DEFAULT RISK WITHIN PEER-TO-PEER CREDIT MARKETS

ASSESSING THE FACTORS AFFECTING BORROWER'S RISK OF DEFAULT

LUCAS RINTALA

MARCUS BRAUN

Bachelor Thesis

Stockholm School of Economics

2020

Default risk within peer-to-peer credit markets

Abstract:

Assessing credit risk is essential for all companies operating on the credit market. Since current literature examining credit risk rarely focus on the peer-to-peer market, this paper examines default risk within the peer-to-peer lending market to contribute to the knowledge of credit risk for investors and peer-to-peer lending platforms. Utilising regressions, hazard models and ROC-curves, this paper has found credit rating to be a better explanatory variable for default than the interest rate. New credit customers have a higher default risk than existing credit customers. Female borrowers perform better than male borrowers. Loans taken for real estate purposes perform the best, whereas loans for loan consolidation perform the worst. Finally, we suggest how key variables can be used by lending marketplaces to focus on the customers with a lower risk of default and how these marketplaces can improve their relationship with investors.

Keywords:

Default risk; risk factors; peer-to-peer lending; credit markets; consumer loan crowdsourcing

Authors:

Lucas Rintala (24264)
Marcus Braun (24035)

Supervisor:

Marcus M. Opp, Associate Professor, Department of Finance

Examiner:

Adrien d'Avernas, Assistant Professor, Department of Finance

Finalised in May 2020.

Bachelor Thesis
Bachelor Program in Business and Economics (enrolment year 2017)
Specialisation in finance, specialisation in management
Stockholm School of Economics

1. Introduction

1.1. Background

Assessing credit risk is essential for all companies operating on the credit market. Understanding the factors determining default risk is a difficult task in the screening process of borrowers since there is a wide stream of information that could influence the risk of default. In this paper, using real data from a peer-to-peer company, several variables are considered when analysing what factors contributes to the risk of a credit default. Loan default cannot be perfectly predicted, but with insight to and understanding of which variables are the most influential, the results of this study can suggest which variables should have a larger weight in the risk predicting models. Analysing the data, it is self-evident that two identical borrowers might show different risk and behave differently in their repayment behaviour. However, this should not be seen as a setback but rather the potential for further research tying credit risk models with decision theory and behavioural finance.

In peer-to-peer lending, in contrast to regular banking, it is not only the lending company that is interested in assessing the risk of the borrower, but also the customer on the other side of the transaction, namely the lender; the investor. This stakeholder does not always have formal training, nor the technical skills to assess the credit risk of the borrowers of their invested capital. The investors are attracted by a well-diversified investment and high expected rates of return. However, this is only true if the borrows they lend their capital to do not default. In the data, interest rates upwards 270 per cent can be observed, but if default risk is assessed incorrectly a high-interest rate can still result in a non-existent return on investment. Best case scenario, investors run regression models themselves, but most probably they use the auto-invest functions without much consideration of the default risk of the borrower. The reputation and livelihood of peer-to-peer lending platforms are that the investors keep funding their loans, which will only happen if they can show positive returns that in one way or another outperforms the stock market, either in return or stability. This is discussed further in *Discussion, managerial implication and suggested further research for business and scholars.*

The peer-to-peer lending platform which forms the basis of this paper is a beneficial marketplace to study since it serves borrowers and investors (lenders) in several countries in Europe. Having access to data from a wide range of borrowers gives us the possibility to draw broader conclusions of our findings and gives weight to our implications. The fact that the site that the data sample is drawn from is being used by a wide range of investors and borrowers with different knowledge and backgrounds but with access to the same financial information also provides a similarity to the global setting of lending markets that this paper wishes to contribute to.

From the data used for this thesis, it can be observed that the borrowers, contrary to the general population, are skewed towards the lower credit ratings (Figure 1 cf. Keys, Mukherjee, Seru and Vig 2010). A similar pattern has been found by Iyer, Khwaja, Luttmer and Shue (2016) on an American peer-to-peer platform (Prosper), indicating that this pattern is a common feature within the peer-to-peer industry and not just in the dataset used in this paper. It is not farfetched to assume that in the future, new peer-to-peer network markets, and therein new peer-to-peer credit markets will surface and grow, it is therefore of great importance to contribute to the understanding of default risk among the sort of customers drawn to these platforms. The sample should, therefore, be representative of the peer-to-peer credit market and not the vast population, which is the case.

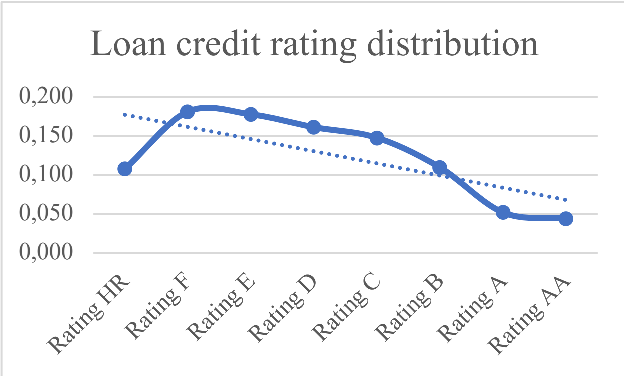


Figure 1. Notes. The graph has a positive skew indicating that lower-rated credit customers to a greater degree either apply for loans or get their loans accepted by Bondora. The data points show the fraction of the loans between 2009 and 2019 in the respective credit rating category. 2.1 per cent of the loans are missing a credit rating and consequently, the fractions in the graph do not sum to one.

1.2. Findings and contribution

Our results find interest rate to be an important variable, with higher interest rates increasing default risks. This is however very understandable since the interest rate is set after an assessment of the borrower has been made. This correlation is, therefore, proof of proper allocation of interest rates. Similar results, and the same explanation, are also found when the credit rating worsens, with a lower credit rating resulting in higher default risks. In contrast to the results found by Iyer et al. (2016), our results show that the area under the receiver operating characteristic (ROC) curve is higher for credit rating than the interest rate, although, one has to note that their results are based on the interest rate and credit score instead. The adjusted R^2 results from the regression with credit ratings is also higher than that of interest rate, contrary to Iyer et al. (2016) findings. Estonia shows the lowest default risks, with a much higher default risk for Finland, and the highest for Spain. The most recently issued loans to Spanish borrowers indicate an increase in loan performance while loans issued in Estonia and Finland during the same period indicate a declining quality of loan performance. The overall loan performance has consequently gone down.

We find that new credit customers have a much higher default risk than existing customers, which is a reasonable finding as an existing customer most likely would not receive a new loan if he or she did not fulfil their obligations on a previous loan.

The data indicate that a longer loan duration leads to a higher risk of default. However, our data also suggest a strong connection between loan duration and credit rating. AA has the shortest average loan duration and F the longest (HR does not follow this trend). It is therefore difficult to tell what the cause is and what is the effect.

Regarding the purpose of loans, we find that loan consolidation has the highest default risk amongst significant variables. This is an interesting contrast to the finding that a higher number of existing liabilities decrease the risk of default. Next, we find that using the loan to purchase real estate has the lowest default risk. This is consistent with the finding that not being a homeowner increase the risk of default.
Not being fully employed, increases the risk of default. This seems reasonable as their income would be less volatile. However, looking at data of matured loans, the fully

employed borrowers have the lowest *total income* on average of all *employment status* categories that are working. On average, fully employed workers earn €767 less than entrepreneurs. With regards to this pattern, the risk of default is assumed to be better explained by a stable income than a high income. Next, we find that higher education lowers the risk of default. Our hypothesis of this was that there would be a connection between high education and high income, however, this could not be seen in the data.

Another finding we could draw from the data is that the work experience of the borrower explains very little in terms of default risks. The same is true for the number of dependents the borrower have in addition to the age of the borrower.

We find that the higher applied amount a borrower wishes to borrow leads to higher default risk. However, the higher actual amount the borrower is able to borrow, the lower the default risk. This is reasonable as a safer borrower would likely be able to borrow a higher amount. However, we do not find a connection between a higher credit rating and larger issued amounts. The data instead suggest that loans with a rating of AA are issued the lowest amounts. When looking at the difference between the applied amount and actual (received) amount the trend is very clear. The better the rating you have, the more likely it is that you will receive the amount that you have applied for. This is seen in the data ranging from less than a €10 difference for AA loans to over a €435 difference for HR loans.

This paper contributes to research and literature aiming towards assessing credit risk and predict default risk but also to businesses, especially within peer-to-peer markets that tend to consist of private, high-risk borrowers and investors without the possibility of assessing the risk themselves. Tang (2019) have looked into the field of peer-to-peer lending, analysing if peer-to-peer marketplaces serve as substitutes or complements to regular banks. Even though this paper does not aim towards answering a similar question, and the findings of risk-factors are not being compared to those of customers of regular banks, a similar study might be conducted in the future with this paper as a starting point. This paper contributes not only to the understanding of risk but also to the understanding of peer-to-peer businesses.

Our paper also contributes to the literature that examines how gender affects the probability of bankruptcy, which Agarwal, He, Sing and Zhang (2016), previously has researched. Consistent with their results we find that females show a lower default probability than men on their loans, although their findings show a larger difference between men and women in terms of risk, finding that the odds of women being involved in bankruptcy events are 28% of those for men.

Our paper complements the literature examining default risks within peer-to-peer lending and specifically examining the predictive power, the goodness-of-fit (R^2) and the effect loan data variables have on default risk, similar to what Iyer et al. (2016) and Duarte, Siegel and Young (2012) previously have looked at. We contribute to these findings by adding more variables and data from a new market.

In this paper, data has only been retrieved from one marketplace, and thereof, only a limited set of loans are examined. The authors believe that the information displayed, and the conclusions made can be used in broader terms in both academia and business. By showing what factors are important to assess risk in these settings the findings can help both future peer-to-peer markets and the research of this industry.

2. Context and Data

2.1. Context

Since the beginning of time, humans have borrowed valuables from each other, and since the early 2000s, this has been done on online peer-to-peer lending platforms. On these platforms, potential borrowers and investors are paired up to receive unsecured loans and returns from interest payments, respectively. Around the world, the peer-to-peer segment of business is on the rise, including businesses in industries such as education, healthcare, communication and information sharing, IT-security and the foci of this paper: lending. In this paper, data from the Estonian peer-to-peer lending platform *Bondora.com* is analysed. Founded in 2007, the company has over 10-years of history of connecting peers in Europe, making Bondora the world's first cross-border lending marketplace (see Appendix A). Over the years, more than 117,000 people have invested over €368 million with total earnings of €42 million.

Bondora currently only offers personal, unsecured loans. The borrowers are from Estonia, Finland, Spain and Slovakia. Up until 2012 they also arranged a small number of loans to businesses. Bondora's loans are private, fixed-rate loans with loan durations stretching between one and 60 months. All loans are currently issued in *euro* (EUR). Loans issued before Estonia joined the eurozone (1 January 2011) and still used *kroon* (EEK) has all been converted to euro by Bondora themselves before the dataset was retrieved. Hence, some loans have very specific amounts. Borrowers can apply for loans of up to €10,000 on the local version of the site (e.g. *Bondora.ee)* and are thereafter funded by bidding. The bidding process works in three ways, by investors manually handpicking loans; using the Bondora API (requiring advanced programming skills by the lender) or using the semi-automated portfolio manager provided by Bondora. Worth mentioning is that not all information needs to be provided by the borrower, one example is the use of the loan. Using the information provided by the applicant and data from third parties, such as credit bureaus, population registries, banks and tax authorities (see Appendix B), each applicant is assigned to one of eight credit ratings: AA, A, B, C, D, E, F and HR (high-risk, the lowest rating). After the loan is funded, it is repaid with interest each month. Should the borrower not pay back and a collection process starts, this is defined as a default. A loan can reach the default state before, at, or after the end of the loan duration (see information on *B Secure* later). When the loan has defaulted and is turned to a civil claim in court or a bailiff, it will not change its status from default. In the following analyses, *Default date* is used as the dependent variable. This is because default is a definite ending and not just a temporary payment issue that might be resolved. This allows for easier handling of data and broader conclusions.

2.2. Data

The data set used to perform the analyses displayed in this paper contains over 99.9% of all loans issued between 2009 and 2019, the few loans not displayed are covered by data-protection laws and therefore unavailable to the public. There is no reason to assume that these loans would dramatically differ from the average. Loans issued in 2020 have been excluded (see *Methodology* for reasoning), consequently, the data set consists of 130,124 loans between 2009-02-28 and 2019-12-31. Of these loans, 26,147 have passed their loan duration (as of 2020-04-09, which was the last time data was retrieved) and are therefore considered matured. The variables included in the dataset range from technical variables such as credit rating and interest rate, to softer ones such as marital status and employment position. The variables useful for this study together with summarised statistics are presented in Table 1. The matured loans are both included in all loans and presented separately. The listed loans are unevenly distributed with more loans issued every year. When looking at all loans the most common year for a loan to be issued is 2019 (last year included) and the most common loan duration is 60 months. However, when looking at all matured loans the most common year is 2014 and most common loan duration 36 months. When analysing only the matured loans this leads to an analysis of shorter and earlier issued loans than the mean. This should allow for a truer image of the risk of default since the default rate of all matured loans is 51.5% (as of 2020-04-09) in comparison to the lower default rate of 37% when looking at all loans from 2009 to 2019. Obviously, a short-term loan issued early in the career of Bondora has had more time to default than a long-term loan issued in recent years. Thus, matured loans should, therefore, be seen to better represent reality.

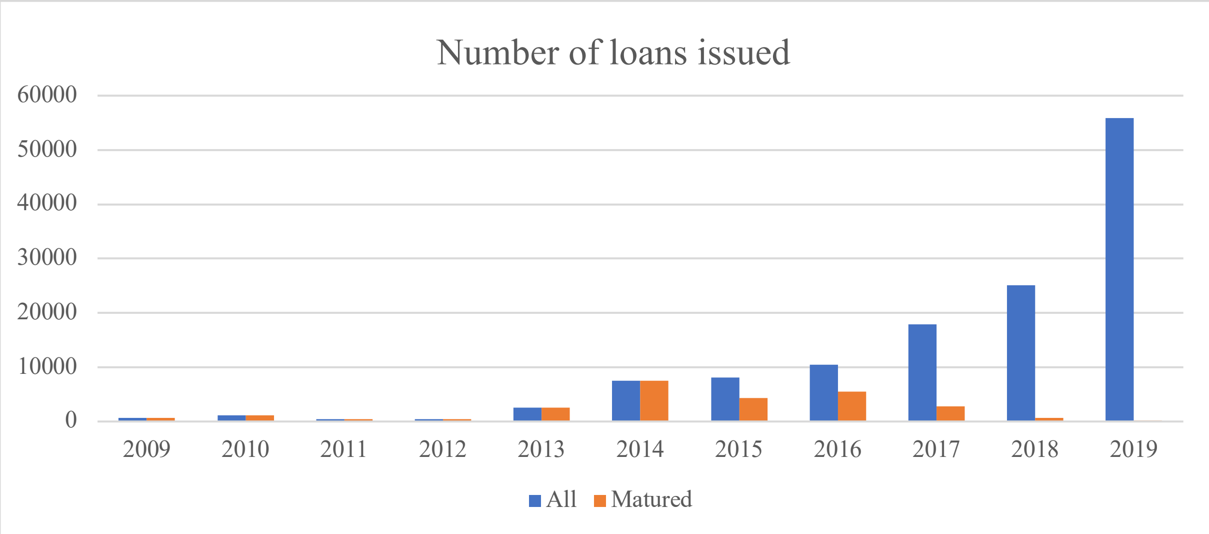


Figure 2. Notes. The number of loans issued per year is uneven with more loans issued every year. All loans issued until and including 2014 have matured. For 2015 and 2016 about half of the issued loans have matured. From 2017 a very small share of the loans has matured.

Table 1 - Summary statistics

	All loans (2009 - 2019)		**Matured loans (as of 2020-04-09)**	
	Mean/fraction	S.D.	Mean/fraction	S.D.
Annual interest rate	0.360	26.509	0.348	26.228
Rating HR	0.768	50.844	0.587	
Rating F	0.553	12.287	0.395	
Rating E	0.358	4.104	0.321	
Rating D	0.287	4.491	0.283	
Rating C	0.218	3.946	0.247	
Rating B	0.161	3.848	0.204	
Rating A	0.131	4.384	0.191	
Rating AA	0.102	3.473	0.167	
Rating not available	0.294	8.360	0.294	
Default percentage	0.370		0.515	
Rating HR			0.573	
Rating F			0.569	
Rating E			0.556	

Rating D			0.528	
Rating C			0.515	
Rating B			0.499	
Rating A			0.457	
Rating AA			0.501	
Rating not available			0.336	
Variables				
Rating distribution				
Rating HR	0.107		0.238	
Rating F	0.181		0.082	
Rating E	0.178		0.096	
Rating D	0.161		0.142	
Rating C	0.147		0.158	
Rating B	0.109		0.110	
Rating A	0.052		0.048	
Rating AA	0.044		0.022	
Rating not available	0.021		0.104	
LoanDateYear				
2009	0.005		0.025	
2010	0.009		0.044	
2011	0.003		0.017	
2012	0.003		0.017	
2013	0.019		0.096	
2014	0.057		0.285	
2015	0.062		0.165	
2016	0.081		0.211	
2017	0.138		0.106	
2018	0.193		0.026	
2019	0.429		0.007	
Country				
EE	0.578		0.634	
ES	0.179		0.201	
FI	0.241		0.154	
SK	0.002		0.011	
Amount requested (€)	2725	2388	2336	2138
Amount received (€)	2534	2177	2181	1972
Age	41	12	37	11
Existing liabilities	3.222	3.444	4.191	3.490
Total monthly liabilities (€)	571	34399	732	1313
Free cash	126	704	427	458
Gender				
Male	0.643		0.553	
Female	0.280		0.392	
Number of dependents	0.723	1.003	0.723	1.003
Loan duration	46	15	32	18
12 months	0.029		0.128	
18 months	0.019		0.071	
24 months	0.039		0.143	
36 months	0.336		0.275	
48 months	0.072		0.076	
60 months	0.477		0.197	
Education				
Primary education	0.087		0.014	
Basic education	0.049		0.118	
Vocational education	0.217		0.183	
Secondary education	0.383		0.416	
Higher education	0.264		0.266	
Not available	0.000		0.002	
Marital status				
Relationship	0.155		0.517	

Not-Relationship	0.126	0.421
Not available	0.719	0.062
Fraction new credit customers	0.578	0.712
Work experience		
LessThan2Years	0.016	0.061
2To5Years	0.039	0.139
5To10Years	0.060	0.205
10To15Years	0.053	0.173
15To25Years	0.061	0.196
MoreThan25Years	0.051	0.164
Not available	0.719	0.062
Employment duration at the current employer		
UpTo1Year	0.180	0.192
UpTo2Years	0.050	0.132
UpTo3Years	0.042	0.110
UpTo4Years	0.028	0.075
UpTo5Years	0.209	0.082
MoreThan5Years	0.387	0.361
Trial period	0.006	0.013
Retiree	0.055	0.001
Other	0.036	0.001
Not available	0.007	0.033
Occupation area		
Other	0.065	0.217
Mining	0.001	0.003
Processing	0.025	0.086
Energy	0.005	0.014
Utilities	0.003	0.010
Construction	0.025	0.082
Retail and wholesale	0.028	0.091
Transport and warehousing	0.019	0.062
Hospitality and catering	0.017	0.055
Info and telecom	0.015	0.052
Finance and insurance	0.009	0.031
Real estate	0.004	0.011
Research	0.004	0.013
Administrative	0.006	0.022
Civil service & military	0.013	0.046
Education	0.011	0.038
Healthcare and social help	0.019	0.059
Art and entertainment	0.005	0.016
Agriculture, forestry and fishing	0.008	0.025
Not available	0.720	0.066
Fraction homeowners	0.498	0.465
Use of loan		
Loan consolidation	0.053	0.198
Real estate	0.007	0.023
Home improvement	0.071	0.222
Business	0.015	0.046
Education	0.010	0.034
Travel	0.014	0.047
Vehicle	0.024	0.086
Health	0.012	0.038
Other	0.075	0.247
Not available	0.719	0.060
Employment status		
Unemployed	0.000	0.000
Partially employed	0.009	0.032
Fully employed	0.231	0.776

Self-employed	0.010		0.033	
Entrepreneur	0.015		0.048	
Retiree	0.014		0.043	
Not Available	0.721		0.069	
Total monthly income (€)	1727	5042	2074	3579
0	0.000		0.002	
1-1000	0.327		0.384	
1000-2000	0.441		0.395	
2000-3000	0.157		0.118	
3000-4000	0.038		0.024	
4000-5000	0.012		0.008	
5000+	0.024		0.070	
Observations	130124		26147	

Notes. For Marital status, a relationship is defined as being either married or cohabitant while not being in a relationship means you are either single, divorced or a widow. A homeowner is classified as either being a sole owner, but it can also be joint ownership, ownership through a mortgage or being an owner with encumbrance. If you are a tenant, living with your parents or living in a council house, this does not count as homeownership. For some of the variables, the fractions do not sum to one, this is either because there are more alternatives than the highlighted in this table or because there are more decimals than shown in the table. When looking at the number of dependents, only numerical values are considered. A small number of borrowers have stated "10+", these have been left out when calculating the average number of dependents. *Use of loan: other* includes the few loans given to businesses before 2012, which was mentioned earlier (these loans are not used in any further analysis concerning the research question). For age, only borrowers with an age of 18 or above have been included.

It should be mentioned that the distribution between the credit rating categories is also uneven. Of all loans between 2009 and 2019, only 35 per cent were rated in the upper half of the rating categories, C or higher, while nearly 63 per cent were rated in the lower half, D or lower. The remaining two per cent does not have a rating available. This paper intends to bring understanding to which variables contribute to the default risk, not only for the customers of Bondora but for the general population as well. When broader conclusions are made, this should be kept in mind.

The variable *Occupation Area* will not be used in the regressions since it does not provide any new information and it is divided into far too many categories. Attempts of recoding the variable have been made but since essential information, such as if, the worker is white- or blue-collar, is missing the variable has been left out of any analysis. The similar variable, *employment status,* is, unfortunately, an uncoded string in foreign languages that have proven difficult to make sense of. The variable *Employment Duration at Current Employer* is also disregarded since it is difficult to get any real meaning from it. It does not display a measure of how loyal you are to an employer in terms of how often you change employer; someone with a short time with the current employer might have the intention of staying there for a long time and vice versa. Instead, *work experience* is looked at to determine a relationship between total time with the employer(s) and the default risk.

Certain loans have a *Contract end date* later than *Loan Date* followed by *Loan Duration*. This also leads to some loans defaulting after the loan duration has ended. This is explained by the fact that borrowers can change their loan schedule – with a system called *B Secure* (see Appendix A). The borrower might change loan duration, monthly payment date and/or take a "principal payment holiday" (still having to pay interest and commission). Making a payment intermission can be one explanation of why the total time of the contract is longer than the loan duration. The authors do not have insight into specific loans as to why they default after the loan duration.

Apart from missing values, some data does not add up. As an example, the borrower with loan number 4057 is fully employed, has worked at their current employer for up to four years, has a work experience of over 25 years but a total income of € 0. The most probable explanation is that the value for the salary is missing but for some reason shows as a "0" and not an empty cell. As a result, the mean total income displayed in Table 1 probably displays a lower value than the true average. However, all "0" could not be removed since some are probably truly missing a total income. When in contact with Bondora they cannot explain this feature but say: "the info[rmation] they [the borrowers] see in the UI [user interface] might not reflect what is actually taken in the calculation". In short, when it comes to this variable, and all others where the source is the borrowers themselves there is not much to do but to trust that the borrowers are being honest and give the true information. This will not be discussed further in this paper since it has been found e.g. by Abeler, Nosenzo and Raymond (2019) that people lie surprisingly little.

The data also, incorrectly shows that some borrowers are under the age of 18 years (see Appendix B). When the mean for age has been calculated only borrowers with a stated age of 18 or above have been included.

3. Methodology

By running sets of simple and multiple regressions, each with different variables, we wish to find which variables affect the default probability of loans issued by Bondora. In addition to these regressions, the Cox Proportional Hazard model, ROC curves and Kaplan-Meier survival estimates will also be used to interpret our data. For the regressions, the quality of a certain regression variable or screening method will be measured based on the R^2 results from the regressions, in addition to ROC curves with area under the curve (AUC) measures, similar to how Iyer et al. (2016) reasoned. The coefficients of the regressions will also be taken into consideration, as long as the coefficients are significant. Iyer et al. (2016) also state that the AUC is "the most common metric used in the credit-scoring industry". Although Bondora does not offer any data on exact credit scores, it is reasonable to expect the AUC to still be useful when looking at credit ratings, which Bondora offers data on. This is because a test using a receiver operating characteristic (ROC) curve can be used when it "is based on an observed variable that lies on a continuous or graded scale" (DeLong, DeLong & Clarke-Pearson, 1988).

A time-fixed variable has not been included in the regressions, mainly since the data did not have any specific year with a large difference in default rates compared to the yearly average. Also, a time-fixed variable would have been most useful for analyses where time is important such as the Cox Proportional Hazard Model and the Kaplan-Meier survival estimates, however as only loans with a 36 months loan duration were included for these models a time-fixed variable was not necessary.

Although a time-fixed variable would mostly have been useful for the above-mentioned ones, it could also potentially have been used in the simple and multiple regressions as well. However, as only matured loans were included in these regressions, it was infeasible to include.

We also did not include any loans from Slovakia, as there are less than 300 loans from this country, making the possibility for errors large. In addition, we also chose to exclude loans with a gender of neither male nor female, as we were interested in interpreting potential gender differences and since the reasons for not stating a true gender might vary amongst borrowers.

3.1. Simple and Multiple Regressions

The reason to conduct simple and multiple regressions was to effectively get an understanding of which variables affect the default risk to the largest extent for the loans in the data set. It is also interesting to make multiple regressions in order to find how several variables would behave in the same regression.

For our regression, we selected the variables from the dataset that we found relevant. Using variables similar to the ones used by Iyer et al. (2016), was one criterion since this would make the comparison between our results and theirs more relevant. Agarwal et al. (2016) analysed the effect of gender on delinquency rates, making it interesting for us to also look at gender as a variable and compare our results with their paper. Some of the variables we did not use were removed due to missing data or that no substantial conclusions could be made. For variables that consisted of several categories, such as credit rating, dummy variables were created for each rating in order to see the difference between each category.

Regressions were mainly made, either with one variable, such as age or interest rate, or with variables consisting of several dummy variables, such as credit rating. For some regressions, however, several variables were included, in order to see how they would behave in the same regression. For some regressions that were made, high Variance Inflation Factor (VIF) values were obtained for certain variables. This was a problem in some cases as it could indicate multicollinearity, however in other cases the variables that had high VIF values were dummy variables with three or more categories, making it possible to safely ignore the high VIF values. (Allison, 2012)

One regression we were interested in conducting, which consisted of several variables, was a finance-oriented regression, consisting of the variables: *applied amount*, *actual amount*, *existing liabilities*, *free cash*, *income total*, *loan duration* and *total liabilities*. When making this regression, however, the VIF values for *applied amount*, the *actual amount* received and the *total income*, all were very high, above 8. Even though this is below a common limit of 10 (Hair, Black, Babin and Anderson 2014) *applied amount*, the *actual amount* and *income total* were removed from the regression, as the risk for multicollinearity otherwise would have been substantial. Thus, the final version of the financial model included the variables *existing liabilities*, *free cash*, *loan duration* and *total liabilities*.

Another interesting regression, consisting of several variables was a regression with *actual amount* and *applied amount* as variables, as it would be interesting to see how these would behave together. However, as the VIF values for this regression was very high, for both variables, it was difficult to say too much about the results. A final regression we were interested in was one with all our selected variables. For this one, we had problems with high VIF values for *applied amount* and *amount*, thus this regression ended up including all variables we were interested in except these two. When making this regression without *applied amount* and *amount*, the significance of several variables overall was reduced, compared to the significance of these variables when making regressions for the variables separately. Thus, we found our other regressions to be more useful, choosing not to include this large regression in our final paper.

For some of the variables in our regressions: *applied amount*, *amount*, *free cash*, *total income* and *total liabilities*, these were in euros, with a change in these variables representing a one-euro difference, which makes their coefficients very small. Because of this, these coefficients were multiplied with 100 to make the coefficients easier interpretable. The constants were, however, not multiplied.

3.2. Cox Proportional Hazards Model

For the Cox Proportional Hazards Model, we constructed the variable *Cox Time Variable*. It is displaying the survival time for each loan. For a loan that has not defaulted, it shows the *loan duration*, and for a loan that has defaulted, it shows the *time until default*. If the loan has not matured nor defaulted it shows the time until the data was retrieved (2020-04-09). *Time until the default* is constructed by subtracting the *Loan Date* from the *Default Date*. For all conversions between days and months, all months are assumed to be 30.44 days long (365.25/12 with two decimals).

The reason for using the Cox Proportional Hazard model is its effective way of looking at hazards, in our case default, where the outcome is binary, either a loan has defaulted, otherwise it has not, in addition to the model looking at the time until failure (default). Also, previous literature such as Gross and Souleles (2002) used hazard functions, when looking at the probabilities of delinquencies. Hazard rates were calculated for the same variables as the ones analysed in the simple and multiple regressions to get a better understanding of our results, by using two different methods. Regarding hazard rates, a hazard rate of less than one implies a decrease in risk from increasing a variable. For example, the hazard rate for the variable *age* was slightly below one, meaning that a higher age slightly decreases default risk. A hazard rate above one for a variable instead indicates a higher default risk. In the same way, as some variable coefficients were multiplied for the simple and multiple regressions, this was also done for the hazard rate results, displaying and multiplying the hazard coefficient for the selected variables with 100, to better interpret the results. The hazard rates, however, were not modified. Also, to complement the hazard rates, Kaplan-Meier survival estimates graphs were used, as these give a good display of survival data, especially when comparing two groups, such as a treatment and a control group. They include the survival fraction for loans, starting at a 100% survival rate when loan duration is zero months and ending at a 0% survival rate after all loans have either matured or defaulted. Two different Kaplan-Meier survival estimates graphs were made, one comparing the survival estimates for countries and whether the loan was matured- or not matured and one comparing whether a credit customer was new or not and whether the loan was matured- or not matured. The comparison between matured- and non-matured loans was simply a case of having a control group (non-matured loans) and a treatment group (matured loans) to control for differences. It is reasonable to expect that these graphs would be the same, if Bondora's loan performance is constant and does not change, in addition to no substantial macroeconomic factors affecting certain years.

Regarding substantial macroeconomic factors, both 2009 and 2020, which are included in our loan data could be counted as years with substantial macroeconomic effects, 2009 due to the financial crisis and 2020 due to the Coronavirus pandemic, however, none of these years are included in our final data set used. Thus, given that Bondora's loan performance has not changed, it is reasonable to expect that both graphs would look the same.

As the data set available from Bondora, unfortunately, do not include any exact credit scores, *credit rating* was instead used as a comparison to interest rates to find how their area under ROC curve compares

3.3. ROC Curves

The reason for using ROC Curves is because of its common usage in commercial financial banking markets (Iyer et al. 2016). As previously said, Iyer et al. (2016) also state that the AUC is "the most common metric used in the credit-scoring industry", which is why the AUC is examined.

As mentioned earlier, ROC curves can be used when the observed variable either lies on a continuous or graded scale (DeLong et al., 1988). As we did not have access to exact credit scores in the data set available from Bondora, we did not have a variable lying on a continuous scale, however by using credit rating, this is on a graded scale, and thus still possible to use for ROC curves. In our ROC curves, credit rating was compared with interest rate to find how their area under the ROC curve compares.
Because we did not have access to credit scores, we were not able to make a robustness check such as the one performed by Iyer et al. (2016), as they can examine the AUC within each credit category, comparing that to the AUC of the credit

score, however, we did not have access to an exact credit score.

The ROC curve itself illustrates a trade-off between the true positive rate (TPR) and the false-positive rate (FPR), where the TPR represents the probability of correctly rejecting bad borrowers and FPR represents the probability of mistakenly rejecting good borrowers. When both the TPR and the FPR equal zero (0%), no borrowers are rejected, however when both the TPR and the FPR equal one (100%), everyone is rejected. In order to measure the accuracy associated with a ROC curve, the AUC will be used. The greater the AUC for ROC curves, the better predictors for default. In an information-scarce environment, a desirable AUC is generally an AUC of 0.6 or greater. For more information-rich environments 0.7 or greater is instead considered to be desirable. As it is difficult to estimate what level to be used, an arbitrary level of 0.65, in between 0.6 and 0.7 will be used for the AUC graphs where all credit ratings are included, as the dataset used for these graphs include the highest amount of observations. For the AUC graphs where only high credit ratings or low credit ratings are included, the information-scarce level of 0.6 will instead be used, because of the lower number of observations used when calculating these graphs. (Iyer et al., 2016)

4. Results

As can be seen in regression (1) from Table 2, higher interest rates increase the risk of default, with significance at the p<0.01 level. Although, the coefficient is quite small (0.008), one has to note that interest rates on Bondora can be very high, with the highest interest rate in our sample being 255.19%, thus causing default probabilities to increase significantly. Also looking at the hazard rate for the Cox hazard model (1) from Table 2, shows a hazard rate, greater than 1, with significance at the p<0.01 level, indicating that a higher interest rate decreases the probability of survival for a loan.

Continuing to regression (2) from Table 2, where both credit ratings and interest rates are included, interest rates still increase the default risk, although the coefficient is much lower than that of regression (1). This could be due to interest rate being explained by credit ratings. This seems to be the case when comparing the adjusted R^2 results between the different regressions in Table 2, since the adjusted shows a marginal increase between regression (3) including credit ratings only and regression (2) including both credit ratings and interest rates in the same model. For the regressions in addition to the hazard rates, there is a clear relationship between a lower credit rating and a higher default rate, since a lower credit rating increases the probability of default, with the base case of credit rating: AA, having the lowest probability of default and the lowest credit rating (HR) having the highest probability of default. Including interest rates and credit ratings in the same regression also gives the same results as when making regressions for interest rates and credit ratings separately, however, the coefficients decrease slightly in size, in addition to the default probability not increasing between credit rating E and F.

Table 2. Interest Rates, Credit Ratings and Loan Performance

	Simple and Multiple regressions			Cox hazard model – Hazard rates		
Do interest rates and credit ratings predict loan performance?	(1)	(2)	(3)	(1)	(2)	(3)
Interest rate	0.008***	0.002***		1.020***	1.010***	
	(0.000)	(0.000)		(0.000)	(0.001)	
Credit rating A		0.091***	0.095***		1.306	1.353
		(0.026)	(0.026)		(0.338)	(0.350)
Credit rating B		0.216***	0.221***		1.966***	2.066***
		(0.023)	(0.023)		(0.468)	(0.492)
Credit rating C		0.275***	0.286***		2.476***	2.722***
		(0.023)	(0.023)		(0.580)	(0.637)
Credit rating D		0.401***	0.417***		3.372***	3.883***
		(0.023)	(0.023)		(0.788)	(0.905)
Credit rating E		0.491***	0.513***		4.017***	4.837***
		(0.024)	(0.023)		(0.944)	(1.133)
Credit rating F		0.483***	0.515***		4.278***	5.563***
		(0.025)	(0.024)		(1.014)	(1.308)

Credit rating HR		0.599***	0.644***		5.812***	8.292***
		(0.024)	(0.022)		(1.371)	(1.923)
Constant	0.252***	0.051**	0.079***			
	(0.008)	(0.022)	(0.021)			
N	19,412	19,412	19,412	6,037	6,037	6,037
Adjusted R^2	0.055	0.128	0.127			

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Notes. Default is the date when the loan went into defaulted state and collection process was started. For the credit ratings, dummy variables have been created. The dummy variable for credit rating AA has been removed, thus representing the base case. For the Cox hazard model, only loans with the same loan duration, in our case matured 36-month loans have been included, as a function of time is taken to be part of the hazard function (Cox, 1972). For the regressions (2) and (3), the VIF is less than 10, which is a common threshold (Hair et al., 2014). The VIF is still high for the credit rating variables, however since credit rating is a dummy variable with three or more categories, the high VIF is not a problem (Allison, 2012).

Looking at the results from Table 3 below, the age of a borrower seems to have a very small negative effect on default probabilities, with the regression and the Cox hazard model for age showing a slightly lower default probability for older borrowers. Thus, lending to older customers seems to be somewhat safer than younger customers. As the results for age shows a low adjusted R^2 however, the quality of this screening method is low.

Continuing, the *marital status* of the borrower, not being in a relationship shows an increase in default probability. Thus, it seems to play a role in determining default risks for borrowers.

Looking at countries, with the base case being Estonia, both Finland and Spain show a much higher default probability than Estonia, Spain having the highest default probability. Lending to Estonian customers thus seems to be the best option when one wishes to limit their risk.

Table 3. Age, Marital Status, Country and Loan Performance

	Linear and Multiple regressions			Cox hazard model – Hazard rates		
Do age, marital status and country predict loan performance?	(1)	(2)	(3)	(1)	(2)	(3)
Age	-0.001**			0.992***		
	(0.000)			(0.002)		
Marital status: Not in a relationship		0.070***			1.292***	
		(0.007)			(0.046)	
Country: Finland			0.331***			2.226***
			(0.009)			(0.094)
Country: Spain			0.379***			3.072***
			(0.009)			(0.142)
Constant	0.516***	0.461***	0.360***			
	(0.013)	(0.005)	(0.004)			
N	19,412	19,412	19,412	6,037	6,037	6,037
Adjusted R^2	0.000	0.005	0.119			

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Notes. Default is the date when the loan went into defaulted state and collection process was started. The dummy variable for Estonia has been removed, thus representing the base case. For the Cox hazard model, only loans with

the same loan duration, in our case matured 36-month loans have been included, as a function of time is taken to be part of the hazard function (Cox, 1972). For the regression (3), the VIF is much lower than 10, which is a common threshold (Hair et al. 2014). For Marital status, a relationship is defined as being either married or cohabitant while not being in a relationship means you are either single, divorced or a widow.

As can be seen in Table 4, lending to people not being fully employed and new credit customers shows a higher probability of default, thus the safest option is to lend to fully employed people and to reoccurring borrowers.

An increasing number of dependents slightly decrease the default probability of a loan. However, it seems more reasonable that a higher number of dependents would increase the default risk. As the results for number of dependents show a very low adjusted R^2 however, the quality of this screening method is low.

Table 4. Employment, New Credit Customer, Number of Dependents and Loan Performance

	Linear and Multiple regressions			Cox hazard model – Hazard rates		
Do employment status, new credit customer and number of dependents predict loan performance?	(1)	(2)	(3)	(1)	(2)	(3)
Employment status: Not fully employed	0.054***			1.082*		
	(0.010)			(0.050)		
New credit customer: True		0.143***			1.394***	
		(0.008)			(0.055)	
Number of dependents			-0.010***			0.950***
			(0.004)			(0.018)
Constant	0.484***	0.389***	0.500***			
	(0.004)	(0.007)	(0.004)			
N	19,412	19,412	19,412	6,037	6,037	6,037
Adjusted R^2	0.002	0.016	0.000			

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Notes. Default is the date when the loan went into defaulted state and collection process was started. Fully employed is the base case. *Not fully employed* includes borrowers that are either unemployed, partially employed, self-employed, entrepreneurs or retirees. When *new credit customer* is true, the base case is that the credit customer is an already existing customer. For the *number of dependents*, the base case is to have no dependents. For the Cox hazard model, only loans with the same loan duration, in our case matured 36-month loans have been included, as a function of time is taken to be part of the hazard function (Cox, 1972).

As can be seen in Table 5 below, females show a lower default probability than men on their loans, which also is in line with the findings by Agarwal et al. (2016), although their findings show a larger difference between men and women in terms of risk, finding that the odds of women being involved in bankruptcy events are 28% of those for men.

Looking at the results for *applied amount* and *amount* from Table 5, the higher the loan amount the customer has applied for, increases the default probability of a loan, whereas the higher the actual amount the customer is able to borrow decreases the risk of default. This seems reasonable, as a better and safer customer would be able to borrow a higher amount than a high-risk customer, however as the VIF for these values are high, one has to be careful to draw too many conclusions from this.

Finally, for Table 5, looking at the results for the effects of education on default probability, with primary education being the base case, higher education seems to decrease the probability of default, although vocational education has a higher probability of default than basic education, which is not in accordance with the assumption of a decrease in default probability from higher education.

Table 5. Gender, Amount, Education and Loan Performance

	Linear and Multiple regressions			Cox hazard model – Hazard rates			
Do gender, amount and education predict loan performance?	(1)	(2)	(3)	(1)	(2.1)	(2.2)	(3)
Gender	-0.040***			0.832***			
	(0.007)			(0.030)			
Applied amount		0.005***			1.000***	0.017***	
		(0.000)			(0.000)	(0.002)	
Actual amount		-0.003***			1.000***	-0.017***	
		(0.001)			(0.000)	(0.002)	
Basic education			-0.154***				0.606***
			(0.037)				(0.095)
Vocational education			-0.079**				0.748*
			(0.037)				(0.115)
Secondary education			-0.272***				0.462***
			(0.037)				(0.070)
Higher education			-0.273***				0.475***
			(0.037)				(0.073)
Constant	0.510***	0.433***	0.710***				
	(0.005)	(0.006)	(0.036)				
N	19,412	19,412	19,412	6,037	6,037	6,037	6,037
Adjusted R^2	0.002	0.014	0.027				

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes. Default is the date when the loan went into defaulted state and collection process was started. Male represents the base case for gender. The dummy variable for primary education has been removed, thus representing the base case. Amount and applied amount have been multiplied with 100 in order to show changes in these variables more clearly. The constant, however, hasn't been multiplied. For the Cox hazard model, only loans with the same loan duration, in our case matured 36-month loans have been included, as a function of time is taken to be part of the hazard function (Cox, 1972). For the Cox hazard models, the hazard coefficients for applied amount and amount are also displayed in (2.2) in order to better display the hazard rates. Their coefficients have also been multiplied with 100 in order to show changes in these variables more clearly. For the multiple regression (2), the VIF is less than 10, which is a common threshold (Hair et al. 2014), although the VIF values are still very high for both variables, making the interpretability of the results low. However, this is also reasonable as our statistics show that applied amount and actual amount are very similar. For the multiple regression (3), VIF is higher than 10, however, since this is a dummy variable with three or more categories, the high VIF is not a problem (Allison, P., 2012).

Looking at Table 6, not being a homeowner increases the risk of default, however, this is, contrary to the findings by Iyer et al. (2016), suggesting that people with homeownership have a higher risk of default.

Examining the use of loans, with *loan consolidation* being the base case for the purpose of applying for a loan, lenders wishing to use their loan for real estate purposes seem to experience the lowest default risk, whereas the use of a loan for loan consolidation seems to have the highest default risk, only looking at variables having a significance of at least p<0.1. The hazard rates show a slightly different result, with real estate purposes having the lowest risk and *use of loan health* having the highest risk, only looking at significant results.

Finally, the level of work experience does not seem to have any strong effects on default probabilities, with slightly higher default probabilities for people with longer work experience, with significant results for a work experience of 2-5 years and 10-15 years. Comparing the regressions to the hazard rates for work experience, the only significant result is that work experience of more than 25 years, implies a lower hazard rate than that of work experience fewer than two years (the base case), however overall for work experience, this variable does not give any clear trend in terms of whether longer work experience is beneficial in terms of default risks or not. Further, the adjusted R^2 results for this screening method is also low.

Table 6. Home Ownership, Use of Loan, Work Experience and Loan Performance

	Linear and Multiple regressions			Cox hazard model – Hazard rates		
Do ownership, loan purpose and work experience predict loan performance?	(1)	(2)	(3)	(1)	(2)	(3)
Home Ownership: Not owner	0.113***			1.540***		
	(0.007)			(0.056)		
Use of loan business		-0.046***			0.969	
		(0.017)			(0.085)	
Use of loan education		0.017			1.117	
		(0.021)			(0.115)	
Use of loan health		0.016			1.155***	
		(0.010)			(0.059)	
Use of loan home improvement		-0.008			0.981	
		(0.010)			(0.051)	
Use of loan real estate		-0.120***			0.691***	
		(0.024)			(0.097)	
Use of loan travel		-0.043**			1.002	
		(0.018)			(0.087)	
Use of loan vehicle		-0.073***			1.076	
		(0.014)			(0.079)	
Work experience 2-5 years			0.029*			1.140
			(0.017)			(0.100)
Work experience 5-10 years			-0.001			1.019
			(0.016)			(0.087)
Work experience 10-15 years			0.037**			0.996
			(0.017)			(0.086)
Work experience 15-25 years			0.026			0.942
			(0.016)			(0.080)
Work experience >25 years			0.002			0.866*
			(0.017)			(0.075)
Constant	0.436***	0.504***	0.476***			
	(0.005)	(0.007)	(0.014)			
N	19,412	19,412	19,412	6,037	6,037	6,037
Adjusted R^2	0.013	0.004	0.001			

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Notes. Default is the date when the loan went into defaulted state and collection process was started. Being a homeowner is the base case. As can be seen in the regression above, not being a homeowner clearly increase the default probability of a loan. Thus, the lowest risk is to lend to homeowners. The dummy variable for use of loan, loan consolidation has been removed, thus representing the base case. The dummy variable for work experience less than two years has been removed, thus representing the base case. The dummy variable for primary education has been removed, thus representing the base case. For the Cox hazard model, only loans with the same loan duration, in our case matured 36-month loans have been included, as a function of time is taken to be part of the hazard function (Cox, 1972).

Looking at Table 7 below, a higher number of existing liabilities for the borrower slightly decreases the risk of default, which seems unreasonable, in comparison to the more reasonable regression result that higher total liabilities for the borrower slightly increase the risk of default. A higher amount of free cash available slightly increase the default risk, which seems unreasonable, as this amount represents the amount available for the borrower to spend after liabilities have been paid. Finally, a longer loan duration slightly increases the default risk.

Table 7. Free cash, Liabilities, Loan Duration and Loan Performance

	Linear and Multiple regressions	Cox hazard model – Hazard rates	
Do free cash, income, liabilities and loan duration predict loan performance?	(1)	(1.1)	(1.2)
Existing liabilities	-0.017***	0.945***	
	(0.001)	(0.006)	
Free cash	0.004***	1.000	0.002
	(0.001)	(0.000)	(0.003)
Loan duration	0.006***		
	(0.000)		
Total liabilities	0.012***	1.000***	0.032***
	(0.001)	(0.000)	(0.003)
Constant	0.247***		
	(0.010)		
N	19,412	6,037	6,037
Adjusted R^2	0.066		

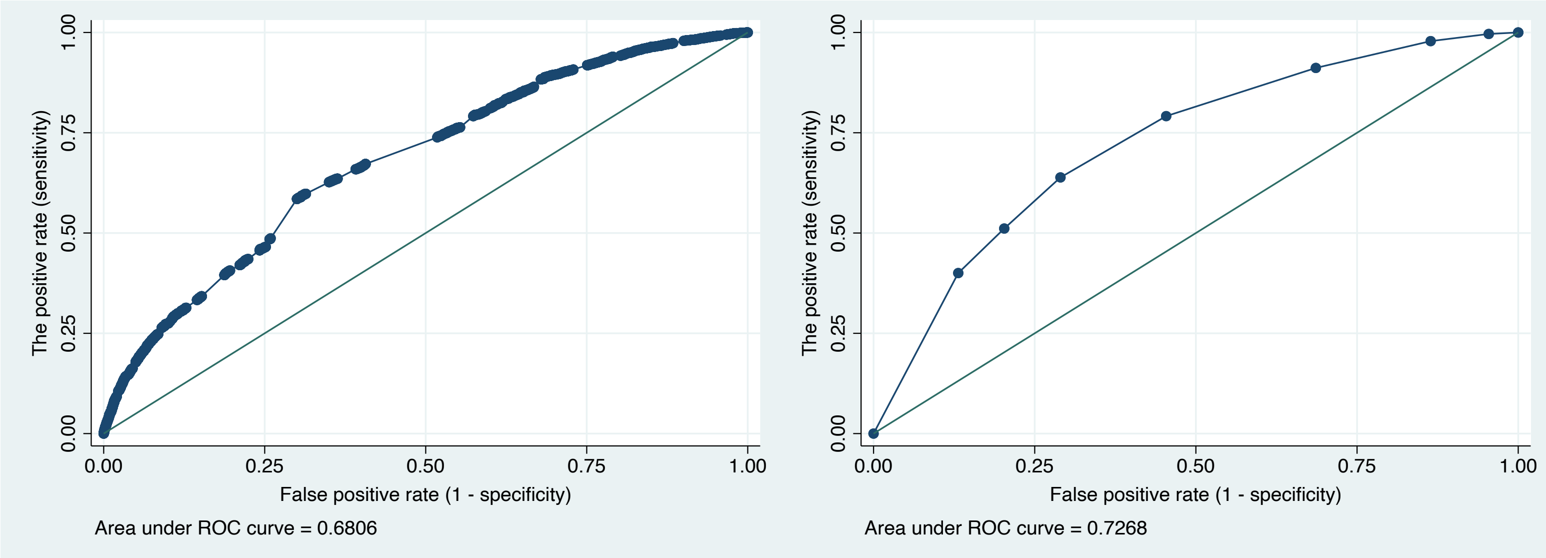
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Notes. Default is the date when the loan went into defaulted state and collection process was started. Male represents the base case for gender. Free cash and total liabilities have been multiplied with 100 in order to show changes in these variables more clearly, however the constant has not been multiplied. For the Cox hazard model in (1.1), the hazard coefficients for free cash and total liabilities are also displayed in (1.2) to better display the hazard rates. These coefficients have also been multiplied with 100 to show changes in these variables more clearly. For the Cox hazard model, only loans with the same loan duration, in our case matured 36-month loans have been included, as a function of time is taken to be part of the hazard function (Cox, 1972). Loan duration is not displayed for the Cox hazard model, as all observations for the model have a loan duration of 36 months, thus omitting loan duration from the model.

As can be seen in Graph 1-3 (A) and (B) below where Graph 1 shows ROC Curves for all credit categories (AA-HR), Graph 2 shows ROC Curves for all low credit categories (C-HR) and Graph 3 finally shows ROC Curves for all high credit categories (AA-B), credit rating has a higher AUC than the interest rate. Thus, being a superior predictor for default. This is the opposite compared to the findings of Iyer et al. (2016). However, as they use credit score, instead of credit ratings as in this paper, some of the difference could be explained in the difference in the variable used. In Graph 1 (A and B), both ROC Curves have a high area under ROC curve results, that is much higher than our suggested desirable level of 0.65 for these graphs. This is also the case for Graph 2 (A and B), having much higher results than our suggested level of 0.60 for these graphs. Finally, for Graph 3 (A and B) only Credit rating show ROC results higher than our suggested level of 0.60.

Graph 1. ROC Curves – All credit categories (AA-HR)

(A) Interest rate

(B) Credit rating

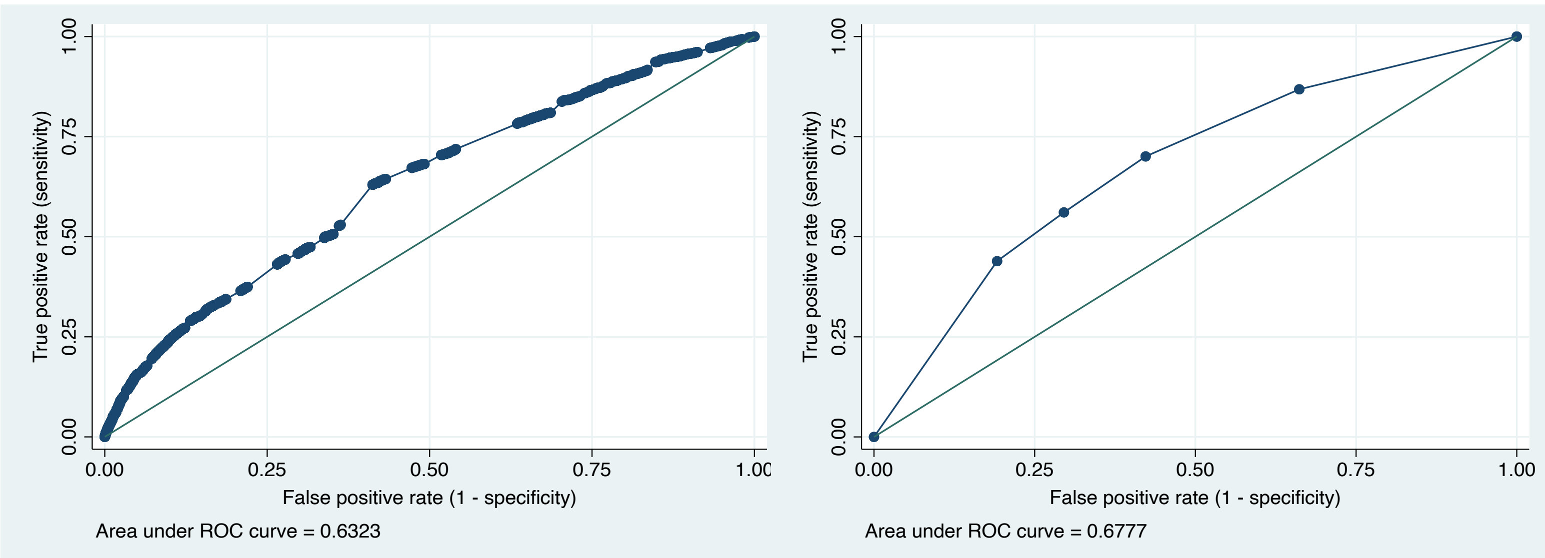


Notes. The number of observations: 23,418, significant at p<0.001. Significance was found using *roccomp* in STATA, in accordance with Iyer et al. (2016).

Graph 2. ROC Curves – Low credit categories (C-HR)

(A) Interest rate

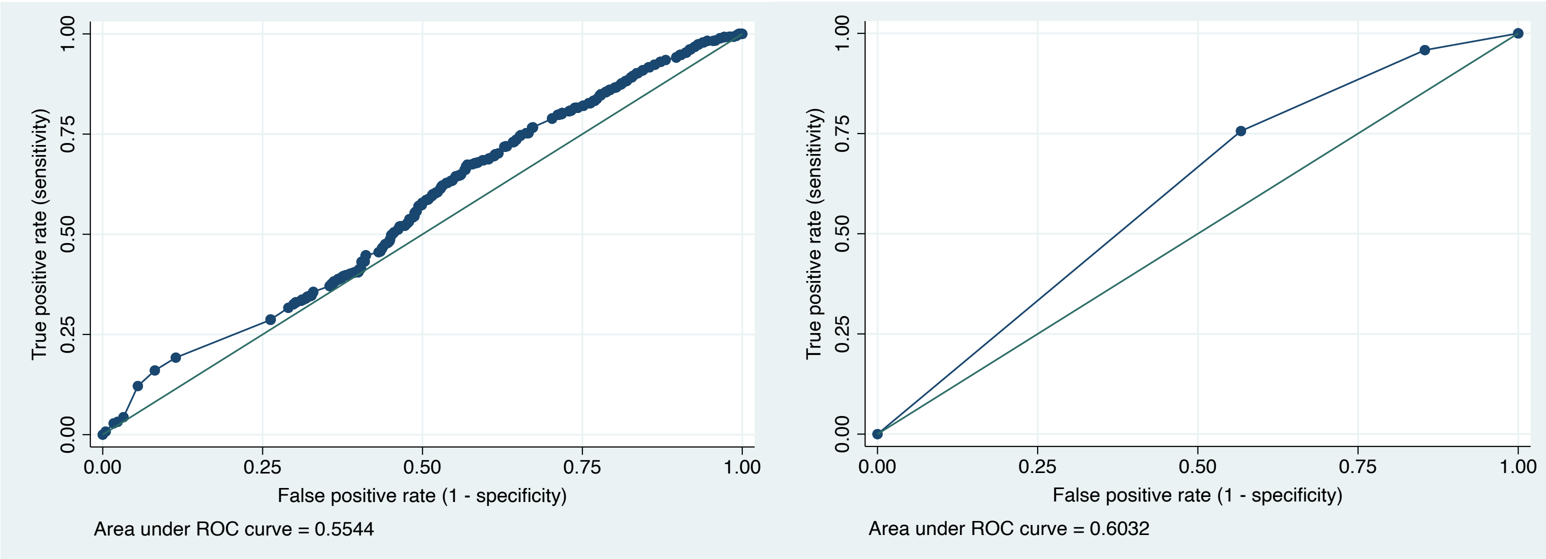
(B) Credit rating



Notes. The number of observations: 18,695, significant at p<0.001. Significance was found using *roccomp* in Stata, in accordance with Iyer et al. (2016).

Graph 3. ROC Curves – High credit categories (AA-B)

(A) Interest rate (B) Credit rating



Notes. The number of observations: 4,723, significant at $p<0.001$. Significance was found using *roccomp* in Stata, in accordance with Iyer et al. (2016).

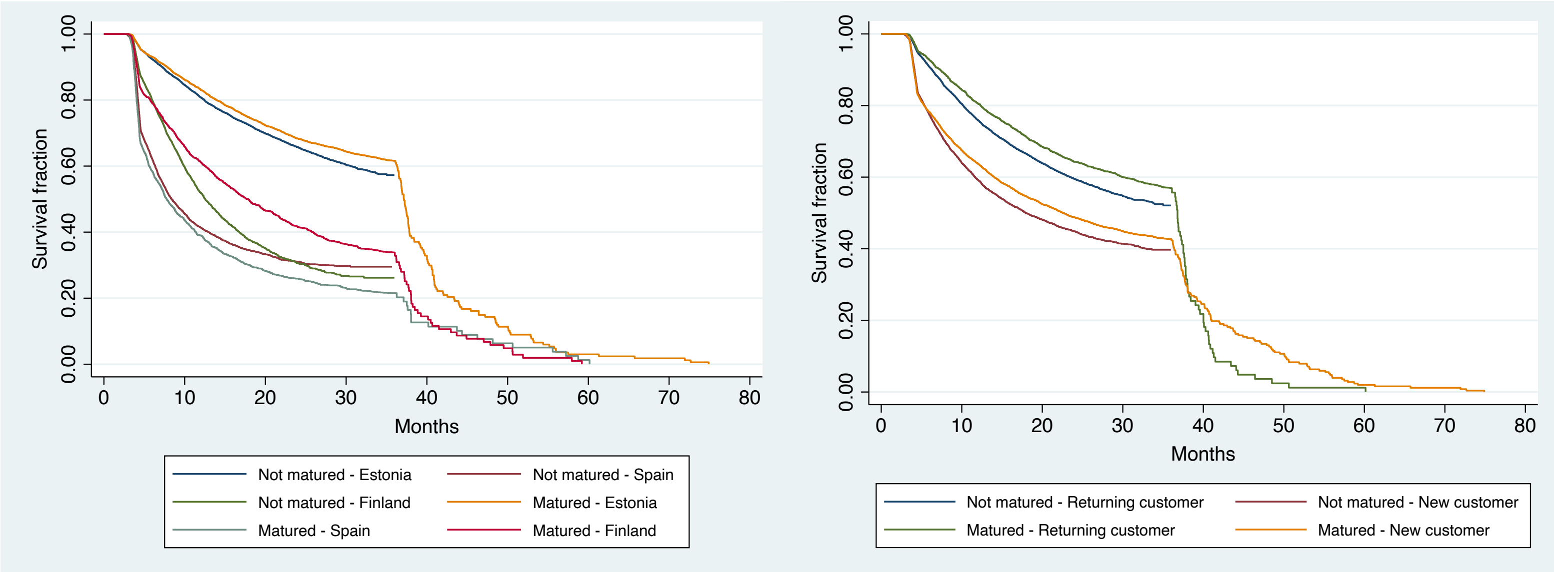
As can be seen in Graph 4 (A) below, Estonian loans have the highest overall survival rate, with Finland following, and lastly Spain. However, it is interesting to see that non-matured loans for Spain, although having a lower survival rate than Finland during the first months of loan duration, surpasses Finland during the last months before month 36, which is the month where the loan duration should have been completed. Overall, the graph shows that newly issued loans for both Estonia and Finland perform worse than already matured loans, indicating that the loan performance of loans from Estonia and Finland have decreased during the last years. For loans from Spain, the opposite has happened, where the loan performance for newly issued loans has improved in comparison to matured loans. Thus, the graph shows that newly issued loans perform worse than matured loans, indicating a decrease in overall loan performance.

Moving on to Graph 4 (B), it is clear that returning customers show a higher survival rate than for new customers. Comparing matured loans with newly issued loans, one can see that newly issued loans for both new customers as well as returning customers perform worse than matured loans, indicating that the overall performance of loans on the platform of Bondora has declined.

Graph 4. Kaplan-Meier survival estimates – Countries, New credit customer

(A) Country

(B) New credit customer



Notes. For the Kaplan-Meier graph, a total of 43,643 observations was included out of which 15,846 were failures (default). Total analysis time at risk and under observation was 622,274.28. Last observed exit was 74.90144 months. All available loans with a loan duration of 36 months have been included except for loans with age missing or age under 18, in addition to loans from Slovakia and loans with missing data on whether they are new credit customers or not. Although all loan durations in the graph have a 36-month loan duration, the graph continues after 36 months, as some loans might default after loan duration has ended, as discussed in 2.2 Data.

5. Discussion, managerial implications and suggested further research for business and scholars

Previous studies on default risk have often been focused on the general credit market and not much spotlight has been put on the peer-to-peer lending market. In a digital and global world, it should be in everyone's interest to get a better understanding of the peer-to-peer industry. This paper has looked into one lending marketplace but aspires to bring an understanding of the entire industry. It is reasonable to assume that one study of one marketplace is not enough to build a complete model, but this paper provides a starting point.

To make this paper relevant, the managerial implications have been kept in mind. An important part of the peer-to-peer business is trust (Duarte et al., 2012), this trust is two-fold: the lenders need to trust the borrowers, but they also need to trust that the platform correctly assesses the borrowers. When a borrower fulfils their payment obligations, both the platform and the investor yield a profit. However, if the borrower default, the risk is with the investor. Hence, both private lenders and peer-to-peer businesses might also find it useful to see what factors are important when assessing a borrower.

The idea of this paper is not to tell peer-to-peer investors and marketplaces to never lend money to people that show features of having a higher risk of default, but to create awareness regarding what factors increase the risk of default and what factors decrease the risk of default to better assess borrowers in the future.

We found credit ratings in addition to interest rate to be important variables in determining default risks, with credit ratings having a greater AUC than the interest rate. Contrary to the findings by Iyer et al. (2016), interest rate has a greater adjusted R^2 than that of credit score, although one has to note that they compare interest rate and credit score instead of our comparison of interest rate and credit rating. This can be seen as an indicator that the interest rates are not solely affected by the credit rating but also third-party sources unavailable to stakeholders.

We found that a higher value of free cash leads to a higher risk of default, which seems unreasonable since the definition of free cash is a residual after liabilities have been paid. We have earlier described how a stable salary (being fully employed) is more important than a high income. This finding, while surprising, points in the same direction, that having more money not significantly reduces the risk of default. When assessing borrowers, the focus should, therefore, be on employability rather than income. It might prove beneficial for marketplaces to consider constructing a variable predicting employability, in order to better predict default. This variable could be constructed by using third-party data from companies producing tests for employment processes.

Also, our results showed mixed results for the amount and number of liabilities a person has, with a higher number of existing liabilities decreasing the default risk, which seems unreasonable, in addition to the more reasonable increase in default risk from a higher amount of total liabilities. For future research, it would be interesting to test if this is true for all of the loans for borrowers with a high number of existing liabilities. Does having a higher number of liabilities decrease the default risk on all of the loans of the borrower or only the newest, oldest, largest or smallest?

Drawing an implication from the variable *use of loan* where *loan consolidation* has the highest risk of default and *real estate* has the lowest, our recommendation for peer-to-peer marketplaces would be not to specialise in loan consolidation but rather focus their marketing campaigns towards real estate.

Seeing that returning credit customers has a lower risk of default can imply two different things: either, Bondora effectively screens how borrowers behave during their first loan and are more restrictive when giving out the second loan. Unfortunately, there is no data available for loan applications that have been rejected by Bondora to test this. Or, Bondora is not more restrictive but rather the returning customers have developed a loyalty to Bondora and are therefore keener on making payments on their subsequent loan. Again, there is no variable on customer satisfaction to test this.

As our data suggest that females have a lower default risk than men, it is suitable to lend to females at a higher degree. Similar results have been found by Agarwal et al. (2016). That being said, ethical concerns could potentially affect this way of lending. In addition, the supply of credit is more than enough to cover loans requested by female customers, thus male customers are also lent money. We do not suggest that males should not be lent capital but rather an implication for peer-to-peer marketplaces would be to angle their external communication to appeal to potential female customers.

With more information regarding who defaults or not, the ethical question of how much of that information should be passed on to the investors arise. Refraining from sharing information about the borrowers might result in more loans being funded. As found by Valle and Zeng (2019), this problem of choosing between increasing the volume or decreasing the adverse selection comes down to the decision of the marketplace. The outperformance of sophisticated investors will probably shrink when less information is provided (Valle and Zeng, 2019). Investors do care about the information, with the knowledge of what variables are the most important this outperformance can be shown to be greater. As discussed earlier, it is important for investors that they can trust borrowers and the marketplace. With this in mind, the argument can be made that trust also is a part of the information sharing consideration which marketplaces are faced with.

In call for action, the findings in this paper do not suggest radical changes in the way which peer-to-peer marketplaces assess borrowers, but rather suggests future investigations of the borrower's behaviour. With regards to the questionable findings that a higher applied amount specified by the borrower increases the risk of default while a higher actual amount received decrease the risk of default, it might be of interest to conduct experiments with borrowers applying for the same amount but systematically giving them different actual amounts. It would be interesting to see if this modification gives results similar to those described in this paper. There is a risk that the borrowers applying for a higher amount than they receive are riskier in general, this is proven in the data where high-risk borrowers are those with the largest gap, on average, between their applied amount and actual received amount. However, on average higher-rated borrowers borrow smaller amounts, in contrast to the finding that a higher actual amount decrease risk. What can be established is that higher rated borrowers are more likely to receive the amount that they apply for, thus incentivizing higher-rated customers to apply for larger amounts could prove beneficial for marketplaces.

Regarding the finding that the customers in Estonia have the lowest default risk, further research is suggested on other cross border peer-to-peer platforms active in Estonia. This is to see if the Estonian borrowers are simply better at paying back their loans or if there is a correlation between a marketplace's home country and the distribution of risk between different markets. Would Estonian customers at a Spanish peer-to-peer marketplace have a higher default rate than the Spanish customers and vice versa? Most probably this is not the case, as previous data shows Spanish loans to perform worse than Estonian loans (European Banking Authority [EBA], 2019). Additionally, our data shows that Spanish loans have improved in terms of a lower default risk during the last years, which has not been the case for Estonian and Finnish loans. Here it would be interesting to see whether Bondora could implement the same strategies that have helped improve their Spanish loans to the Estonian and Finnish loan market as well, to improve the overall performance of Bondora. One explanation for the better performance of Spanish loans could be that loans are issued to customers with a higher credit rating on their newer loans, proven by the data showing that recent loans issued for Spanish borrowers have on average, a better credit rating. 77% of the matured loans in Spain have an HR-rating, the same number for the non-matured loans is only 29%. The majority of the active loans now instead have an F-rating (53%).

This thesis highlights that not just one or a few variables should be taken into consideration when assessing borrowers, but instead a multitude of variables should be used. Peer-to-peer platforms should use this information as a call to keep screening for new variables to include in their models.

When collecting data, the lack of proper explanations of variables was discovered. If lenders and platforms strive to assess borrowers accurately, this is an area that needs more focus. The authors have been in contact with Bondora that have agreed to improve their data explanations (see Appendix B). A general suggestion to all peer-to-peer platforms would be to not only provide access to loan data for investors but also clearly describe the data.

An insight from the work with this paper is not only the lack of interest in the topic from scholars but also the surprise in interest from Bondora. It should be in Bondora's interest to uphold an interest in the way borrower's assessment from the investors since an efficient assessment is the main selling point of lending platforms. However, this is very vaguely communicated. A large improvement could be made on how marketplaces communicate with investors regarding the borrower assessment since both parties would benefit from a better understanding of the optimal borrower assessment process.

6. Conclusion

This thesis has looked into the peer-to-peer lending industry by analysing loans provided by the Estonian peer-to-peer marketplace Bondora. Analyses have been conducted to assess borrowers and evaluate what variables contribute to the default risk.

Peer-to-peer lending provides possible access to capital for borrowers with limited access to other sources of lending. Either cheaper (Tang, 2019) or more expensive (Morse, 2011). In order for peer-to-peer marketplaces to stay relevant, a correct assessment of borrowers is vital. This paper has therefore looked into which variables have the most important effects on the default risk of borrowers.

This paper has found credit rating to be a better explanatory variable for default than the interest rate. New credit customers have a higher default risk than existing credit customers. Out of the countries Estonia, Finland and Spain, Estonian loans perform the best and Spanish loans perform the worst. Loan performance for female borrowers performs better than male borrowers. The performance of loans taken for real estate purposes performs the best, whereas loans for loan consolidation perform the worst.

Future studies within the peer-to-peer industry should investigate data sets from other marketplaces to see what results are specific to Bondora and which results can be used to draw general conclusions, with a focus on the questions discussed in *Discussion, managerial implications and suggested further research for business and scholars*.

7. Acknowledgements

The authors are thankful for the answers and clarifications when in contact with Bondora, especially with Kairi and Michelle at *Investor Relations*. The authors thank Frank Emericks for helpful comments during the write-up. The authors thank the management of the Stockholm School of Economics for keeping their premises open during the Covid-19 pandemic and providing a safe working environment for the authors. The authors thank supervisor Marcus M. Opp for constructive feedback during the thesis process. The authors thank the department of finance and the department of management at the Stockholm School of Economics for making this cross-department thesis possible.

The views expressed in the thesis are the author's own. The data from Bondora has been provided without any requirements. Any faults in the collection of data and unclarities of definitions are to be answered by Bondora. Any faults in interpreting the data and when conducting calculations are to be answered by the authors.

Finally, the authors want to encourage more research within the field of peer-to-peer business. Both when it comes to lending but also other industries.

8. References

ABELER, J., NOSENZO, D. and RAYMOND, C., 2019. Preferences for Truth-Telling. Econometrica, 87(4), pp. 1115-1153

AGARWAL, S., HE, J., SING, T.F. and ZHANG, J., 2016. Gender Gap in Personal Bankruptcy Risks: Empirical Evidence from Singapore. Review of Finance, 22(2), pp. 813-847

ALLISON, P. 2012. Statistical Horizons. Retrieved 2020 May 1 from https://statisticalhorizons.com/multicollinearity

COX, D. R., 1972. Regression Models and Life-Tables. Journal of the Royal Statistical Society, 34(2), pp. 187-220

DELONG, E. R., DELONG, D. M. and CLARKE-PEARSON, D. L. 1988. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics. 44(3). Pp. 837-845

DUARTE, J., SIEGEL, S. and YOUNG, L., 2012. Trust and Credit: The Role of Appearance in Peer-to-peer Lending. The Review of Financial Studies, 25(8), pp. 2455-2483

European Banking Authority. (2019). EBA report on NPLs (2019). Retrieved from https://eba.europa.eu/risk-analysis-and-data/npls

GROSS, D. B., SOULELES, N. S. 2002. An Empirical Analysis of Personal Bankruptcy and Delinquency. The Review of Financial Studies, 15(1), pp. 319-347

HAIR, J. F., BLACK, Jr. W. C., BABIN. B. J., ANDERSON, R. E. 2014. Multivariate Data Analysis (Seventh Edition). Place of publication: Pearson.

IYER, R., KHWAJA, A.I., LUTTMER, E.F.P. and SHUE, K., 2016. Screening Peers Softly: Inferring the Quality of Small Borrowers. Management Science, 62(6), pp. 1554-1577

KEYS, B., MUKHERJEE, T., SERU, A. and VIG, V., 2010. Did Securitization Lead to Lax Screening? Evidence from Subprime Loans. The Quarterly Journal of Economics, 125(1), p. 307

MORSE, A., 2011. Payday lenders: Heroes or villains? Journal of Financial Economics, 102(1), pp. 28-44

TANG, H., 2019. Peer-to-Peer Lenders Versus Banks: Substitutes or Complements? The Review of Financial Studies, 32(5), pp.1900-1938

VALLE, B. and ZENG, Y., 2019. Marketplace Lending: A New Banking Paradigm? The Review of Financial Studies, 32(5), pp. 1939-1982

9. Appendix

9.1. Appendix A

Explanations of the variables used, and supplementary data can be found at:

https://www.bondora.com/en/public-reports

Information on how *B Secure* (changing a loan) works can be found here:

https://www.bondora.ee/en/about-bondora/faq/#how-do-i-apply-for-a-new-payment-schedule

The history of Bondora:

https://www.bondora.com/en/road-to-100-million

For numbers concerning Bondora that is not backed up with a reference we refer to information that can be found on the Bondora webpage: https://www.bondora.com/en

9.2. Appendix B

Quotes from email conversations with *Investor Relations* at Bondora*:*

1. "externally validated data we get from credit bureaus, population registries, banks and tax authorities"

Explaining where third party data is collected from.

2. *"*Minimum age to apply for a loan is 18."

Regarding the fact that some loans show an (borrower) age of 0, 1 or 2.

3. " We need to review the translation key for the term 'Homeless' as this is incorrect, it means that we have no information on this, not that the person is a homeless person."

Regarding that, the variable can take the value "0 = homeless" when it should be "0 = not available. As well as other mistake and unclear definitions.