Using bank transactions to assess credit risk

A quantitative study on how bank statement information can be used to predict delinquency in consumer loans

Authors

Hugo Korsell (50432) & Oscar Samuelsson (50442)

Stockholm School of Economics

Bachelor thesis - Retail Management

Submission

2019-05-16

Supervisor

Mariya Ivanova

Examinator

Johan Graaf

ABSTRACT

A lack of proper credit risk assessment can have catastrophic consequences, not only for individual borrowers or lending banks, but for society as a whole, which was the case in the financial crisis of 2008. Therefore it is essential that any risk assessment process has as much information as possible at hand so that the delinquency predictions are as accurate as they can be. This study uses real loan data to investigate whether incorporating information about gambling, making collection payments, and being dishonest in reporting income in risk assessment models improves their power to predict delinquency of loan payments. Prior literature has considered both hard and soft information in credit risk assessment, and this study contributes to that by using a unique dataset which provides the opportunity to test whether the behavioural patterns mentioned above, which have not previously been used in credit risk assessment, is associated with delinquency. The results shows that there is no significant connection between the researched factors and delinquency of loans, but that borrowers with a history of gambling or collection payments still tend to have a higher perceived risk which is reflected in the interest rate of their loans.

Keywords: Credit risk, bank statement, private loans, gambling, dishonesty

1. Introduction

1.1. Background

The loan market is a central function in today's society that enables people to use capital without having to save up for it. Loans can be used for shifting the time of consumption, buying property or investment in businesses, making the society function more efficiently. For lenders, there is always a risk of borrowers not repaying the loan, which means credit risk assessment is an essential part of the process to make sure the borrowers are qualified for the loan according to the risk appetite of the lender. It also plays a big part in the global economics in a wider perspective. It can easily be argued that the financial crisis of 2008 was a result of poor credit risk assessment in the US housing market, leading up to a bad spiral with decreasing house prices and increasing unemployment rates, eventually causing the global financial crisis to start off (The Balance (2019), The Economist (2019)).

Banks have been a big part of society for hundreds of years, but their methods of risk assessment have not always been structured and systematic. Before the 20th century, banks generally had a standardized interest rate offered to all customers regardless of their creditworthiness. A basic risk assessment was made when an applicant applied for a loan. If the risk level was deemed appropriate, the loan was granted, and if not it was denied. In recent years, most banks have adapted risk-based pricing models using information about the lenders in order to determine their risk levels and set the interest rate accordingly, so that the lending party can make sure the credit risk levels are appropriate and sustainable, both for the lending company and for society as a whole (Edelberg, 2006). The use of risk-based credit models has been possible thanks to digital improvement and possibility to store and access data to a greater extent. The effect is that the interest rate today is more accurate and reflective of the actual risk of granting the loan.

1.1.1. The private loan market in Sweden

The market of private loans, also called blanco loans, unsecured loans or personal loans, is growing rapidly in Sweden. In December 2018, the total amount of lent out blanco loans in

Sweden reached 253,1 billion SEK and had an annual growth rate of 8 % (Finansmarknadsstatistik december 2018, Statistiska centralbyrån). These loans are characterized by the fact that they are loans without security, which means the borrowing party can use the loan for whatever purpose they require without the interference of the lending party, which is often a bank. Secured loans are different in the way that the lending party can claim the security, typically a house or a car, should the borrower not be able to repay the loan. Not having security indicates a higher risk for the lending party, which generally would result in a higher interest rate.

In general, the average interest rates of blanco loans in Sweden depend on two main factors - competitive pressure and the Riksbanken prime rate. Competitive pressure has increased considerably in the past decade due to the prominence of loan comparison services. They are companies that work as an intermediary between consumers and banks - offering a comparison where a customer can view interest rates and terms from many banks at the same time, only using one credit check. In theory, this creates a more efficient market which results in greater competition between banks since the information gap where customers only know what offer they may get from one bank decreases, which thereby lowers interest rates for customers due as a result of the increased availability of information. The revenue of the comparison agencies are mainly made up by commission from the banks, who pay the comparison companies when they provide a customer to the bank (Direkto, 2019). As of 2018 around 13 % of all private loans are provided through a loan comparison provider, and the market for loan comparison services is still growing rapidly (Kvalitetsindex, 2019).

Figure 1 below shows the development of the interest rates in the market for unsecured private loans in Sweden from 2006 to 2019, with official data from Statistiska Centralbyrån, the government authority of statistics with respect to the general trend in the interest rates, which is shown in the dotted line. It shows that the interest rates for this type of loans are declining. Although it is not clear what is behind this trend, it could be in line with classical economics theory that competition lowers prices. The market of loan comparison services has also increased dramatically during this period, with immense growth in the private loan sector in general. With the private loan market growth, the interest rates might continue falling, in favor of customers. However, the interest rate decline is in line also with the decline in the

prime interest rates in Sweden that has also shown a major decline since 2012, which indicates that the interest rate levels of unsecured blanco loans are following the prime rate rather than level of competition (Riksbanken, 2019).

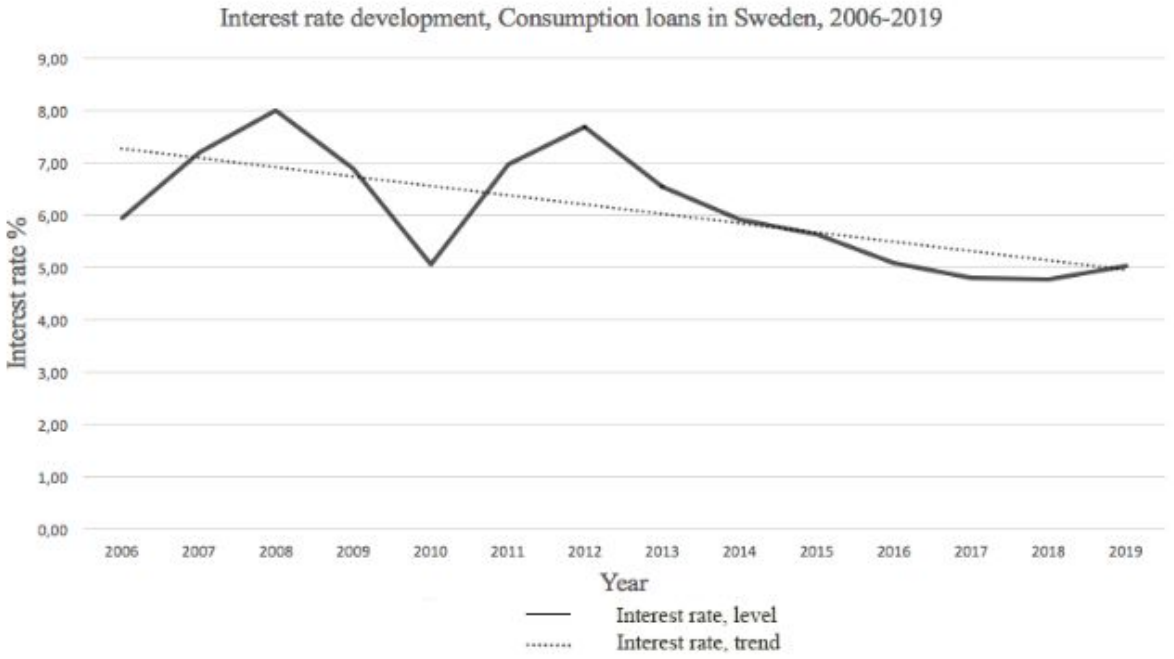

Figure 1, Interest rate development, Consumption loans in Sweden, 2006-2019

Since the product any lender is selling, the loan, or more specifically the ability to use money in advance for a cost, is generic, the price is a very important tool to differentiate from competitors. The offered products are practically indistinguishable between competitors, and although Swedish customers historically have been loyal to their bank, the loyalty is declining - with customers switching or adding banks more frequently than before (Dagens Industri, 2019). Since the entry of private loan comparison websites, the information asymmetry of the market has decreased and it has become more clear for the borrower where s/he will get the lowest interest rate, which means that the lending companies must remain competitive with respect to price to attract customers. This is different from other kinds of markets where a company can benefit from a differentiated product to attract customers. While brand strength and service reasonably affects the customer to some extent, it is reasonable that a lower interest rate than the competitor makes a lender more attractive since it is the main feature of the offered product. This means that loan comparison companies reasonably put a pressure on the banks to lower the interest rates offered to the end customer.

This will in turn pressure the banks to improve their risk assessment, in order to cut costs related to delinquency and loan defaults so that they can still make a positive result despite the decreased interest rate earnings. With an improved risk assessment, the banks can lower their costs and therefore their prices without reducing their margins. If a bank is better at risk assessment than a competitor they can give low-risk customers a lower interest rate and high-risk customers can be correctly priced or avoided altogether.

1.1.2. Implications for the pricing of loans

A fundamental assumption in finance is that an interest rate is meant to reflect the risk of the underlying asset, plus a profit margin or systemic risk premium added by the lender. This indicates that the interest rate of a private loan should reflect the risk of default of that specific loan, plus a potential premium added on by the lender, and to determine the risk level and thereby set an interest rate to a loan, credit models are used. In Sweden, what is most commonly used for this is Upplysningscentralen, hereafter UC, which is a company that uses information such as the applicants' income history, credit history, marital status and more and sells this information to lenders, including a credit score calculated based on those factors. This is further specified in Appendix 1. The credit score determined by UC is based on the information available to them, which mainly consists of so called hard values such as income and wealth, credit history etc. However, there is information that is not used by UC but that could still prove to be relevant to determine the creditworthiness and thereby also the risk level of a specific application. The probability of default of a loan may increase following things such as high levels of gambling, frequent payments to collection companies or loan application dishonesty and such factors should, if proven to have an effect on risk, be added on top of other credit scores by the bank when setting interest rates for customers.

1.2. Purpose of the study

The main aim of this thesis is to investigate whether incorporating information about certain behavioral patterns of borrowers such as gambling, making collection payments, and being dishonest in reporting income in risk assessment models improves their predictive power by using a unique dataset consisting of unsecured private loans issued by a lending firm in Sweden. Hereafter, to protect the identity of the firm, it will be referred to as The Bank. A

further description of The Bank and its credit process can be found in 3.2. Currently, risk classification and thereby interest rates of unsecured private loans are completely based on the Upplysningscentralen credit score (UC Score) of the applicant. The applicant is assigned an internal risk class, which is determined by intervals in UC Scores. These were created by The Bank to make pricing easier, otherwise it would be too hard to manage separate interest rate tables for every single value in the UC Score, since it has 1-2 digits and three decimals. However, a drawback of the UC credit score is that it is limited to official information from government institutions and does not consider behavioral patterns or other factors that may be identifiable through additional sources such as bank statements. Using relevant information that is not available to UC and including it in a credit score should reasonably increase the accuracy of said model. Things such as gambling problems, history with collection companies and application dishonesty can be identified and possibly used for the risk assessment of the loan through the use of transactional data.

Therefore, the purpose of this study is to investigate any potential connections between information in the bank statements and the repayment ability of the respective borrowers in order to discover whether these can be used for measuring risk and should therefore be incorporated in risk models. This will make the interest rates reflect the real risk better and thereby benefit the low-risk borrowers, since an unproblematic behavior regarding gambling, collection and income cheating is likely to prove to indicate lower risk and thereby lower interest rate. With the pricing model that is currently in place, borrowers in lower risk classes are probably collectively punished since there are high risk borrowers hidden in their risk class, since their problematic behaviors are not identified, which boosts the expected default rate and therefore also the interest rate of the specific risk class. By including more data points, these problematic behaviors will be identified and likely put in a different risk class, reducing the default rate and thereby also the expected default rate and interest rates. By the same logic, a more accurate risk assessment would increase the interest rates of the loans to borrowers with problematic behaviors, making them pay the real price relative to their risk.

Incorporating behavioral patterns in lending decisions would also allow banks to set more competitive interest rates to these ”good” customers that the other banks who are not using bank statements might consider ”bad” ones and therefore give high interest rates, which

means in the long run that banks can generate higher loan volumes since the offer is improved. The interest rates can be lowered either by decreasing the operational expenses so that the lender can add a smaller profit margin on the interest rate and still be profitable, or by improving the risk assessment thus making the pricing more efficient and therefore more accurate, making it easier for credit managers to set appropriate interest rates and therefore be more in control of the exact levels of expected credit losses. By making the risk assessment more accurate, the actual risk of miscalculating risk itself decreases.

1.3. Research question

Motivated by the background above, the general research question for this study is as follows.

"Does incorporating personal behavioral patterns improve credit risk assessment in the context of Swedish private loans?

More specifically, to address the research question this thesis examines whether gambling, payments to collection companies, and dishonesty are associated with the probability of default.

1.4. Contribution

1.4.1. Contribution to existing theory

This study will provide new insights regarding specific factors that can affect the accuracy of credit scoring models, whereas these factors have not been researched from a risk assessment perspective before. The study further provides valuable insights to the theory regarding factors affecting the ability to repay loans, with the factors being gambling, income statement dishonesty and collection agency payments. The prior theory on gambling does in part cover the effect that gambling has on a person's private economy, but not how it affects the ability to repay loans, and this research provides new insight through that perspective. It is therefore contributing to both gambling theory and credit risk theory. Furthermore, the results of this study also contribute to the existing theory on credit risk in the area of application dishonesty, that is relatively unexplored today. While this research only measures the level of truthfulness

in the income statement, the shown results opens up possibilities to explore how level of truthfulness in other parts of a loan applicant's life affects the credit risk. This research also contributes by developing a new direct measure of application dishonesty that is based on actual bank statements, as opposed to previous research which only used indirect estimations and proxies to assess application dishonesty. Finally, the research contributes to the litterature on collection payments where this research provides insights to the way collection payments can be used in a risk assessment process, where previous research has provided knowledge on how it affects other parts of a person's private economy.

1.4.2. Practical implications

A more accurate credit score calculation, which means a more accurate pricing of the loans, would arguably have implications on the interest rates of the applicants, both by decreasing and increasing the interest rates. For applicants that are today classified with a lower risk than they really are based on their behavioral patterns, the interest rate would increase since the higher hidden risk would be identified and thereby added to their interest rate. However, this would be highly useful for the lender since they are no longer approving loans with an interest rate that does not properly reflect the borrower's risk. In the same way, applicants who are classified as more risky than they actually are would be given a lower interest rate than today which is obviously beneficial for them. This is also in the lenders' interest, since the competitive landscape in the market due to the entry of loan comparison services and the nature of the product discussed above makes it essential to be able to give low interest rates to be competitive and stay in business. The revenue that is lost in interest rate, is gained by lending greater volumes. The applicants who might be worse off from having a more accurate risk assessment of private loans are the applicants that are today classified as a lower risk than they should be, since their interest rates will increase. However, one could argue that they are today getting a better price than they should be getting and that improving the accuracy will only bring their price closer to an equilibrium of sorts.

1.5. Assumptions

Several assumptions were made in this thesis. One key assumption made in this research is the one that the sample is reflecting the Swedish loan market as a whole, that the key findings

can be used to accurately assume characteristics for all private loan applicants in the country. The main issue for lending banks is presumably borrowers not paying back, but since the data does not have enough samples defaulting on the whole loan, a delinquency of at least 30 days is assumed to show the loans that have a high risk of defaulting. To add to this, the assumption is made that banks want to know more about factors affecting their customers' ability to repay their loans, and reflect the known risk in the offered interest rates. Another assumption is made for the salary analysis, where the loan applicant's stated income is compared with the actual income found in the bank statement. The assumption is that that the income has been taxed at an average Swedish tax rate, with a formula using 32.12 % tax rate with additional high income tax brackets taken into consideration. This due to the self-reported income being in gross numbers and the bank statement data being in net numbers, and to be able to identify dishonesty, the stated gross income must be converted to net income. Furthermore, we assume The Bank's classification of transaction types is completely accurate, i.e., it captures all gambling, collection and income transactions without neglecting any of them or classifying them wrongly in the bank statement data.

1.6. Limitations

A major shortcoming about only assessing the bank statement data manually, as in the credit process today, is that the risk of the applicant's behavior is not properly quantified. There are no numbers supporting what limit of e.g. gambling should be accepted and to what price. Theoretically, the purpose of any interest rate is to reflect the risk of the underlying asset, and this means for all levels risk there must also be a corresponding interest rate to compensate for the risk. While banks may identify an applicant that gambles for a large portion of their income every month as a high risk customer by manual assessment, the risk can theoretically be quantified and converted into an interest rate for that applicant that would make the risk level acceptable for a bank with a high enough risk appetite. It all comes down to the risk appetite, but there is surely a price that would reflect the actual risk – and we intend to measure that risk instead of simply disregarding the ”bad” applications without calculating the actual risk it brings.

There is substantial probability that the applicants with the most risky behaviours when it comes to gambling, collection agencies and falsely stated income have been denied a loan and therefore are not included in the data set and hence not measurable in terms of loan defaults or late payments, but since applicants with low or medium levels of those behaviours have likely not been denied loans, the default probability might still prove to be connected to it.

Another limitation is the fact that the study does not consider the applicants who have not shared their bank statement information. Out of The Bank's borrowers, not all actually share their bank statement data with The Bank. Since this is voluntary for some customers, and mandatory only in certain cases, one could assume the bank statements available for this study belong to high risk borrowers to a greater extent than to low risk customers, which indicates the average risk of the study sample might be higher than the average risk of the whole loan stock of The Bank, which could affect the results of this study. Furthermore, the loans used in this study are only loans that have been approved and that are only part of The Bank's loan stock. Since The Bank's risk appetite most likely is not identical to other lending actors' in the market, this study will not cover 1) the loans that are with other banks, and 2) the customers that are outside of The Bank's acceptable risk spectrum. Therefore the results of this study will not be able to give a complete and 100 % exact picture of the risk of the entire private loans market in sweden, but it will still be possible to give a good indication.

Another limitation induced by the bank statement data is that it only covers the transactions from the bank account where the borrower receives his or her salary. It is impossible to say whether a customer has a bank account or credit card with another bank, and therefore we can assume not all of the transactions of every borrower are captured in this study.

Furthermore, most of the loans used in this study have not yet matured. This means the study can only see whether the borrower has been delinquent so far, and cannot tell whether delinquency will occur before the loan is fully repaid. This is obviously a big issue, but it is reasonable to believe that most delinquency occurs early in the lifetime of a loan, making the study still relevant.

1.7. Delimitations

This research uses a specific data set from a bank in Sweden, and not using all loans that they have ever had. While the data set is used as a proxy for the whole loan market in Sweden, the data set is only using loans that are paid out to the borrowers, that are active and that have existed for at least two months. When retrieving the data set from the bank, a mutual decision was made to only include delinquency data for no more than 90 days. This is because the number of loans that have been more delinquent than that make up a fraction of the total sample that would be too small for any useful analysis to be made. This means this study will not actually look at loans that have defaulted, which can seem problematic since being delinquent per se does not bring with it any loss, it only makes a time shift for the payback transaction, given that the borrower eventually pays back the loan, and the default rate is what is eventually measured when calculating the probability of default, i.e. risk, the determinant of the price of the loan. However, one can assume that it would be in The Bank's interest to keep delinquencies to a minimum, which means the results of this study will still be useful. Furthermore, the study does not look in to calculations such as *loss given default*, which is a measurement of how much a defaulted loan actually costs The Bank. When a loan defaults, it is sent to a collection company whose goal is to recover as much of the outstanding debt as possible from the borrower and then return it to The Bank, minus their recovery fee. We decided to leave this out of the study since we will not be looking at defaulted loans, but we still think it is worth to mention this.

1.8. Definition of terms

Applicant - A person applying for a loan with the bank, but that has not yet been approved

Bank statement - A document containing all transactions to and from a bank account

Borrower - A person who has a loan with The Bank

Credit check - A document produced, usually by UC, and sold to banks containing information about a person's historical income, debts, UC risk score and more

Credit process - The process where a loan application is processed in order to be either approved or denied

Credit team - The team of employees at The Bank making loan decisions, granting and denying applications

UC - Upplysningscentralen, the biggest company in Sweden for providing data about people's financial situation

UC Score - UC's risk score, defined as a probability of delinquency (at least 1 day) within the next 12 months, with higher score meaning higher risk

2. Theoretical framework

This section discusses different aspects of risk and credit risk models and presents literature relevant to the use of hard and soft information in risk assessment made by lending banks. The literature covers both an overview of credit risk and more specific traits that are to be analyzed in the research, gambling, collection payments and dishonesty in a borrower's loan application. These are factors that may affect the risk of a loan.

2.1. Defining risk

Sitkin & Pablo (1992) propose a way to define risk by looking at the characteristics of an outcome. They argue the risk of a certain outcome is defined as the *uncertainty*, *expectations* and *potential* of the given outcome.

Outcome uncertainty

The outcome uncertainty is what is most commonly associated with risk, and is defined as "*variability of outcomes, lack of knowledge of the distribution potential outcomes and the uncontrollability of outcome attainment*" (p. 11). They argue variability of outcomes increases risk since higher variance makes it more difficult to accurately predict the actual outcome. Lack of knowledge increases risk since the prediction is unaware of all possible outcomes and the likelihood of each of these outcomes, and without being able to see the whole picture it is impossible to correctly calculate the expectancies of each outcome. They also mean that outcomes that are uncontrollable and that happen by chance are riskier than controllable and influenceable outcomes since the decision maker can affect them.

Outcome Expectations

The risk in terms of the outcomes' expectations is located in the gap between the aspirations of the stakeholders and the mean of the distribution of the expected outcomes, meaning the stakeholders might often have unrealistic expectations of the outcomes while the actual expectancy is lower. This means even positive outcomes, i.e. a stock that gives a positive return, might still be disappointing and perceived as a negative outcome if the return is not as big as expected.

Outcome potential

Individuals tend to overweight the extreme outcomes even if their probability is very low, e.g. lottery ticket buyers often overvalue the probability that they win the grand prize and therefore have unrealistic expectations of the expected return on buying the lottery ticket. Therefore, both low and high extremes in the outcome distribution increase the risk of miscalculating the expected return. For example, a bigger loan might appear to be higher risk in the qualitative assessment by the credit team since a big loan that defaults inflicts a bigger loss than the default of a small loan, even though the actual risk might be smaller, and it might therefore be denied or given a higher interest rate.

A study by Jorion (2009) on the risk management lessons from the financial crisis of 2008 brings to the table another definition of risk, and suggests that the risk factors can be divided into three different categories: *known knowns*, *known unknowns* and *unknown unknowns.* The *known knowns* are defined as risk factors where the risk managers or decision makers are able to identify and correctly measure everything related to the risk itself, and can therefore manage risk/return levels perfectly based on this knowledge. However, Jorion also argues "*Risk management, even if flawlessly executed, does not guarantee that big losses will not occur. Big losses can occur because of business decisions and bad luck*" (p. 932). By this, he means that even in an ideal scenario where all risk factors would be known knowns, you could still end up with great losses due to either business decisions such as the exposure and beta value of a stock portfolio, and simply bad luck such as a market crash.

Known unknowns are factors in the risk measurement system that are either known factors that are overlooked or not considered, or that are measured or weighted incorrectly. An example of a known unknown in credit risk assessment could be the bank statement data, which this report is investigating. The third category of risk, *unknown unknowns*, are factors that cannot be measured or predicted, such as regulatory changes.

His conclusions are that because of these three types of risk, risk management can never be perfectly executed since there will always be factors that are practically impossible to predict and measure. However, by reducing the known unknowns and unknown unknowns, making

them known knowns, risk assessment will become more accurate and reflect reality better, making risk management more effective. Manually assessing bank statement factors, for example gambling habits, as is done today at The Bank, makes the gambling risk a known unknown, since the credit team knows it exists but does not know to what extent it affects risk or how it should be weighted. By measuring this and incorporating it in the pricing model, i.e. to set the interest rate, it will be converted to a known known which means the risk in the credit process will decrease.

2.2. Credit models and pricing of risk

2.2.1. Fundamental credit model

Steenackers et al. (1989) lay the foundation for how a risk-based credit score model can be built. The credit score model uses factors that have proved to affect probability of repaying a loan and weigh them differently based on how big their effect on the repayment is. They show both a model based on linear regression and a model based on logistic regression. The credit score models are used by a credit manager to get an automatic estimate of a loan applicant's' ability to repay a loan, rather than manually and subjectively assessing the different factors, such as income, credit history or net worth.

The mathematical models are built in a way that they create a score that is higher if an applicant has a high probability of paying back their loan on time. These models can be used to deny applications that are deemed too risky. The two models are different, but based on the same logic: that there are underlying factors affecting an applicant’s ability to repay a loan and that they are of different weight. The study analyzed loans from a Belgian credit company, and divided them in “good” and “bad” loans to define how different factors affected the repayment ability. In this case, “bad” loans were those where three or more payment reminders were sent to the borrower. Mathematically, a credit scoring system can be expressed as a decision rule based on a linear function as shown in model 1 below.

Model 1 - Basic credit scoring model

$$f(X_1, ..., X_k) = b_1X_1 + b_2X_2 + ... + b_kX_k,$$

where

$X_i =$ relevant characteristic,

$b_i =$ weight or score corresponding to characteristic X_i.

Model 1 gives a weighted credit score, but it is not compatible with categorical variables such as marital status. Therefore, the model is developed to take all different kinds of factors in consideration, i.e. giving a higher score to a person who own a house in contrast to a person who is not a homeowner. The developed model is shown below.

Model 2 - Developed basic credit scoring model

$$p_x = \frac{e^{b_n+b_1X_1+...+b_kX_k}}{1+e^{h_0+h_1X_1+...+b_kX_k}},$$

where

$X_{i=}$ relevant characteristic,

$b_i =$ corresponding weight.

Credit scoring models such as the ones presented above are used to define good and bad clients in order to improve accuracy of risk assessment and thereby interest rates and denial of applicants that are too risky, all to improve the economic profit of the lender (Řezáč & Řezáč, 2009).

2.2.2. Risk-based pricing based on hard information

Edelberg (2006) published a quantitative study that uses data on interest rates in USA during the 1990's to determine risk premium spreads increased during the period. This is linked to the fact that the majority of US banks had a "House Rate", an interest rate that was the same for all customers. The risk assessment that the banks did was only done to deny risky customers, but around the 1990's they started to adjust interest rates to base them on the individual risk of every customer. This enabled banks to be more competitive to low-risk customers and to avoid charging high-risk customers less than their risk level justifies. Working with individual interest rates makes the work with all factors affecting risk

important, every identified factor will make pricing more correct. The logic is that risk should always be reflected in the interest rate.

Edelberg concludes that the risk-based pricing affected different risk groups in different ways. On one hand, very high-risk customers could get a loan that they would previously be denied, albeit with a high interest rate. On the other hand, low-risk customers saw their interest rates go down since the banks realized that there was a gap between the paid risk-premium and the actual risk and had to adjust to competition in lowering the interest rates for low-risk customers. High-risk customers saw their relative premiums go up and changed their borrowing in response. Overall, more knowledge about risk creates a more competitive and efficient market.

There is substantial evidence that an applicant's income affects their ability to repay a loan. This is used by banks that have analyzed their historical data to make credit scores, where low-income samples got lower score than high-income samples (Mester, 1997). Furthermore, Kočenda & Vojtek (2011) and Volkwein & Szelest (1995) show that other factors such as a borrower's financial resources, the purpose of the loan, marital status and education level and field all prove to affect a borrower's repayment ability and the probability of default of a loan.

2.3. Risk assessment using alternative data

Abdou et al. (2006) investigate how creditworthiness can be assessed in a developing country, in this case Cameroon, where access to financial information that is usually accessible in developed countries, such as verified income information and credit history, is limited or not available. This can be a bridge to motivating how not only such financial information can matter for credit scoring, but that other factors such as whether the applicant owns a cell phone or not, are relevant as well, which indicates looking at bank statement data will probably have an explanatory value for determining the creditworthiness of a loan applicant. However, this study is conducted in an environment that is fundamentally different from the Swedish market which our study is focusing on. Therefore, the results of this study are not expected to be directly transferable to the Swedish market, but they are included as

proof that other factors than those usually considered for credit assessment can be used to build an accurate credit model and that such factors are relevant to consider when building such a model. Further supporting evidence of this can be found in Wongnaa and Awunyo-Vitor (2013) which conduct a similar study among farmers in Ghana, and similarly to Abdou et al. (2006) manage to build a relatively accurate credit model even though the scoring data was limited.

Furthermore, Wang et al. (2018) study how bank card transaction data can be used for credit assessment in the Chinese microcredit market. Their new improved credit model, containing not only hard financial information but also transactional information, improved the credit assessment accuracy by 13,6 %. While the Chinese market and socioeconomic environment differs from the Swedish, the results of this study indicates transactional data could also be relevant for credit assessment of private loans in Sweden, which further motivates the relevance of this study on the Swedish market.

Miller (2015) further concluded in her study that access to more information about the loan applicants will result in a more accurate credit model, and thereby making the risk levels more manageable for the risk managers at the lending parties. Friendship and networks can also affect the ability to get a loan and the interest rate given in peer-to-peer lending where individual lenders give unsecured microloans to individuals. A quantitative study was done by Lin et al. (2013) using data from an online peer-to-peer lending service, measuring financial effects of the website's friendship and network system, where website users can establish contact and form a formal network bond such as a "friendship". The study finds that a friendship on the website works as a signal of credit quality. The applicants with friendships have a higher probability of getting a loan, and get lower interest rates. However, the friendships are not false signals of creditworthiness, they actually improve applicant's ability to repay the loan on time. This evidence suggests that soft factors such as social capital has an effect both on the perceived credit quality and the actual credit quality of applicants. The exact relationship between social capital and credit quality is not necessarily quantifiable, but this still proves that the hard factors, such as income or net worth, are not the only ones that matter when assessing credit risk.

Furthermore, Khandani et al. (2010) made research on credit card transactions combined with credit bureau information to more accurately predict risk levels of applicants using machine learning algorithms and managed to reduce credit losses by 6-25 % for major commercial banks in the US. This is further evidence that transactional data, even debit card transactions which is to be used for this study, as opposed to credit card transactions which were used by Khandani et al., is likely to improve the accuracy of the credit risk assessment for Swedish private loans.

2.4. The effects of gambling on private economy

The main focus of this study is to determine whether gambling, collection agency payments and application dishonesty can be used to predict the probability of loan default in credit risk assessment. In this subsection, literature related to gambling and its effects on private economy are discussed.

2.4.1. Traits associated with gambling

Problematic gambling is linked to specific personality traits and behaviours. Internet gambling frequency is significantly correlated with poor mental health (Petry & Weinstock, 2007). There is also a connection between the personality trait of sensation seeking and the frequency of gambling (Fischer & Smith, 2008). Furthermore, pathological gambling is shown to be connected with impulsivity in a person. The impulsivity negatively affected the pathological gambler's ability to make wise financial decisions, i.e. causing the gambler to choose $500 now over $1000 in one year (Alessi & Petry, 2003). These findings indicate that gambling is not just a problem when it comes to the actual money losses affecting the gambler's private economy, but that it also brings other underlying problems with it, regardless if gambling causes those problems or if it is caused by them. Those problems can be assumed to affect the ability to repay a loan negatively.

2.4.2. Gambling-related debt

A qualitative study by Downs & Woolrych (2010) shows the relationship between problematic gambling and family and work life. Problematic gambling is here defined as the gambling of a person with gambling-related debt. The debt can be direct, with borrowed

money being used in gambling, or indirect, with gambling causing a shortage of money leading to the person taking out a loan to cover other expenses. The debt levels for the defined problematic gamblers were between 2,000 and 144,000 GBP. The study shows that people who have a problematic gambling behaviour experience negative consequences in their private finances as well as in their family life. Being a problematic gambler leads to a lessened ability to keep or get a job, with gamblers reporting that they have a hard time concentrating on their work. This in turn leads to employers proceeding with disciplinary actions related to absenteeism, misuse of computer facilities and theft in the workplace. The fact that problematic gambling leads to both gambling debt and problems to keep or get a job indicates that there may be a connection between gambling and the ability to repay a loan. The problematic gambling also leads to social costs in form of losing a home and friends as a result of a forced sale caused by debt, as well as mental health problems with anxiety about debt. In conclusion, problematic gambling can lead to severe consequences in the family and work life, affecting both the problematic gamblers themselves and the people around them. This problematic gambling behaviour can be seen as a major risk factor financially, but the study does not present quantitative thresholds as to what levels of gambling are identifiable as financially problematic. However, this indicates that gambling habits in general could be connected to the risk of a person having problems repaying their loans. The study does not discuss gamblers that are not seen as problematic, so the effects of the "normal" gambling are unknown, making it relevant to investigate further.

2.4.3. Personal bankruptcy

Another study (Barron, 2002) investigates the relation between the presence of casinos and rates of personal bankruptcies. Areas with casinos present or nearby are compared to areas without casinos to show how the presence affects the population. The study shows that areas with casinos have higher rates of personal bankruptcy than areas without. After controlling for other factors such as social and economic stigma of filing for bankruptcy, previous debt and expense shocks it can be told that the volume of gambling has a direct relation to the level of personal bankruptcies. The research model shows that the areas with casinos would have their bankruptcy rate 5,4 % lower if there was no casino gambling at all in the area. Since the study is focused on offline gambling the conclusions concerning local effects of gambling will work out differently in a time when internet gambling is available in

practically all areas of a modern country. This is noted in the study, with the remark that local effects will be very hard to analyze since there will be very few areas without gambling (Barron, 2002).

The Downs & Woolrych (2010) and Barron (2002) studies show coherent results with the general conclusion that gambling prevalence in a society will lead to severe economic consequences for certain people. While Downs & Woolrych (2010) show specific problems, especially private financial ones, connected with problem gambling, Barron (2002) show that a prevalence of gambling leads to a higher rate of personal bankruptcies. This indicates that gambling, that is currently available to practically all adults with internet access, will cause a share of all participating individuals to have severe economic problems. Since Barron (2002) only measure the share of people who file for bankruptcy, the rate of people experiencing less severe economic problems are unknown, but Downs & Woolrych (2010) indicate that there are different levels of consequences, with different problematic gamblers experiencing different problems. This could mean that, on top of the bankruptcies, there are problematic gamblers with smaller problems, such as missed loan payments. All in all, the studies on gambling show that it is very relevant to quantify the relationship between gambling habits and the ability to repay a loan. Specifically, it is proposed that gambling increases the probability of delinquency.

H1: Gambling is positively associated with the probability of delinquency.

2.5. Dishonesty in loan applications

2.5.1. Effect on loan performance

Regarding dishonesty in loan applications and its effect on payment performance, there is a link between assumed dishonesty and payment delinquency. For loan applications with unverified assets of the applicant, applications with assumed dishonesty were almost 25 percentage points more likely to cause subsequent delinquency (Garmaise, 2015). This suggests that the dishonesty of a loan applicant has direct or indirect effects on payment ability, and that it is interesting for a lending bank to discover said dishonesty. However, it

should be noted that this previous research on financial dishonesty does not have a direct measure of dishonesty, but instead uses a proxy of dishonesty. In our study, actual dishonesty of monthly income will be checked for with the help of bank statement data.

In the early 2000's the total mortgage credit in low income areas in the USA increased considerably. The stated income of loan applicants from certain low income areas with negative income growth grew at a high rate, and it is suggested that this was not due to certain individuals experiencing income growth, but rather due to fraudulent loan applications (Mian et al, 2017). This shows that dishonesty is something that occurs in the credit business, and that it is not captured by the credit screening if not specifically looked at since the applicants hide the fact that they are lying. Based on the theory and evidence provided by Garmaise (2015) we propose that loan application dishonesty increases the probability of delinquency.

H2: Dishonesty is positively associated with the probability of delinquency.

2.6. Collection payments

2.6.1. The effect on loan risk

A debtor with a history of collection agency payments may imply a higher risk for a lender. When collection payment agencies are reviewed, it is found that the achieved collection rate of the debt is lower if the debtor in question has gone through a collection process with the debt collection agency before, indicating that prior experience of debt collection is a sign of low ability or willingness to pay (Beck, 2017). This shows the relevance of investigating if loan applicants have made payments to debt collection agencies in the past, since it presumably increases the risk of the applicant. It could be related to factors such as income and wealth, that poor people are more prone to find themselves in debt collection processes, but it could also be that collection payments have a direct effect on loan payment ability. Thus, it is proposed that collection payments may indicate higher probability of delinquency.

H3: Making collection payments is positively associated with the probability of delinquency.

2.7. Summary of reviewed literature

The literature reviewed covers relevant parts of how risk factors are assessed in a lending process and relevant factors that can affect the risk. It is showed that risk can be measured in different ways in order to be priced as an interest rate. Different kinds of risk are shown to be affecting the repayment ability, with both hard factors such as income and soft factors such as personal traits. The literature on the specific factors gambling, collection payments and application dishonesty show that it is likely that these specific soft factors can be used to assess risk and to improve the risk process in banks. The literature also shows that it is very likely that the mentioned factors will affect the borrowers ability to repay a loan negatively, with gambling, collection payment history and application dishonesty increasing the risk of delinquency of a customer.

2.8. Hypotheses

Supported by the previous literature on the subject, the three hypotheses formulated for this thesis have been listed below.

H1: Gambling is positively associated with the probability of delinquency.
H2: Dishonesty is positively associated with the probability of delinquency.
H3: Making collection payments is positively associated with the probability of delinquency.

There are numerous types of variables in a borrowers' bank statement, ranging from the monthly spend on food or gasoline for their car, but some of those variables are not as clearly connected to the responding borrowers' credit risk. Partially because they are not connected in the same way to risky behaviour, but also because those costs can vary greatly based on family size, geographic location of their home and more. It would of course be interesting to dig deeper into this data, but due to a lack of resources and time for this study, it had to be limited to only a few of these variables. This thesis chose not to include all types of transactions and focus on the three ones mentioned previously; gambling, collection and dishonesty. The reason we chose to include gambling and dishonesty is because it can be seen

as a compromising behaviour which indicates an unstable way of living and handling your private economy, which would in turn indicate higher risk. These are furthermore factors that can in no way be incorporated in the current risk models, why they are deemed highly relevant. Collection payments are included since they indicate previous problems with payments, which would reasonably increase the risk of future payment problems.

3. Methodology

We are doing a quantitative study to examine the association between gambling, collection payments and dishonesty in stated income and credit risk of private loans. The data for the analysis is provided by The Bank, as described below.

3.1. Data

3.1.1. General information about The Bank

The bank used in this data set is a Swedish loan provider in the Swedish market, and has private loans as its sole product. For privacy reasons, the name of The Bank or any other identifying information will not be disclosed in this study. It is a relatively small actor, but has a large and growing customer base which makes the data useful and applicable to the market as a whole.

3.1.2. Risk assessment process

In order to determine whether a loan application should be approved or denied, The Bank uses three main sources of information. Firstly, a credit check from UC is used. Based on the information in the credit check, there are a number of credit rules that are used to make it easier for the credit team to get an overview as well as for the system make automatic approvals or denials and thereby lower the workload. The credit check also includes a credit score which is calculated based on the variables in the credit check. This credit score is what determines which risk class the applicant is given, and since the risk class is what sets the interest rates at The Bank, the credit score indirectly sets the interest rate of the loan the applicant is applying for.

Secondly, a budget calculation is made using the applicants' income and costs in order to make sure there is room in their monthly budget to pay for the loan they are applying for. It includes absolute numbers provided by the applicant, such as accommodation costs and income, as well as standardized costs such as a flat-rate cost for each child or car as well as living costs. This is then put into the calculation together with the monthly cost of the loan

the applicant is applying for to make sure the applicant has room left in their monthly budget to pay for the loan. The loans at The Bank are repaid through monthly installments during a specific time period chosen by the borrower, between 1-15 years. Once the loan has been granted, the time period is static and cannot be changed.

Thirdly, bank statement data is used. The Bank is collecting the transactions on the applicants' bank statement for the past six (6) months using a digital tool to assure it can not be manipulated by the borrower. This is mandatory for some customers, and voluntary for others depending on which credit rules are hit by the application. For privacy reasons, The Bank does not disclose which rules are forcing the bank statement data collection. The Bank is currently using the bank statement information to a limited extent, i.e. as a supplementary source of information to corroborate the information provided by the potential borrower, but the bank statement information is not affecting the terms of the loan such as the interest rate, which should reflect the real risk, neither for low or high risk behavioural patterns. These transactions are then automatically categorized by type of transaction, such as income, housing costs, gambling and more. This information is then manually reviewed by the credit team and used to manually assess whether a loan should be granted. It is not included in the credit model, but in the qualitative and manual part of the process. The use of bank statement information is not widespread in the credit assessment processes by Swedish loan providers.

3.2. Variable description

3.2.1. Delinquency variables

The three delinquency variables, *Ever_30DPD, Ever_60DPD* and *Ever_90DPD* are binary variables that indicate whether a loan has ever been delinquent by 30, 60 or 90 days past due date at the same time. It is coded as 1 for yes and 0 for no. It also means that if a loan has the value 1 for *Ever_90DPD*, it must also have the value 1 for both *Ever_30DPD* and *Ever_60DPD* since it must at some point have been delinquent by 30 and 60 days in order to eventually reach 90 days.

3.2.2. Main independent variables

To measure the effects of gambling, collection and income inflation, i.e. dishonesty, the variables *DummyColl, DummyGamb* and *IncOverstate* were created. The borrowers were divided into groups based on our three main areas of study - gambling, collection payments and dishonesty when stating income. The variable *DummyGamb* is set to Yes (1) if the borrower has made at least one transaction to a gambling company and No (0) if he or she has not, *DummyColl* is set to Yes (1) if the borrower has made at least one transaction to a collection company and No (0) if s/he has not, and *IncOverstate* is set to Yes (1) if the stated income is at least 10 % higher than the actual income identified in the bank statement data, and No (0) in every other case.

3.2.3. Demographic variables

The variables defined as demographic variables are age and gender. These variables are collected from the borrowers' Swedish personal identification number and are therefore very reliable.

3.2.4. Loan related variables

The variables defined as loan related variables are *Maturity, AppliedAmount* and *APR*. The first two, *Maturity* and *AppliedAmount* are entered by the borrower when making the loan application and *APR* is the interest rate set by the bank. The Maturity is fixed and once a loan agreement is signed it cannot be changed. The AppliedAmount is the amount that was paid out to the borrower once the loan was granted. The APR is, as discussed, based on the UC Score, see below.

3.2.5. UC Score

Upplysningscentralen (UC) is Sweden's biggest credit rating company, making credit scores for people and businesses. Their main service for private loans is a prediction of the probability of missed payments of at least 1000 SEK. The estimation is based on hard factors such as taxed income, family status, payment history etc that is mostly collected through official Swedish government authorities (UC Kredit, 2019). The collected information is

assumably weighed together with UC's internal risk model - however the exact algorithm is confidential. The product is a credit score available for purchase for companies and people (UC Om oss, 2019). The Bank uses the UC risk score as a determinant to set customer interest rates, and this leaves a gap for other factors that might affect repayment probability. The factors missed by a credit rating company such as UC are presumably soft factors that are not available through the government institutions.

3.2.6. Variables stated by borrower

The variables *MaritalStatus, Purpose* and *YearlyIncome* have in common that this information is stated by the loan applicant upon application. The values in the first two variables are text values. The latter is stated by the applicant and are therefore not confirmed to be reliable. The variable NetMonthlyIncome is calculated using *YearlyIncome* and calculated as the net income using the average Swedish tax rate of 2018, 32,12 %, and also accounting for the high income tax brackets of the same year. (Skatteverket)

3.2.7. Bank statement related variables

Furthermore, there are a number of unused variables created based on the bank statement information. A description of these variables are available in Appendix 3. These variables were used in multiple different combinations to attempt creating a more accurate logistic regression model, but did not show any significant results. The results of these models were decided to be left out due to its irrelevance, but the variables are included for inspiration for further studies on the subject on other datasets.

3.3. Research design

To test our hypotheses, we build a logistic regression model to predict the probability of default of loans using bank statement data combined with data that is already in use today in order to make the prediction more accurate, hence being able to adjust pricing to a more fair level. The main model is as follows:

$$\text{logit}(P(\text{Delinquency} = 1)) = \alpha + \beta_1 Gambling + \beta_2 CollectionPayments + \beta_3 Dishonesty + Controls + \varepsilon$$

where Delinquency is the main dependent variable and is *Ever_30DPD, Ever_60DPD* or *Ever_90DPD* depending on the model. Hypotheses H1, H2 and H3 predict that the coefficients on the variables of interest will be positive and significant. We include controls for the UC score, demographic variables, and loan related variables as described above.

Other than building a regression model, we will also use t-tests and mean comparisons to analyze how good The Bank and UC are at predicting risk. We will do this by exploring whether there is any difference in delinquency between those who gamble, pay to collection companies and/or lie about their income, and those who don't as well as comparing this to the risk perceived by the bank and UC which would be reflected in the APR and UC Score respectively.

3.4. Sample description

The sample is randomly selected from The Bank's loan stock and consists of a total of 5956 observations, of which 2328 has shared their bank statement information with The Bank. The loans represented in the sample were initiated between 2015-08-21 and 2019-01-31. Table 1 presents the descriptives of the full sample, while table 2 includes only the subsample of borrowers that have shared their bank statement information with The Bank. In the descriptives for the full sample, the bank statement variables are not included since they are only available for the subsample.

	Full sample					
	N	**Minimum**	**Maximum**	**Mean**	**Median**	**Std. Dev**
Gender	*5956*	*0,00*	*1,00*	*0,37*	*0,00*	*0,48*
Age	*5956*	*18,00*	*84,00*	*39,43*	*37,00*	*13,20*
YearlyIncome	*5956*	*144000,00*	*7200000,00*	*352266,20*	*324000,00*	*173974,90*
AppliedAmount	*5956*	*5000,00*	*500000,00*	*113324,00*	*98000,00*	*88196,05*
Maturity	*5956*	*6,00*	*180,00*	*8971,00*	*84,00*	*48,86*
APR	*5956*	*2,95*	*17,48*	*10,18*	*10,43*	*3,54*
UCScore	*5956*	*0,0001*	*24,97*	*5,10*	*2,11*	*6,38*
Ever_30DPD	*5956*	*0,00*	*1,00*	*0,07*	*0,00*	*0,26*
Ever_60DPD	*5956*	*0,00*	*1,00*	*0,04*	*0,00*	*0,20*
Ever_90DPD	*5956*	*0,00*	*1,00*	*0,03*	*0,00*	*0,18*
NetMonthlyIncome	*5956*	*8145,60*	*267627,92*	*19574,66*	*18327,60*	*7319,88*

Table 1 -Descriptive statistics for the full sample of loans

When looking at the demographics of the borrowers, the table shows that 37 % of the borrowers are female and 63 % are male, since the gender variable is coded 0 for male and 1 for female. The average age is around 39 years, and the yearly income is on average roughly 352 KSEK. Apparently, at least one observation does not have information about age, since the minimum value is 0. However, since age will not be one of the main variables to analyze for this study, we choose to still include those observations since they still have valid value for the relevant variables. The maximum yearly income is very high, 7,200 KSEK, and this might be problematic. However, since it is what is registered in the data set that comes from a reliable source, that observation will still be included in the analysis.

Looking at the loan information, the average loan amounts to 113,324 SEK and has a maturity of almost 90 months, i.e. about 7,5 years. The average UC Score is 5,10 % which generates an average interest rate of 10,18 %. However, the median UC Score is much lower than the average, 2,11 %, which indicates that most of the observations are lower than the average, with the higher UC Scores having a big impact on the average score. About 7 % of the loans are at some point delinquent by at least 30 days, or two missing payments, ~4 % are at some point delinquent for at least 60 days, or three missing payments, and ~3 % are ever delinquent by at least 90 days, or four payments.

The demographics of the subsample including only the borrowers who shared bank statement are similar to those of the whole sample. There are no drastic differences to be noted. However, the UC Scores as well as the APR are slightly higher for the subsample than for the whole sample. This might be a result of the fact that the highest risk customers are forced to share their bank statements when applying for a loan.

Looking at gambling habits, you can see that the average borrower has 6,24 gambling transactions per month, amounting to on average 1,116.88 SEK per month. The borrower who spent the most money on gambling has an average net sum of all inbound and outbound transactions of -115,082 SEK per month, and the most frequent gambler makes more than 200 gambling transactions per month. 63 % of all borrowers has had at least one transaction to or from a gambling company.

For collection payments, 40 % of the borrowers in the sample have made at least one transaction to a collection company in the time of available bank statement data. The average borrower makes 0.19 transactions to collection companies per month, averaging 178.91 SEK per month. The highest amount spent per month on collection payments is 25,929.55 SEK and the most frequent collection payer makes 4.04 transactions per month.

On average, the borrowers inflate their income by 5.56 %, and 46 % of the borrowers have inflated their income by more than 10 % and are therefore considered to have overstated their income in our analysis.

	Sample of interest (shared BS)					
	N	Minimum	Maximum	Mean	Median	Std. Dev
Gender	2328	0,00	1,00	0,36	0,00	0,48
Age	2328	19,00	80,00	39,58	37,00	13,00
YearlyIncome	2328	144000,00	3960000,00	355062,28	324000,00	165198,52
AppliedAmount	2328	5000,00	500000,00	115264,80	100000,00	87257,59
Maturity	2328	12,00	180,00	91,59	84,00	48,99
APR	2328	2,95	17,48	10,30	10,73	3,55
UCScore	2328	0,0001	24,96	5,36	2,37	6,50
Ever_30DPD	2328	0,00	1,00	0,06	0,00	0,25
Ever_60DPD	2328	0,00	1,00	0,04	0,00	0,20
Ever_90DPD	2328	0,00	1,00	0,03	0,00	0,17
NetMonthlyIncome	2328	8145,60	151851,92	19703,39	18327,60	7080,12
No_Months	2328	1,00	12,00	7,15	6,00	2,98
SumGambMonth	2328	-115082,13	95086,50	-1116,88	-18,67	6310,76
FreqGambMonth	2328	0,00	201,17	6,24	0,42	17,20
AvgGamb	2328	-120000,00	28000,00	-389,50	-33,71	3125,91
SumCollMonth	2328	-25929,55	0,00	-178,91	0,00	747,97
FreqCollMonth	2328	0,00	4,04	0,19	0,00	0,37
AvgColl	2328	-102801,00	0,00	-388,36	0,00	2306,14
BS_SalaryMonth	2328	0,00	159149,33	20439,67	19466,99	11899,42
BS_FkassaMonth	2328	-6094,14	26798,27	1108,69	0,00	2987,68
BS_IncomeMonth	2328	-2344,86	159149,33	21548,36	20193,17	11783,61
MonthlyIncomeRatio	2328	0,00	868,34	105,56	107,83	55,12
GambPerIncomeMonth	2328	-2546,05	12887,83	12,31	0,06	277,72
CollPerIncomeMonth	2328	-249,46	871,70	2,06	0,00	28,34
DaysSinceGamb	2328	0,00	700,00	15,67	0,00	40,60
DaysSinceColl	2328	0,00	601,00	24,31	0,00	50,07
GambLossAvg	1432	-120000,00	-16,00	-1494,22	-340,00	4341,77
GambLossSumMonth	1432	-267539,78	-1,58	-8939,43	-816,88	22934,52
GambLossFreqMonth	1432	0,08	183,00	7,13	1,83	15,05
DummyGamb	2328	0,00	1,00	0,63	1,00	0,48
DummyColl	2328	0,00	1,00	0,40	0,00	0,49
IncOverstate	2328	0,00	1,00	0,46	0,00	0,50
Valid N (listwise)	**1432**					

Table 2 -Descriptive statistics for the subsample that has shared bank statement data

Furthermore, the distributions of the variables Family Status and Purpose are shown in the tables below. Since their values are not quantifiable, they were excluded from the full descriptives table and shown separately. The most common family status among the borrowers is single (42.49 %), followed by married (29.13 %) and sambo (23.35 %), which is living with a partner to whom you are not married. A majority (59.12 %) of all loans issued by The Bank are for consolidating other loans or credits into one loan, probably to a lower interest rate. This would indicate The Bank has competitive pricing and frequently steal customers from their competitors. The second most common loan purpose is for buying a vehicle (14.34 %).

Purpose	***# of loans***	***% of loans***
LoanConsolidation	*3521*	*59,12%*
Vehicle	*854*	*14,34%*
Other	*512*	*8,60%*
HomeImprovement	*339*	*5,69%*
Vacation	*195*	*3,27%*
KitchenAppliances	*112*	*1,88%*
PersonalHealth	*93*	*1,56%*
Move	*78*	*1,31%*
Renovation	*77*	*1,29%*
Education	*61*	*1,02%*
Marriage	*42*	*0,71%*
HomeElectronics	*40*	*0,67%*
Divorce	*18*	*0,30%*
None	*10*	*0,17%*
Business	*4*	*0,07%*
Total	***5956***	***100%***

Table 3 -Distribution of loan purposes (full sample)

Family status	***# of loans***	***% of loans***
Single	*2531*	*42,49%*
Married	*1735*	*29,13%*
Sambo	*1391*	*23,35%*
Divorced	*213*	*3,58%*
Apart	*58*	*0,97%*
Widower	*28*	*0,47%*
Total	***5956***	***100%***

Table 4 - Distribution of family statuses (full sample)

3.5. Previous knowledge

Since this is the final thesis of a Bachelor of Science in economics for the two authors, they have also acquired a base of knowledge in finance and economics that will prove useful for this report. They also have professional experience from working at a lending company and a collection company respectively.

4. Results

4.1. Gambling, collection and dishonesty effect on delinquency

Since both the behavioural variables and the delinquency variables, *Ever_30DPD, Ever_60DPD* and *Ever_90DPD*, are binary, univariate logit tests were run to look for any potential connections between these behaviours and the performance of the respective loans of these borrowers. The delinquency variables, as mentioned above, indicated whether a loan has ever been delinquent for 30, 60 or 90 days respectively. The results of these tests are shown in table 5 below. The coefficients are not significantly different than 0 indicating tha gambling, making collection payments and dishonesty in application does not have any effect on the delinquency of the loans, and that they therefore are unlikely to add additional information when calculating the interest rates of the loans.

Behaviour var	**Delinquency var**	**Coef.**	**Sig.**
DummyGamb	Ever_30DPD	-0.008	0.962
DummyGamb	Ever_60DPD	0.058	0.795
DummyGamb	Ever_90DPD	0.127	0.628
DummyColl	Ever_30DPD	0.161	0.346
DummyColl	Ever_60DPD	0.070	0.747
DummyColl	Ever_90DPD	0.026	0.918
IncOverState	Ever_30DPD	-0.212	0.214
IncOverState	Ever_60DPD	-0.255	0.239
IncOverState	Ever_90DPD	-0.061	0.808

Table 5 - Logit tests for main variables of interests and delinquency

T-tests comparing the means between the delinquencies and the behavioural variables, see table 6 below, further strengthens the conclusion that there are no clear connections between gambling, collection or dishonesty and delinquency.

Tested variable	Grouping variable	Mean Delinquency No	Mean Delinquency Yes	Diff	Sig (1t)	Sig (2t)
Ever_30DPD	DummyColl	0,061	0,070	-0,010	0,173	0,346
Ever_60DPD	DummyColl	0,039	0,041	-0,003	0,373	0,745
Ever_90DPD	DummyColl	0,029	0,029	-0,001	0,459	0,918
Ever_30DPD	DummyGamb	0,649	0,644	0,005	0,519	0,962
Ever_60DPD	DummyGamb	0,038	0,040	-0,002	0,398	0,795
Ever_90DPD	DummyGamb	0,027	0,030	-0,003	0,314	0,628
Ever_30DPD	IncOverstate	0,070	0,058	0,013	0,893	0,214
Ever_60DPD	IncOverstate	0,044	0,034	0,010	0,881	0,238
Ever_90DPD	IncOverstate	0,030	0,028	0,002	0,596	0,808

Table 6 - T-tests comparing delinquencies for gamblers, collection payers and income overstaters

4.2. Logistic regression model

As mentioned in section 3.2. we initially attempted to build a multivariate logistic regression model that would be able to predict delinquency based on bank statement information. For this, we used our three main variables of interest, *DummyGamb, DummyColl* and *IncOverstate* as well as adding four control variables, *Gender, Age, YearlyIncome* and *UCScore* which can reasonably affect the ability to repay a loan. For example, a younger person might be more irresponsible and forget to pay their bills etc, and a borrower with a high income will probably repay a loan better than someone with a low income. For *Age* and *Yearlyincome,* the natural logarithm of their values were used in the regression since they are a better fit for conducting a logistic regression when all values are greater than 1. The correlations between the variables are shown in table 7 below, where *, ** and *** indicate significance at the 10 %, 5 % and 1 % levels respectively.

	Ever_60DPD	Ever_90DPD	DummyGamb	DummyColl	IncOverstate	Gender	Age	YearlyIncome	UCScore
Ever_60DPD	1.0000								
Ever_90DPD	***0.8554	1.0000							
	0.0000								
DummyGamb	0.0054	0.0101	1.0000						
	0.7954	*0.6278*							
DummyColl	0.0067	0.0021	***0.0551	1.0000					
	0.7469	*0.9178*	*0.0079*						
IncOverstate	-0.0245	-0.0050	0.0191	0.0371	1.0000				
	0.2378	*0.8083*	*0.3569*	*0.0736*					
Gender	**-0.0332	-0.0225	***-0.0724	***0.1133	***-0.0556	1.0000			
	0.0104	**0.0821*	*0.0005*	*0.0000*	*0.0073*				
Age	-0.0257	-0.0298	-0.0281	***-0.0913	***-0.0561	***0.0647	1.0000		
	0.0469	***0.0214*	*0.1748*	*0.0000*	*0.0068*	*0.0000*			
YearlyIncome	-0.0135	-0.0083	-0.0294	0.0157	***-0.0743	***-0.0893	***0.1218	1.0000	
	0.2986	*0.5229*	*0.1568*	*0.4484*	*0.0003*	*0.0000*	*0.0000*		
UCScore	-0.0088	-0.0025	0.0760	***0.1428	0.0105	***0.0538	***-0.1923	***-0.1016	1.0000
	0.4975	*0.8441*	****0.0002*	*0.0000*	*0.6112*	*0.0000*	*0.0000*	*0.0000*	

Table 7 - Table of correlations for variables used in regression analysis

These variables were then used to build the regression models shown in the tables below. Table 8 shows the models' performance in regards to predict 60 days delinquency, and table 9 shows the same models for 90 days. The results for 30 days are not included partially because they showed similarly insignificant results, partially because 60 and 90 days delinquency are to be looked at more seriously and being able to predict these will have a greater impact than being able to predict 30 days delinquency. Model 1 includes all seven variables, model 2 includes only the demographic variables and model 3 includes only the bank statement related variables. For both dependent variables, *Ever_60DPD* and *Ever_90DPD,* the model using only the demographic variables prove to be more useful which shows by the higher log likelihood, higher LR Chi2 and lower prob > chi2, i.e. significance.

Dep var	**Ever_60DPD**					
Model no	1		2		3	
Variable	**Beta**	**Sig**	**Beta**	**Sig**	**Beta**	**Sig**
Constant	*-1,556*	*0,747*	*2,606*	*0,402*	***-3,144*	*0,000*
DummyGamb	0,042	0,757			*0,058*	*0,795*
DummyColl	0,068	0,195			0,077	0,724
IncOverstate	-0,284	0,668			-0,259	0,232
Gender	-0,1	0,145	**-0,335	0,018		
AgeLN	-0,512	0,945	**-0,464	0,023		
YearlyIncomeLN	0,027	0,820	-0,304	0,222		
UCScore	-0,004	0,474	-0,011	0,282		
N		2323		5956		2323
Pseudo R2		0,0054		0,0069		0,0021
Log likelihood		- 385,112		- 1054,954		- 386,401
LR Chi2		4,19		0,1458		1,61
Prob > Chi2		0,758		0,006		0,657

Table 8 - Regression model results for 60 days or more delinquency

Dep var	**Ever_90DPD**					
Model no	1		2		3	
Variable	**Beta**	**Sig**	**Beta**	**Sig**	**Beta**	**Sig**
Constant	*-5,929*	*0,261*	*-0,358*	*0,918*	***-3,578*	*0,000*
DummyGamb	0,080	*0,762*			0,126	0,629
DummyColl	0,003	0,991			0,021	0,932
IncOverstate	0,083	0,741			0,063	0,799
Gender	-0,299	0,293	-0,233	0,147		
AgeLN	-0,643	0,121	**-0,559	0,019		
YearlyIncomeLN	0,373	0,375	-0,073	0,793		
UCScore	0,001	0,613	-0,001	0,577		
N		2323		5956		2323
Pseudo R2		0,0090		0,0053		0,0010
Log likelihood		- 300,881		- 843,839		- 303,447
LR Chi2		5,44		8,97		0,31
Prob > Chi2		0,607		0,062		0,958

Table 9 - Regression model results for 90 days or more delinquency

Furthemore, the models show that none of the bank statement related variables have a significant impact on delinquency, neither 60 nor 90 days, regardless of whether the demographic variables are included or not. This goes in line with the results of the univariate tests in 4.1. Also, an interesting observation is that when including only the demographic variables, age has a significant (5 % level) impact on delinquency both for 60 and 90 days, with quite strong negative effect of -0,464 and -0,559 respectively. Nevertheless, the aim of this study is to look at how bank statement related variables can be used, not how age affects delinquency, which makes this finding irrelevant in the current context.

4.3. Performance of The Bank's pricing today

In order to evaluate the accuracy of the pricing model used at The Bank today, a t-test was made to compare whether the groups that prove to eventually be delinquent were also predicted to be more likely to be delinquent according to the credit model, and hence given a higher interest rate.

Delinquency	**Mean APR if Delinquency = 0**	**Mean APR if Delinquency = 1**	**Diff**	**Sig (1t)**	**Sig (2t)**
Ever_30DPD	10.067	11.570	-1.503	0.000	0.000
Ever_60DPD	10.106	11.687	-1.581	0.000	0.000
Ever_90DPD	10.121	11.803	-1.682	0.000	0.000

Table 10 - Differences in mean interest rate for delinquent borrowers

The results showed that borrowers who pay higher interest rate also have a high probability of delinquency, i.e. the delinquent loans are in fact priced higher than the non-delinquent loans, and the more delinquent a loan gets, the higher the interest rate is generally, which means the pricing model functions as it should. Since the interest rates at The Bank are determined by the UC score of a borrower, this should mean that the UC score will also have the same type of ability to predict delinquency. However, this proves not to be the case. The results of the same tests but replacing APR with UCScore are shown below, and it shows that the UC Score, which is a variable where a higher value indicates higher risk, is not only lower for the loans that are 60 and 90 days late, but the results are also not significant. Hence, it is clear that the UC Score is not very effective for predicting delinquency, while The Bank still manages to predict risk more accurately. This implies that The Bank should not be relying on UC as its only source of information to set interest rates.

Variable	Mean UC Score if Delinquency = 0	Mean UC Score if Delinquency = 1	Diff	Sig (1t)	Sig (2t)
Ever_30DPD	5.097	5.165	-0.068	0.416	0.832
Ever_60DPD	5.113	4.839	0.274	0.751	0.498
Ever_90DPD	5.104	5.013	0.091	0.578	0.884

Table 11 - Differences in mean UC Score for delinquent borrowers

To investigate whether gambling, collection payments and dishonesty affect the interest rate the borrower is given, t-tests were used to compare the means between the two groups - Yes and No - for each of the three behavioural variables. The results of this is shown below, including the mean differences and 1-tailed and 2-tailed significances, and it is clear that gamblers and collection payers get a substantially higher interest rate, while dishonesty does not have any significant effect. However, as shown in table 6, these behaviours did not prove to have any effect on the actual delinquency, indicating the gamblers and collection payers gets a price premium that is not motivated by the data.

Variable	Mean APR No	Mean APR Yes	Diff	Sig (1t)	Sig (2t)
DummyColl	9.86 %	10.96 %	1.10 %	0.000	0.000
DummyGamb	9.99 %	10.48 %	0.49 %	0.001	0.001
IncOverstate	10.25 %	10.36 %	0.11 %	0.229	0.456

Table 12 - Differences in mean interest rate for gamblers, collection payers and income overstaters

To see whether these differences in interest rates are also reflected in the UC Score, thereby indicating these behaviours would be incorporated in the UC Score, the same tests were run using the UC Scores instead of APR and the results are presented in table 13 below. As can be seen, the differences in UC Score reflect the differences in APR quite well for the three groups, where collection has the highest effect and income inflation is unrelated. This might be explained by the fact that UC has access to collection history, while they do not have access to gambling habit information.

Variable	Mean UC No	Mean UC Yes	Diff	Sig (1t)	Sig (2t)
DummyColl	4.602 %	6.498 %	1.896	0.000	0.000
DummyGamb	4.712 %	5.735 %	1.023	0.000	0.000
IncOverstate	5.292 %	5.430 %	0.137	0.306	0.611

Table 13 - Differences in mean UC Score for gamblers, collection payers and income overstaters

5. Discussion and analysis

5.1. Collection Payments

The results shows a significant relationship between the APR and a history of collection payments, those who has made payments to collection companies on average get a 1.10 % higher interest rate, even though there was no difference to be found in regards to how well a borrower who made collection payments actually repay their loans. There is also a significant relationship between collection payments and UC score, which is what sets the interest rate, where those who have made collection payments on average have a 1.896 % higher UC score than those who have not, meaning that the borrowers with a history of collection payments are deemed riskier by UC. UC uses data from collection companies, which is probably the explanation behind this difference. The UC score is in turn used at The Bank to set the APR for a borrower, so these results are expected. This goes in line with Beck's (2017) study, deeming borrowers with a collection history as riskier. However, there is no significant relationship between a history of collection payments and payment delinquency at 30, 60 or 90 days, i.e. collection payers does not become delinquent more often than those who have no collection transactions. This is surprising since collection payments were expected to affect the borrowers risk according to Beck, that collection payments would be linked with further delinquency. And if you look at it, the collection payment itself indicates there has been previous delinquency on another loan or other similar payment that created the borrower's collection debt, which would presumably indicate a high probability it would reoccur.

These findings indicate that borrowers with a history of collection payments are in fact not more risky than those without, even though previous theory as well as UC's risk assessment says the opposite. This either means the data used in this study was biased, or a behavioral change in large scale has taken place, meaning that people who get sent to collection companies have started repaying their loans to a greater extent than before. The first scenario seems more likely, but the second scenario cannot be ruled out. While he results of this study were achieved with a fairly large data set of 2,326 observations of loans who had shared their bank statements, it is nowhere close to the entire population. It can also be biased since the

loans used for this study have gone through the manual screening of the bank statements in The Bank's credit process, indicating their levels of collection payments and gambling were decided to be on acceptable levels, and that the worst cases were never approved and hence are not included in the dataset. However, if the second scenario were to be true, it would mean the collection payers are given an interest rate that is unfairly high and they get an unmotivated risk premium that The Bank should look at removing since they would be able to attract more customers with collection payments if they could offer a lower interest rate.

5.2. Gambling

The results when looking at gambling habits are similar to those of collection payments, with on average a 0.49 % higher interest rate and 1.023 % higher UC score for gamblers compared to non-gamblers. And just as is the case with the collection payments, gambling does not seem to affect repayment or delinquency of the underlying loan. Also, this goes against what previous theory said, (see Downs & Woolrych (2010), Barron (2002)). However, Downs & Woolrych (2010) focus on "problem gamblers", while the gambling variable in this study includes everyone who have gambled at least once in the past six months, no matter how small the amounts or how low the frequency. That, combined with the fact that "problem gamblers" are likely to have been denied loans in the manual screening done by The Bank due to high levels of gambling identified in the bank statement data, indicates that while there is no connection between gambling and delinquency in this data set, gambling could still prove to affect risk. For that to be correctly measured, it would be require all banks and lenders in the market to accept all applicants, no matter their perceived risk or gambling habits, and measure how their loans perform, while also possibly distinguishing problematic gamblers from non-problematic, which has not been done in this study. .

The findings indicate that The Bank captures the presumed risk of gambling through their risk assessment process, where they use UC risk score to set the APR. That is supported by the fact that there is a significant relationship between gambling and UC score, meaning that the borrowers with a history of gambling were predicted by UC to have higher risk than those who do not gamble. This could be explained by UC having other variables that are correlated to gambling and therefore predict risk in the same way but with underlying causes. Since UC

are not using data from internet casinos or transactional data according to their list of sources, it seems that the gambling is connected to an underlying risk factor captured by UC.

5.3. Dishonesty in application

Dishonesty in the application, in contrast to the collection and gambling, does not seem to affect the UC Score nor the APR. The UC score uses taxed income as opposed to stated income and has no way to assess application dishonesty, why it is reasonable that there is no connection between this and dishonesty. Since the UC Score is what sets the APR, the lack of relationship between by dishonesty and APR can be explained by the same reason.

Furthermore, the study also suggests there is no difference in delinquency rates for those who overstate their income and those who don't, for neither of the delinquency variables. This means income inflation does not affect neither the actual nor the perceived risk, neither by UC nor by The Bank. However, Garmaise (2015) indicates that dishonesty in loan applications would be linked to risk, which goes against the results of this study, at least when it comes to stating income which is the only measurement of dishonesty that is used here. The shown results might be explained by the fact that The Bank reasonably prefers to look at net income when assessing whether to approve a loan application, and that they only approve the loans with an appropriate level of net income identified in the bank statements. Those who has stated a certain level of income but fail to verify it through their bank statements are most likely to have been denied. The previous theory captured borrowers who got loans on false pretenses, where applicants had lied to get loans, but if the income dishonesty is captured by The Bank before giving out a loan, the loan will be based on the true numbers, i.e. what is shown in the bank statement. This means that even if the dishonesty as a trait is not identified to be a factor indicating risk, the results of the dishonesty could be a risk factor if it leads to borrowers getting loans on false pretenses where banks think that income is higher than it is. Because this kind of loans on false pretenses is not expected to be found in large numbers in our data sets, the results are reasonable.

5.4. APR and UC Scores

No significant relationship was found between UC score and delinquency, which is a surprising result since the UC score's sole purpose is to predict risk. This means that the UC cannot indicate which of the approved loans will show problems with payments. While the results are surprising, they can possibly be explained with the fact that the data analyzed obviously only shows borrowers who got their loan applications approved, since the data set captures real loans that have been administered by The Bank rather than a full population containing all applicants. The Bank has reasonably denied applications of high risk, and it is reasonable to assume that the UC score is a prominent way to measure risk in order to deny applications of customers deemed too risky.

When looking at the interest rates of the delinquent loans, the interest rates get higher the more delinquent a loan has been, with on average 1.503 % higher for loans that are ever delinquent at least 30 days, 1.581 % for those ever delinquent by at least 60 days and 1.682 % for those ever delinquent by at least 90 days, which is shown in table 10. The loans have an interest rate that is supposed to reflect the risk of the borrower, and this is proven by the fact that delinquent loans were given a higher interest rate than non-delinquent ones, and that the more delinquent a loan is, the higher the interest rate. This shows that The Bank's credit model captures the risk even better than UC, since the UC scores did not show similar results, even though the UC Scores are what set The Bank's interest rates. This might be a result of the qualitative part of the credit process, i.e. the manual bank statement screening which filters out the risk factors that are not included in UC's calculations, as discussed by Jorion (2009) making the unknown unknowns into known unknowns. UC has no idea of the distribution of e.g. gambling nor its effect on actual risk, i.e. it is an unknown unknown, and hence it is not included in their risk score. On the other hand, The Bank knows the distribution of gambling with the borrowers who share their bank statements, but they don't know the exact effect on the risk, which makes it a known unknown.

In theory, the higher APR could be the cause of higher delinquency since it increases the monthly cost of the loan, but since the APR is set to reflect the risk of the borrower it is

reasonable to assume that the APR is the result of a high risk and not the other way around. This means that the APR is reasonably not a risk factor in itself, which is also clear since it is not a part of the borrower's person in the same way as income, gambling or other behavioural factors. Furthermore, even if the APR would be seen as a risk factor, it would not make sense to use it in the model since the APR will actually be set based on the outcome of the model. The APR is set according to UC risk scores deeming risk level of a borrower. The UC risk score is largely based on classic hard factors such as income and wealth, showing the point that these findings go in line with Edelberg's conclusion (2006), that hard factors have an impact on the level of risk of a borrower.

Furthermore, when looking at the descriptives in tables 1 and 2, it shows a slightly lower delinquency rate for the borrowers who shared their bank statements compared to those who did not, since 6 % of bank statement loans were at least 30 days delinquent compared to 7 % for the whole sample, and reasonably even higher rate if the non-bank statement loans would have been isolated. This indicates bank statements, even though not proven to be useful for actual risk measurement, can be used as a tool to improve risk assessment and increase the quality of approved loans. The studies by Miller (2015), Abdou et al (2006) and Wongnaa & Awunyo-Vitor (2013) all show proof that additional information that is not commonly used in credit models should be included to improve the accuracy when predicting default or delinquency, which further strengthens the conclusion that bank statements have a positive impact on risk assessment. This is something The Bank, and other banks who deal with private loans, should exploit further and it would be beneficial for these actors if bank statements were a standard procedure for all loans when making the credit risk assessment.

5.6. Connection to theory

The findings can be used to change the way these specific factors of gambling, collection payments and dishonesty are categorized and used in a risk assessment process. In line with Jorion's (2009) way of risk dividing risk factors in the three categories *known knowns*, *known unknowns* and *unknown unknowns*, these factors can to some extent change from being *known unknowns* to *known knowns* since they are now measured and proven, even though possibly to a great enough extent to rely completely on. The way these factors relate to

delinquency was on one hand not shown with significant results, but the fact that they did not show significant results shows that the non-effect that they have on delinquency is now known.

This means that a risk assessment can be more precise now that gambling, collection payments and application dishonesty do not increase the risk of a borrower. Of course, these findings might not be applicable if the data set would contain a full population instead of only containing borrowers that actually got through the application process and were granted loans. It is very interesting that the theory connected to gambling, collection payments and application dishonesty only partly went in line with the findings in this research. They largely went in line with theory with gambling and collection payments being connected to UC scores, i.e. perceived risk, but not when it came to predicting actual delinquency, i.e. actual risk.

Further, the findings on application dishonesty did not go in line with previous theory since it did not show a connection to UC score or actual delinquency. Again, these results may very well be the cause of the risk analysis done before loan approval or denial at The Bank, skewing the data set towards less risky clients since the riskier ones reasonably are denied loans. To conclude, neither of the hypotheses formulated based on previous literature proved to be true in this study.

5.7. Areas for future research

For future research it would be interesting to investigate a sample of loans that have not been through a credit process, and thereby also including loans that The Bank usually would deny to see how these loans perform. In order to build a successful model, we believe the entire spectrum of risk, where probability of default measured can be higher than the acceptable level at The Bank, and bad loans are not denied and thereby sorted out, is necessary to identify any clear connections and variables to predict default rates. It would require a bank or lending company to accept high risk by lending to everyone for a long period to generate a large enough number of observations, but the resulting model could make it a good investment in the long run, leading to an improved credit model. It would also be beneficial

to do this type of analysis using data from several different banks and lending companies since they differ in for example their risk appetite and risk assessment when making loan decisions, which would mean The Bank used in this study might only capture a small niche of the market as a whole and a specific type of borrower that is not representative for the whole market. When doing this, it would be helpful to use a machine learning algorithm that could identify and use data points that humans might not think of exploring.

5.8. Ethical considerations

Using personal financial data to improve the accuracy of risk assessment in a bank is of course a delicate subject. The integrity of every borrower is very important, both in matters of laws such as GDPR and ethically. The bank should not use findings in personal transaction data from bank statements to delve too deep into the personal space of an applicant, and it is essential that individuals are not analyzed in a way that disrupts their integrity. While the applicants are not anonymous per se, their personal data has to be handled in a way compatible with the current laws. Further, there is an ethical grey-area regarding where to draw the line when it comes to bank statement analysis. To analyze the amount transferred to online casinos every month may not be too personal and sensitive, but much more personal information could be explored by the bank, e.g. analyzing if the applicant has withdrawn money from an ATM in a “problematic” area or if the person has sent bank transfers to people regarded as “not good”. This topic has become increasingly relevant in the past few years as a result of data breaches from big companies as well as new well discussed regulations like GDPR, and must be handled with great care.

6. Conclusion

The aim of the study was to determine whether the incorporation of personal behavioural patterns would improve risk assessment in the context of Swedish private loans. These patterns are specifically the behaviour regarding gambling, collection payments and application dishonesty.

Neither gambling, collection payments or salary dishonesty of borrowers improve the risk assessment process, as they do not explain delinquency of the borrowers. The borrowers with those behavioural patterns do not induce a higher level of risk for the lender when looked at as isolated factors from bank statement data, neither when combined in a logistic regression model.

While APR did not seem to be significantly related to dishonesty, the APR of the loans was higher for the group who had gambling transactions or collection payments in their bank statement, meaning that the borrowers who showed a history of gambling or collection payments actually do pay a higher interest rate on their loans. This could be a result of direct factors, e.g. the bank denying the loans with gambling or collection transactions that don't have a high enough interest rate to compensate for the perceived risk, or due to indirect factors where other factors which are incorporated by the UC Score, which sets the interest rate, has a correlation with the behaviours in question, causing the interest rates to be higher for the applicants with higher perceived risk.

Those borrowers' interest rates are higher, which should reflect a higher probability of default, but our findings show that there is no significant difference in probability of default between the different groups with the identified behaviours. Since there is no significant difference in delinquency, the risk is the same, but since the people with history of gambling or collection payments pay more money in interest rates, they have a higher perceived risk which actually make them seem to be more valuable customers for The Bank. The borrowers who gamble or have a history of collection payment in their application are not significantly connected to a higher risk of default. However, it should of course be noted again that the

data set only contains approved applications and that high risk-applications are very likely to have been denied.

The APR is set to reflect risk, and is connected to the delinquency, which indicates that The Bank is doing a good risk analysis since a higher probability of default generates a higher interest rate. The fact that the dataset might be biased is of course of high relevance since the denial of risky applicants has changed the structure of the loan database. The data available for analysis is genuine bank data, which means that the loans have already gone through a risk assessment process where applicants deemed too risky being denied loans. It is very possible that the results of the study would be notedly different if the data would include a full population and not only borrowers that have not shown signs of high perceived risk.

7. References

Electronic sources

Dagens Industri. (2016, November 28). Otroheten ökar bland svenska bankkunder. Retrieved from https://www.di.se/nyheter/otroheten-okar-bland-svenska-bankkunder/

Emmylou Tuvhag. (2017, June 20). Blancolån ökar dramatiskt. Retrieved from https://www.svd.se/blancolan-okar-dramatiskt--mellanhander-vinnare

Hemad Razavi. (2017, October 28). En Miljardmarknad till fördel för låntagare. Retrieved from https://www.direkto.se/laneformedling_en_miljardmarknad_till_fordel_for_lantagare/

Kimberly Amadeo. (2019, March 30). What caused the 2008 global financial crisis. Retrieved from https://www.thebalance.com/what-caused-2008-global-financial-crisis-3306176

Kvalitetsindex. (2018, December 10). Personlig service utmanar digitala tjänster. Retrieved from http://nyheter.kvalitetsindex.se/documents/svenskt-kvalitetsindex-om-bolaan-fastighetslaan-privatlaan-och-sparande-2018-84284

Riksbanken. (2019, April 25). Reporänta, in- och utlåningsränta. Retrieved from https://www.riksbank.se/sv/statistik/sok-rantor--valutakurser/reporanta-in--och-utlaningsranta

SCB. (2017, December 5). Kommunalskatterna. Retrieved from https://www.scb.se/hitta-statistik/statistik-efter-amne/offentlig-ekonomi/finanser-for-den-kommunala-sektorn/kommunalskatterna/pong/statistiknyhet/kommunalskatterna-2018/

Skatteverket. (n.d.) Svar på vanliga frågor. Retrieved from https://www.skatteverket.se/privat/etjansterochblanketter/svarpavanligafragor/inkomstavtjans

t/privattjansteinkomsterfaq/narskamanbetalastatliginkomstskattochhurhogarden.5.10010ec103545f243e8000166.html

The Economist. (2013, September 7). Crash course. Retrieved from https://www.economist.com/schools-brief/2013/09/07/crash-course

UC. (n.d.). Kreditscore Person. Retrieved from https://www.uc.se/kreditscore-person/

UC. (n.d.). Om oss. Retrieved from https://www.uc.se/om-uc/om-oss/

Articles

Abdou, H. A., Tsafack, M. D. D., Ntim, C. G., & Baker, R. D. (2016). Predicting creditworthiness in retail banking with limited scoring data. *Knowledge-Based Systems*, *103*, 89-103.

Alessi, S. M., & Petry, N. M. (2003). Pathological gambling severity is associated with impulsivity in a delay discounting procedure. *Behavioural Processes*, 64(3), 345-354.

Beck, T., Grunert, J., Neus, W., & Walter, A. (2017). What Determines Collection Rates of Debt Collection Agencies?. Financial Review, 52(2), 259-279.

Barron, J. M., Staten, M. E., & Wilshusen, S. M. (2002). The impact of casino gambling on personal bankruptcy filing rates. *Contemporary Economic Policy*, *20*(4), 440-455.

Blaszczynski, A. P., Wilson, A. C., & McConaghy, N. (1986). Sensation seeking and pathological gambling. *British Journal of addiction*, 81(1), 113-117.

Downs, C., & Woolrych, R. (2010). Gambling and debt: the hidden impacts on family and work life. *Community, Work & Family*, *13*(3), 311-328.

Edelberg, W. (2006). Risk-based pricing of interest rates for consumer loans. *Journal of Monetary Economics*, *53*(8), 2283-2298.

Fischer, S., & Smith, G. T. (2008). Binge eating, problem drinking, and pathological gambling: Linking behavior to shared traits and social learning. Personality and individual Differences, 44(4), 789-800.

Garmaise, M. J. (2015). Borrower misreporting and loan performance. The Journal of Finance, 70(1), 449-484.

Jorion, P. (2009). Risk management lessons from the credit crisis. *European Financial Management*, *15*(5), 923-933.

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, *34*(11), 2767-2787.

Kočenda, E., & Vojtek, M. (2011). Default predictors in retail credit scoring: Evidence from Czech banking data. *Emerging Markets Finance and Trade*, *47*(6), 80-98.

Lin, M., Prabhala, N. R., & Viswanathan, S. (2013). Judging s by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending. *Management Science*, *59*(1), 17-35.

Mester, L. J. (1997). What's the point of credit scoring?. *Business Review*, 3(Sep/Oct), p9.

Mian, A., & Sufi, A. (2017). Fraudulent income overstatement on mortgage applications during the credit expansion of 2002 to 2005. The Review of Financial Studies, 30(6), 1832-1864.

Miller, S. (2015). Information and default in consumer credit markets: Evidence from a natural experiment. *Journal of Financial Intermediation*, *24*(1), 45-70.

Petry, N. M., & Weinstock, J. (2007). Internet gambling is common in college students and associated with poor mental health. The American Journal on Addictions, 16(5), 325-330.

Řezáč, M., & Řezáč, F. (2009, August). Measuring the quality of credit scoring models. In Credit Research Conferences.

Sitkin, S. B., & Pablo, A. L. (1992). Reconceptualizing the determinants of risk behavior. *Academy of management review*, *17*(1), 9-38.

Steenackers, A., & Goovaerts, M. (1989). A credit scoring model for personal loans. *Insurance: Mathematics & Economics*, *8*(1), 31-34.

Volkwein, J. F., & Szelest, B. P. (1995). Individual and campus characteristics associated with student loan default. Research in Higher Education, 36(1), 41-72.

Wang, H., Zhou, J., & Huang, D. (2018). RFMS Method for Credit Scoring Based on Bank Card Transaction Data.

Wongnaa, C. A., & Awunyo-Vitor, D. (2013). Factors affecting loan repayment performance among yam farmers in the Sene District, Ghana. *Agris on-line Papers in Economics and Informatics*, *5*(665-2016-44943), 111.

Appendix 1 - UC's sources

These are included to show where UC gets their information and what information is incorporated in their risk score. As seen, there are no sources of information that related to gambling nor dishonesty, while there are sources that are related to collection payments which can be seen at the bottom of the list.

Uppgift	Källa	Uppdatering	Lagring
Namn & adress på fysiska personer	Skatteverket/SPAR	5 ggr/ vecka	När uppdaterad info kommer från källa
Civilstånd	Skatteverket/SPAR	5 ggr/ vecka	När uppdaterad info kommer från källa
Taxerad inkomst	Skatteverket	4 ggr/ år (juni/ aug/ sep/ dec)	Två senaste taxeringsåren
Taxerad inkomst, omräkningar	Skatteverket	2 ggr/ månad	Två senaste taxeringsåren
Fastighetsuppgifter (taxeringsuppgift)	Skatteverket	1 gång per år	När uppdaterad info kommer från källa
Fastighetsuppgifter, lagfart etc	Lantmäteriet	5 ggr/ vecka	När uppdaterad info kommer från källa
Äktenskapsförord	Skatteverket	Varje vecka	När uppdaterad info avseende civilstånd kommer från källa
Personlig förvaltare	Tingsrätt via PoIT	5 ggr/ vecka	När uppdaterad info kommer från källa
Aktiebolag	Bolagsverket	5 ggr/ vecka	
Styrelse, VD mm i aktiebolag	Bolagsverket	7 ggr/ vecka	Gallras efter 60 månader
Revisorer i aktiebolag	Bolagsverket	7 ggr/ vecka	
Händelseuppgifter i aktiebolagsuppgifter (ej beslutade ärenden i Bolagverkets diarium)	Bolagsverket	7 ggr/ vecka	
Årsredovisningar aktiebolag,	Bolagsverket	5 ggr/ vecka	
Årsredovisningar koncern	Bolagsverket	5 ggr/ vecka	
Nyckeltal för olika branscher och storleksklasser	UCs egna beräkningar	9 ggr/ år	
Koncernstruktur	Årsredovisningar samt källor som tidningar	5 ggr/ vecka	
Besläktade företag	UCs egna analyser	5 ggr/ vecka	När uppdaterad info kommer från källa
Börsnoterad	Börslistor på www	5 ggr/ vecka	
Serveringstillstånd Kommuner	Folkhälsomyndigheten	5 ggr/ vecka	När uppdaterad info kommer från källa
Certifieringar bl.a. ISO- och EMAS	Ackrediterade utfärdade	1 ggr/ månad	
Tullkrediter	Tullverket	5 ggr /vecka	Gallras efter 36 mån
Kunglig Hovleverantör	Hovlev.com	Löpande	När uppdaterad info kommer från källa
Yrkestrafiktillstånd Länsstyrelser	Transportstyrelsen	2 ggr/ månad	När uppdaterad info kommer från källa
Miljösanktionsavgift	Kammarkollegiet	Varje månad	När uppdaterad info kommer från källa
Förkommen ID-handling	Förlustanmälan	Löpande	Gallras efter 24 mån
Ställda frågor och frågeställare	UC	Löpande	Gallras efter 12 mån
Sökt/prövad kredit	UCs kunder	Löpande	
Handels- och kommanditbolag; Firma, adress m.m.	Bolagsverket och SCB	5 ggr/ vecka	Historisk firma gallras ej Övrig information gallras när uppdaterad info kommer från källan
Handels- och kommanditbolag; Nya och avregistrerade, firmatecknare, delägare, prokurist, parallell- och bifirma	Bolagsverket	7 ggr/ vecka	Delägare gallras 60 månader efter avregistrering. Övrig information gallras när uppdaterad info kommer från källan
Andra företagsformer	SCB BASUN	Varje vecka	
Arbetsställen	SCB	Varje vecka	När uppdaterad info kommer från källa
Juridisk person/enskild firma, namn och adresser	SCB	Varje vecka	Firman gallras inte Adress gallras när uppdaterad info kommer från källa
Juridisk person/enskild firma; namn och adresser	Bolagsverket	5 ggr/ vecka	Firman gallras inte Adress gallras när uppdaterad info kommer från källa
Juridisk person/enskild firma	Telefon SCB	Varje vecka	När uppdaterad info kommer från källa
Näringsgrenstillhörighet (SNI)	SCB	Varje vecka	När uppdaterad info kommer från källa
Branschbevakning	SKI	Var 3e månad	
Företagsrekonstruktion och rekonstruktör	Kronofogden och tingsrätt, PoIT	5 ggr/ vecka	Gallras efter 60 månader
Enskilda näringsidkare	SCB BASUN	Varje vecka	
Reg. för moms	Skatteverket	5 ggr/ vecka	Gallras ej
Reg. för F-skatt	Skatteverket	5 ggr/ vecka	Gallras ej
Reg. för arbetsgivaravgift	Skatteverket	5 ggr/ vecka	Gallras ej
Företagsinteckningar	Bolagsverket	5 ggr/ vecka	När uppdaterad info kommer från källa
Konkursborgenärer (oprioriterade konkursfordringar)	Konkursförvaltare/ tillsynsmyndighet	5 ggr/ vecka	Efter 24 månader
Konkurser och konkursförvaltare	Kronofogden och tingsrätt via PoIT	5 ggr/ vecka	

Kreditengagemangsuppgifter (beviljade lån/ krediter)	Bank, kreditmarknadsbolag, kontokortsföretag	Varje månad / dagligen	Efter 12 mån
Anmärkningar; restförda skatter och avgifter	Kronofogden	5 ggr/ vecka	Efter 36 mån
Anmärkningar; tredskodomar	Tingsrätt	5 ggr/ vecka	Efter 36 mån
Anmärkningar; konkursansökningar	Tingsrätt och Kronofogden	5 ggr/ vecka	Efter 36 mån
Anmärkningar; betalningsföreläggande n	Kronofogden	Varje vecka	Efter 36 mån
Anmärkningar; utmätningsförsök/återtaget gods	Kronofogden	5 ggr/ vecka	Efter 36 mån
Missbrukade bankkonton	Banker	5 ggr/ vecka	Efter 36 mån
Misskötta krediter och hypotekslån	Banker och finansbolag	5 ggr/ vecka	Efter 36 mån
Missbrukade kontokortskrediter	Banker, finansbolag och kontokortsföretag	5 ggr/ vecka	Efter 24 mån
Näringsförbud Tingsrätten	Bolagsverket	5 ggr/ vecka	Efter 36 mån (efter upphörande)
Skuldsaldo	Kronofogden	Varje vecka	Efter 24 mån, Aktuellt skuldsaldo visas löpande
Skuldsanering	Kronofogden	5 ggr/ vecka	Inledd skuldsanering gallras efter 36 mån.
UC Riskklass	UCs egna beräkningar	Dagligen	Efter 24 månader
UC Riskprognos	UCs egna beräkningar	Dagligen	Efter 24 månader
UC Betalmönster	Samarbetande företag	Löpande	
Bokslutsrapport och Redovisningskonsult	Årsredovisning	5 ggr/vecka	

Appendix 2 - List of variables

Variable	Description
Gender	Gender of borrower (0 = male, 1 = female)
Age	Age of borrower at time of loan application. Based on personal identification number.
YearlyIncome	Monthy income of borrower at the time of loan application. Stated by borrower in loan application.
MaritalStatus	Family status of borrower at the time of loan application. Stated by borrower in loan application.
AppliedAmount	The loan amount of a granted loan.
Maturity	The duration of the loan, stated in months
APR	Interest rate (annual percentage rate) of the loan. The interest rate is constant during the enitre maturity of the loan.
UCScore	UC Risk Score, defined as "probability (%) of delinquency within 12 months" according to UC
Purpose	Purpose of the loan, stated by borrower when applying for loan
Ever_30DPD	Binary variable stating whether a specific loan has ever had two or more delinquent payments at the same time (30 days)
Ever_60DPD	Binary variable stating whether a specific loan has ever had three or more delinquent payments at the same time (60 days)
Ever_90DPD	Binary variable stating whether a specific loan has ever had four or more delinquent payments at the same time (90 days)
NetMonthlyIncome	The net monthly income (SEK) of a borrower at time of application

	according to customer. Calculated using the average Swedish tax rate, 32.12 %, and taking into consideration the higher tax rates for the higher income brackets, using 2018 bracket levels and rates (Skatteverket)
DummyGamb	Binary variable that takes the value 1 if the borrower has made at least one gambling transaction in the available bank statement data, and 0 if not
DummyColl	Binary variable that takes the value 1 if the borrower has made at least one collection transaction in the available bank statement data, and 0 if not
IncOverstate	Binary variable that takes the value 1 if the borrower has overstated his/her income by at least 10%, and 0 if not

Appendix 3 - Variables unsuccessfully used to attempt building a regression model

Variable	Description
No_Months	The number of months of available bank statement data.
SumGambMonth	The net sum of gambling transactions per month of available bank statement data
FreqGambMonth	Number of gambling transactions per month of available bank statement data (both deposits and withdrawals)
AvgGamb	Average amount per gambling transaction
SumCollMonth	Net sum of collection transactions per month of available bank statement data
FreqCollMonth	Number of collection transactions per month of available bank statement data (positive and negative)
AvgColl	Average amount per collection transaction
BS_SalaryMonth	Monthly net income from salary per month according to the bank statement
BS_FkassaMonth	Monthly net income from government support per month according to the bank statement
BS_IncomeMonth	Total monthly net income per month according to the bank statement
MonthlyIncomeRatio	The ratio between NetMonthlyIncome and BS_Income (NetMonthlyIncome / BS_Income * 100). A value of 100 means they are identical and a value of 120 means the bank

	statement income is 20% greater than the stated income.
GambPerIncomeMonth	The share of net income spent on net gambling. A value of 0,1 means a borrowers spends 10% of his/her income on gambling each month
CollPerIncomeMonth	The share of net income spent on collection. A value of 0,1 means a borrowers spends 10% of his/her income on collection payments each month
DaysSinceGamb	Days between last gambling transaction and loan payout date
DaysSinceColl	Days between last collection payment and loan payout date
GambLossSumMonth	Sum of **negative** gambling transactions per month of available bank statement data (i.e. only counting losses, not withdrawals)
GambLossFreqMonth	Number of **negative** gambling transactions per month of available bank statement data (i.e. only counting losses, not withdrawals)
GambLossAvg	Average amount per **negative** gambling transaction (i.e. only losses)
Last_gamb	Days between last gambling transaction and loan application, to see whether recency has any effect
Last_coll	Days between last collection transaction and loan application, to see whether recency has any effect