://dictionary.cambridge.org/dictionary/english/default

Cawley,. G.C. & Talbot, N.L.C. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research*. p.2079-2107. Retrieved 19.10.18 http://www.jmlr.org/papers/volume11/cawley10a/cawley10a.pdf

Christensen, J. (2018, September 27). Ferske SSB-tall: Norske husholdninger har 3.393 milliarder kroner i gjeld. *Dagens Næringsliv*. Retrieved 14.12.18 from https://www.dn.no/makrookonomi/kreditt/kredittvekst/gjeld/ferske-ssb-tall-norske-husholdninger-har-3393-milliarder-kroner-i-gjeld/2-1-430542

Christiansen, H. (2018, March 22). Her bor noen av Norges dårligste betalere. *Aftenposten*. Retrieved 08.12.18 from https://www.aftenposten.no/familieogoppvekst/Her-bor-noen-av-Norges-darligste-betalere-11772b.html

De Giorgi, E. G., Hens, T., & Bachmann, K. (2018). *Behavioural finance for private banking: from the art of advice to the science of advice* (2nd ed.). Wiley.

DnB (2018) *Consumer loans*. Retrieved 06.12.18 from: https://www.dnb.no/en/personal/loans/consumer-loans.html

Dobbin, K.K. & Simon, R.M. (2011). Optimally splitting cases for training and testing high dimensional classifiers. *BMC Medical Genomics, 4:31*. Retrieved 15.10.18 from https://bmcmedgenomics.biomedcentral.com/articles/10.1186/1755-8794-4-31

Draper, N.R. & Smith, H. (1998) *Applied Regression analysis* (3rd edition). Wiley.

Ekeseth, F.C. (2018, October 9). Nordea selger milliardportefølje med misligholdte lån. *Dagens Næringsliv*. Retrieved 7.12.18 from https://www.dn.no/marked/bank/nordea/finans/nordea-selger-milliardportefolje-med-misligholdte-lan/2-1-438678

Finansdepartementet. (2016). Forskrift om krav til nye utlån med pant i bolig (boliglånsforskriften). Retrieved 08.12.18 from https://www.regjeringen.no/no/dokumenter/forskrift-om-krav-til-nye-utlan-med-panti-bolig-boliglansforskriften/id2523973/

Finansforetaksloven. (2015). Lov om finansforetak og finanskonsern (LOV-2015-04-10-17). Retrieved from https://lovdata.no/dokument/NL/lov/2015-04-10-17. §19-1 – 19-4.

Finansportalen. (2009, November 30). Verdt å vite om inkasso og betalingsanmerkninger. Retrieved 07.12.18 from https://www.finansportalen.no/andre-valg/artikler/verdt-a-vite-om-inkasso-og-betalingsanmerkninger/

Finanstilsynet. (2010). Rapport for finansinstitusjoner 1.-4. kvartal 2010. Retrieved 27.11.18 from https://www.finanstilsynet.no/contentassets/94357d68a476416d83f6cbe7786b45b4/2010_4_kvartal_rapport_for_finansinstitusjoner.pdf

Finanstilsynet. (2014). Resultatrapport for finansforetak 2014.Resultat Retrieved 18.10.18 from https://www.finanstilsynet.no/contentassets/f80b92ae926b4a19982a93a1907f1b5d/2014_1_4_kvartal_resultatrapport_for_finansinstitusjoner.pdf

Finanstilsynet. (2016). Resultatrapport for finansforetak 2016. Retrieved 18.10.18 from https://www.finanstilsynet.no/contentassets/3a01ecd5c9784a20b3b5de0811001160/rapport-4q-2016-offentlig_versjon.pdf

Finanstilsynet. (2017a) Pressemelding: Retningslinjer for forbrukslån. Retrieved 29.11.18 from https://www.finanstilsynet.no/nyhetsarkiv/pressemeldinger/2017/retningslinjer-for-forbrukslan/

Finanstilsynet. (2017b). Rundskriv: Finanstilsynets retningslinjer for forsvarlig utlånspraksis for forbrukslån. Retrieved 05.11.18 from https://www.finanstilsynet.no/contentassets/324f40afaaee445d8b4950ad4b930dac/rundskriv-5-2017-retningslinjer-for-forbrukslan.pdf

Finanstilsynet. (2017c). Beregningsgrunnlaget. Retrieved 13.12.18 from https://www.finanstilsynet.no/tema/kapitaldekning/beregningsgrunnlaget/

Finanstilsynet. (2017d). Resultatrapport for finansforetak 2017. Retrieved 18.10.18 from https://www.finanstilsynet.no/contentassets/de14e53407744ba1ba315efbb0c67c3c/resultatrapport-for-finansforetak.pdf

Finanstilsynet. (2018a). *Resultatrapport for finansføretak.* Retrieved November 2, 2018, from https://www.finanstilsynet.no/publikasjoner-og-analyser/resultatrapport-for-finansforetak/#22547

Finanstilsynet. (2018b) Finansielt utsyn – juni 2018. Retrieved 09.12.18 from https://www.finanstilsynet.no/contentassets/de1768219a1c47f88db65a096c8ab862/finansielt-utsyn-juni-2018-16.21.pdf

Finanstilsynet. (2018c). Finansielt utsyn – desember 2018: Kapittel 3 - figurer og bakgrunnsdata: Bankene (xlsx). Retrieved 9.12.18 from https://www.finanstilsynet.no/publikasjoner-og-analyser/finansielt-utsyn/#24805

Finanstilsynet. (2018d). Finanstilsynet foreslår forskriftsregulering av forbrukslån. Retrieved 18.11.18 from https://www.finanstilsynet.no/nyhetsarkiv/pressemeldinger/2018/finanstilsynet-foreslar-forskriftsregulering-av-forbrukslan/

Finanstilsynet. (2018e) Finansielt utsyn – desember 2018. Retrieved 09.12.18 from https://www.finanstilsynet.no/contentassets/a023382a3e38475ebfd503a1c24fb016/finansielt-utsyn-desember-20182.pdf

Forbrukerrådet. (2018). Inkasso. Retrieved 18.10.18 from https://www.forbrukerradet.no/forside/okonomi-og-betaling/inkasso-2/

Foreldelsesloven. (1979). Lov om foreldelse av fordringer. (LOV-1979-05-18-18). Retrieved 21.10.18 from https://lovdata.no/dokument/NL/lov/1979-05-18-18?q=fire%20%C3%A5r

Garside, M.J. (1965). The Best Sub-set in Multiple Regression Analysis. *Journal of the Royal Statistical Society – series C*. p.196-200. Retrieved 02.11.18 from https://www.jstor.org/stable/2985341?seq=1#metadata_info_tab_contents

Goodhart, C.A.E. (2008) The background of the 2007 financial crisis. Springer-Verlag 2007. Retrieved 25.11.18 from https://link.springer.com/content/pdf/10.1007/s10368-007-0098-0.pdf

Hagen, M., Turtveit, L.T., Vatne, B.H. (2017, January 30). Sterk vekst i forbrukslån. "Aktuell kommentar" - 2017. *Norges Bank*. Retrieved 16.10.18 from https://static.norges-bank.no/contentassets/ebe82c99213e4e8592a0415cc387cb45/aktuell_kommentar_1_2017.pdf?v=03/09/2017123521&ft=.pdf

Hair, J.F., Black, W.C., Babin, B.J & Anderson, R.E. (2014). Multivariate Data Analysis (7th edition). Pearson.

Hansen, A., og Lillebø, P. (2016). *Det norske forbrukslånsmarkedet: empiriske analyser av forbrukslånkundenes finansielle soliditet og drivere for veksten i forbrukslånmarkedet*. Norges Handelshøyskole

Hosmer, D.W., Lemeshow, S. & Sturdivant, R.X. (2013) Applied Logistic Regression (3rd edition). Wiley.

International Accounting Standards Board. (2014). IFRS 9 Financial instruments. In *International financial reporting standards (2018)***.** Retrieved 16.10.18 from http://eifrs.ifrs.org/eifrs/PdfAlone?id=19282&sidebarOption=UnaccompaniedIfrs

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2017). *An Introduction to Statistical Learning with applications in R*. Springer.

Kapitalforskriften. (2014). Forskrift om kapitalkrav for banker, kredittforetak, finansieringsforetak, holdingforetak i finanskonsern, verdipapirforetak og forvaltningsselskaper for verdipapirfond mv. (FOR-2006-12-14-1506). Retrieved 19.12.18 from https://lovdata.no/dokument/SF/forskrift/2006-12-14-1506

Kisgen, D.J. (2006) Credit ratings and Capital Structure. *The Journal of Finance, vol 61.* p.1035-1072. Retrieved 19.11.18 from https://onlinelibrary.wiley.com/doi/full/10.1111/j.1540-6261.2006.00866.x

Komplett Bank AS. (2018). Leter du etter et forbrukslån på timen?. Retrieved 18.12.18 from https://www.komplettbank.no/om-komplett-bank/

Lånehjelpen. (2018). Oversikt over billig forbrukslån I Norge. *Nettavisen.* Retrieved 01.11.18 from https://www.nettavisen.no/forbruksl%C3%A5n/

Lavik, R. & Borgeraas, E. (2015). Forbrukstrender Økonomi 2015. *Statens institutt for forbruksforskning.* Retrieved 17.10.18 from http://www.hioa.no/extension/hioa/design/hioa/images/sifo/files/file80276_sifo_forbrukstrender_economy_2015.pdf

Lewis-Beck, M.S. (1980) Applied regression: an introduction (2nd edition). SAGE Publications.

Lindquist, K.G., Solheim, H. & Vatne, B.H. (2017). Husholdningenes gjeld og koplinger til boligmarkedet. Konsekvenser or finansiell stabilitet. “Aktuell kommentar” – 2017. *Norges Bank* Retrieved 17.10.18 from https://static.norges-bank.no/contentassets/cf6c78a9b36c491b9008a6e648cf9b89/aktuell_kommentar_7_2017.pdf?v=12/06/2017153845&ft=.pdf

Meier, S., & Sprenger, C. (2010). Present-Biased Preferences and Credit Card Borrowing. American Economic Journal: Applied Economics, 2(1), 193-210. Retrieved 15.12.18 from http://www.jstor.org/stable/25760198

Mikalsen, K.E. (2016, November 1). Tydelig trendskifte: Nå kjøper stadig flere nordmenn melk, bøker, medisiner og parkering på kreditt. *Bergens Tidende*. 15.12.18 from https://www.bt.no/nyheter/okonomi/i/GAwA9/Tydelig-trendskifte-Na-kjoper-stadig-flere-nordmenn-melk_-boker_-medisiner-og-parkering-pa-kreditt

Nordea Bank. (2018a). Nominell og effektiv rente - hva er forskjellen. Retrieved 09.12.18 from https://www.nordea.no/privat/vare-produkter/lan-og-kreditt/nominell-og-effektiv-rente.html

Nordea Bank. (2018b). Refinansiere lån – samle lån og spar penger. Retrieved 19.12.18 from https://www.nordea.no/privat/vare-produkter/lan-og-kreditt/refinansiering-av-gjeld.html

NRK. (2018, October 31). «31 Oktober - Dagsnytt atten», video, (25:20)

Null TDN Finans. (2015, September 29). DNB selger misligholdte lån til Lindorff. *Dagens Næringsliv*. Retrieved 07.12.18 from https://www.dn.no/dnb-dnb/dnb/dnb-selger-misligholdte-lan-til-lindorff/1-1-5471473

Ohlson, J.A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting research.* p.109-131. Retrieved 13.11.18 from https://www.jstor.org/stable/pdf/2490395.pdf

Regjeringen. (2017). Forskrift om markedsføring av kreditt. Retrieved 17.12.18 from https://www.regjeringen.no/no/dokumenter/forskrift-om-markedsforing-av-kreditt/id2548392/

Regjeringen. (2018). Forsinkelsesrente. Retrieved 07.12.18 from https://www.regjeringen.no/no/tema/okonomi-og-budsjett/renter/forsinkelsesrente-/id429404/

SAS. (2018). Machine Learning. Retrieved 20.12.18 from https://www.sas.com/en_us/insights/analytics/machine-learning.html

Shao, J. (1993) Linear model selection by Cross-Validation. *Journal of the American Statistical Association, vol.88, No.422.* p 486-494. Retrieved 12.11.18 from https://www.jstor.org/stable/2290328?seq=1#metadata_info_tab_contents

Skatteetaten. (2018). 3.3.1 Gjeldsrenter. Retrieved 1011.18 from https://www.skatteetaten.no/person/skatt/skattemelding/finn-post/3/3/1/

SIFO. (2017). Forbrukstrender: SIFO survey 2005-2017. *Oslo: Forbruksforskningsinstituttet SIFO*. Retrieved 10.11.18 from http://www.hioa.no/Om-OsloMet/Senter-for-velferds-og-arbeidslivsforskning/SIFO/Publikasjoner-fra-SIFO/Forbrukstrender-SIFO-survey-2005-2017

Singstad, J. (2018, November 2). Finanstilsynet. Email Communication.

Statistisk sentralbyrå. (2018a). Arbeid. Retrieved 18.10.18 from https://www.ssb.no/arbeid-og-lonn/faktaside/arbeid

Statistisk sentralbyrå. (2018b). Statisikk: Lønn. Retrieved 11.11.18 from https://www.ssb.no/arbeid-og-lonn/statistikker/lonnansatt/aar

Stiglitz, J.E. and Weiss, A. (1981) Credit Rationing in Markets with Imperfect information. *The American Economic Review Vol. 71, No. 3.* p. 393-410. Retrieved 10.11.18 from https://www.jstor.org/stable/1802787?seq=1#metadata_info_tab_contents

Trumpy, J. (2018, November 18). Evry står klar med gjeldsregister på forbrukslån – men venter på ett hinder for å trykke på start. *Dagens Næringsliv*. Retrieved 20.12.18 from https://www.dn.no/marked/forbrukslan/datatilsynet/evry/evry-star-klar-med-gjeldsregister-pa-forbrukslan-men-venter-pa-ett-hinder-for-a-trykke-pa-start/2-1-478845

Trumpy, J. & Christensen, J. (2018, October 30). Bank Norwegian lemper over misligholdte lån på John Fredriksen-selskap. *Dagens Næringsliv.* Retrieved 07.12.18, from https://www.dn.no/marked/bank-norwegian/finansdepartementet/norwegian-finans-holding/bank-norwegian-lemper-over-misligholdte-lan-pa-john-fredriksen-selskap/2-1-462977

Tvangsfullbyrdelsesloven. (1992). Lov om tvangsfullbyrdelse. (LOV-1992-06-26-86). Retrieved 21.10.18 from https://lovdata.no/dokument/NL/lov/1992-06-26-86?q=tvangsfull

Vedeler, M. (2016, December 3) Slik fikk E24 to mill. I forbrukslån på fire dager. *E24.* Retrieved 23.10.18 from https://e24.no/boers-og-finans/bank/slik-fikk-e24-to-mill-i-forbrukslaan-paa-fire-dager/23858253

VG. (2017, March 27). Forbrukslån på ville veier. *VG.* Retrieved 17.10.18 from https://www.vg.no/nyheter/meninger/i/5Vw0rX/forbrukslaan-paa-ville-veier

Will Kenton. (2018, December 13). Return On Average Assets - ROAA. Retrieved 01.11.18 from https://www.investopedia.com/terms/r/roaa.asp

Woolridge, J.M. (2012) Introductory Econometrics – A modern approach (5th edition). Cengage Learning.

yA Bank. (2018). Ofte stilte spørsmål - Hvor lang tid tar det før lånet utbetales? Retrieved 15.12.18 from: https://ya.no/forbrukslan#humany-generell-lightbox=/47/hvor-lang-tid-tar-det-for-lanet-utbetales/

Appendices

Appendix A: Definitions

Borrower – An individual or institution that borrows something, for instance money from a bank (Cambridge dictionary, 2018). A person granted a consumer loan is a borrower for that loan.

Ceteris paribus – All other things held equal. A phrase used to underline that an argument only holds when all other conditions remain constant.

Consumer bank – A bank that focuses on issuing unsecured debt.

Consumer loan – An unsecured loan provided by a lender. Includes one-time loans and credit card loans.

Correlation - A measure of the linear relationship between two sets of values (Woolridge, 2012, p.739). A high correlation coefficient might indicate a causal relationship between two variables.

Collateral – Something a borrower put up as security for a loan, which the lender receives if the loan defaults. Consumer loans do not include collateral.

Default – Failing to do something that you are legally required to do (Cambridge University Press, 2018). When the borrower has failed to repay a debt payment, such as an interest charge, 90 days after it was due, the loan is considered defaulted (IASB, 2014) and unlikely to be repaid in full.

Default rate – The proportion of defaults to the total amount of defaults and non-defaults.

Delinquency – A situation where a borrower is late or overdue on a payment (Investopedia, 2018). Extended loan delinquency leads to loan default.

Effective interest rate – The total cost of a loan, as a percentage of the loan amount. The effective interest rate is a combination of the nominal interest rate and all additional fees (Nordea, 2018).

IFRS – The abbreviation for International Financial Reporting Standards. All listed firms within the EU and the EEA are required to comply with the most recently revised standards of the IFRS.

Lender – Someone that lends something to a borrower, with an agreement of receiving it back. A bank that provides consumer loans is a lender for those loans.

Machine learning – A method of data analysis that automates analytical model building, using a programming language such as R (SAS, 2018).

Nominal interest rate – The interest rate paid on a regular basis on the loan, excluding any additional fees. Often denoted in annual terms, although payments are normally paid monthly.

Principal amount – The borrowed loan amount. Monthly payments on a loan is the monthly interest rate times the remaining, unpaid principal amount.

Secured debt – Debt backed by a collateral that the lender receives if the borrower fails to repay the loan. Consumer loans is an example of an un-secured debt, which holds greater risk than secured debt since the lender might be left with nothing if the borrower fails.

Appendix B: Cleaning the data

Step 1: Excluding recent observations

Since default is not recognized until a payment is 90 days overdue, all loans provided within 104 days[44] of the date the dataset was collected has been removed.

Step 2: Removing insignificant variables

Some of the variables in the data were not relevant for the thesis questions, and was therefore removed. One of these was agent ID, which is an identifier for which agent approved the loan. While useful for the lender in order to evaluate the agents, it is not useful for modelling or statistical purposes. Another insignificant variable was the date of the loans. After having removed the most recent loans, the date provides no additional value to the analysis. Finally, only five observations had a yearly bonus, so that variable was also removed along with the five observations.

[44] First payment is due 14 days after the loan is approved. No loan can default before 14+90=104 days has passed.

Although the loan amount the borrower originally applied for is interesting for the analysis, the variable was converted to a new variable called Difference from Applied for two reasons. The first is that it is more interpretable to look at the actual difference between what the borrower applied for and what they received. The second reason is that the correlation coefficient between Loan Amount and Applied Amount was 0.9, meaning the variables were strongly correlated and both should not be present during analysis (Lewis-Beck, p.60, 1980). Applied amount was therefore replaced with Difference from Applied, calculated by subtracting the values in Loan Amount with the values in Originally Applied.

Step 3: Dealing with missing data

Another common stepping-stone when cleaning data is dealing with missing values. This could be a result of improper storing of data or lacking answer by respondents (Hair Jr. et al, p.32, 2014). Missing values are dealt with in one of three ways: changing the value, removing the observation or removing the variable.

In the dataset from the bank, 3 753 of the remaining observations had a missing value for Education. Since Education is likely to be significant for the analysis, the observations with missing data were removed. This also applied to 104 observations that had missing values in either Employment, Living Arrangement or Civil Status, 94 observations with missing values for wealth, and seven observations with missing values in Applied amount.

Some of the numeric variables also contained missing values that the contact at the bank ensured corresponded to zero, including Children under 18, Rent Income and Other Income. These values were converted to zero.

Step 4: Standardizing income and expenses

Income and expenses also needed to be standardised. While Gross Income was on a yearly basis, Rent Income, Other Income and Rent Expenses were on a monthly basis. All of the monthly variables were annualised to make the data more interpretable, by multiplying their values with 12.

Step 5: Removing outliers

The final step of the cleaning process was removing outliers. Outliers are observations that are substantially different from the majority of the observations for one or more of the variables. Because of their extreme values, they might distort statistical testing and make the sample less representative of the population (Hair Jr. et al, p.63, 2014). Univariate detection[45] and a Z-value[46] of 3.0 was used to remove outliers from the dataset.

Appendix C: Table of the cleaning process

	Original	*Step 1*	*Step 2*	*Step 3*	*Step 4*	*Step 5*	**Final**
Observations	17 784	16 190	16 185	12 227	12 227	10 836	10 836
Variables	29	29	26	26	26	26	26

[45] Examining the distribution of the observations for each variable and removing observations that fall outside the given range (Hair Jr. et al, p.64, 2014).

[46] Assuming a normal distribution, excluding outliers outside of 3.0 Z-range will omit 0.26% of the observations, or one in every 385 observation. (Woolridge, p.831-832, 2012).

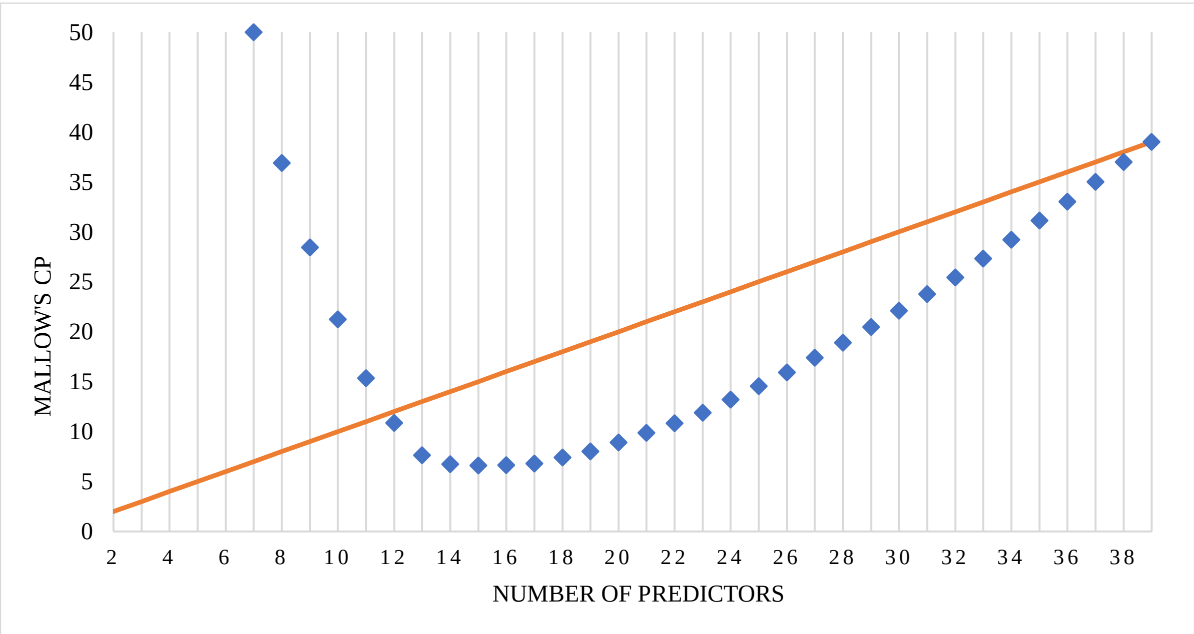
Appendix D: Correlation Matrix

	Defau~90	LoanAm~t	Mortgage	OtherL~s	Studen~n	Origin~t	Princi~s
Default90	1.0000						
LoanAmount	-0.1181*	1.0000					
Mortgage	-0.0813*	0.1647*	1.0000				
OtherLoans	-0.0710*	0.3323*	0.1251*	1.0000			
StudentLoan	-0.0174	-0.0150	0.0501*	-0.0007	1.0000		
Originally~t	-0.0973*	0.8942*	0.1200*	0.3401*	-0.0233	1.0000	
Principalf~s	0.0496*	0.1403*	0.0420*	-0.0462*	0.0039	0.1535*	1.0000
InterestRate	0.1682*	-0.4377*	-0.2728*	-0.1324*	0.0370*	-0.3634*	-0.0032
Co_signer	-0.0924*	0.3286*	0.0039	-0.0539*	-0.0211	0.2963*	0.0027
Age	-0.1156*	0.0162	0.0832*	0.1607*	-0.2501*	0.0114	-0.0393*
Man	0.0336*	0.0843*	0.0810*	0.0268*	-0.0878*	0.0708*	0.0025
Master	-0.0578*	0.2419*	0.1493*	0.0796*	0.1800*	0.2216*	0.1161*
Bachelor	-0.0039	0.0725*	0.0464*	-0.0138	0.1712*	0.0586*	0.0371*
PrivateOwner	-0.0857*	0.0669*	0.5915*	0.1024*	-0.0060	0.0325*	0.0246

	Intere~e	Co_sig~r	Age	Man	Master	Bachelor	Privat~r
InterestRate	1.0000						
Co_signer	-0.2033*	1.0000					
Age	-0.2235*	-0.0111	1.0000				
Man	0.0083	-0.0222	-0.0945*	1.0000			
Master	-0.1623*	0.0361*	0.0359*	-0.0382*	1.0000		
Bachelor	-0.0153	0.0114	-0.0493*	-0.0440*	-0.2859*	1.0000	
PrivateOwner	-0.2819*	0.0326*	0.1901*	0.0091	0.0425*	0.0366*	1.0000

*Significance level: * p<0.01*

Appendix E: Mallow's Cp for model without lender input



Appendix F: Coefficients and VIFs of two models w/ lender inputs

	With Loan Amount		Without Loan Amount	
Predictors	Coefficient	VIF	Coefficient	VIF
Loan Amount	-1.05E-06***	1.50		
	(-2.67E-07)			
Mortgage	-9.78E-08*	1.38	-7.903E-0.8*	1.36
	(4.10E-08)		(4.04E-08)	
Student Loan	-1.78E-06***	1.08	-1.69E-06***	1.08
	(4.82E-07)		(4.79E-07)	
Principal free Months	0.014***	1.04	0.013***	1.01
	(2.37E-03)		(2.33E-03)	
Interest Rate	0.112***	1.37	0.135***	1.14
	(0.014)		(0.013)	
Co-signer	-0.804***	1.15	-0.946***	1.08
	(0.150)		(0.145)	
Age	-0.023***	1.19	-0.022***	1.18
	(3.13E-03)		(3.12E-03)	
Married	-0.395***	1.41	-0.396***	1.39
	(0.083)		(0.083)	
Children_U18	0.143***	1.26	0.159***	1.25
	(0.040)		(0.039)	
Gross Income	-5.33E-07**	1.57	-8.05E-07***	1.42
	(2.18E-07)		(2.09E-07)	
Man	0.217***	1.10	0.208***	1.10
	(0.073)		(0.073)	
Domestic Partner	-0.183**	1.21	-0.183**	1.21
	(0.081)		(0.081)	
(Intercept)	-2.129***		-2.595***	
	(0.303)		(0.288)	

*Significance levels: *** p<0.01 ** p<0.05 * p<0.1*

Standard Errors in pharanteses

Appendix G: Model accuracies using 15 percent threshold

	With lender input	Without lender input
Correct NO	2 243/3 214	2 224/3 214
Correct YES	278/471	260/471
False NO	194/471	211/471
False YES	970/3 214	990/3 214
Correctly classified	**68.41 %**	**64.27 %**
Captured default	**59.02 %**	**55.20 %**

Appendix H: Model accuracies using 20 percent threshold

	With lender input	Without lender input
Correct NO	2 716/3 214	2 759/3 214
Correct YES	167/471	149/471
False NO	304/471	322/471
False YES	498/3 214	455/3 214
Correctly classified	**78.23 %**	**78.91 %**
Captured default	**35.46 %**	**31.63 %**