University of Warsaw

“What factors determine investor’s gain or loss on p2p lending market? ”

Prepared by

Stanisław Smyl 353838

Iryna Bazaka 363461

Warsaw, 2018

Abstract

The main goal of our project is to analyse how available information on the p2p lending platform about borrowers can determine gain or loss. We found, that significant variables are amount, real interest rate, total income, Estonia and Slovakia countries, undefined gender, home ownership types homeless and mortgage and secondary education. Also, our main hypothesis, that higher amount decreases the probability to gain is improved. However, our second hypothesis, that borrowers from relatively rich region has greater probability of paying off the debt, is not improved. Besides, we found that the interaction between region variable and amount of existing liabilities of borrower is statistically significant.

Contents

Introduction

The widespread availability of various peer-to-peer lending solutions is rapidly changing the landscape of financial services. Because of its easy and fast online application process and also lower interest rates compared with some credit cards, more people are likely to take a loan using the p2p platform rather than traditional financial institutions.

Peer-to-peer (P2P) lending, also known as "social lending", lets individuals lend and borrow money directly from each other. Just as eBay removes the middleman between buyers and sellers, P2P lending companies like Zopa and Prosper eliminate financial intermediaries like banks and credit unions.[1]

As the popularity of online peer-to-peer lending increases among borrowers and investors, so we decided to build an econometric model which could describe factors, that determine whether lender gain or loss money at given loan.

We test the main hypothesis, that the higher amount of loan decreases the probability of gain. We test second hypothesis, that being a borrower from rich region, often including capital city of country, increases the probability of gain. Second hypothesis is based on the intuition that the richer region, the higher average salary and more opportunities for finding new, better job. Results of our research can be used to support investor's decision making process.

The remainder of this essay is organized as follows. Section I reviews the literature about peer to-peer online lending platforms. Section II describes the raw dataset and methods used for cleaning and features engineering. Section III present the data explanatory techniques, and section IV reports modeling part. Section V includes findings and possible future improvements.

I.Literature Overview

Many researches have focused on understanding the general dynamics of the online peer-to-peer lending marketplace. Hulme and Wright (2016) provide a study of peer-to-peer lending focusing on "Zopa.com". Ceyhan et al. (2011) study the dynamics of bidding behavior in a P2P loan auction website, Prosper.com. They investigate the change of various attributes of loan requesting listings over time, such as the interest rate and the number of bids.

[1] Investopedia: https://www.investopedia.com/articles/financial-theory/08/peer-to-peer-lending.asp

Garman (2008) writes if a borrower wants to get fully funded loans, verification processes of his ability to pay, including verification of income have been implemented directly by some online platforms, for example prosper.com. Herzenstein et al. (2008) and Galloway (2009) announce that in this kind of lending model the mediation of financial institutions is not required.

Another part of authors concentrates on the factors influencing peer-to-peer default and credit risk, etc. Herzenstein et al. (2008) study the determinants of funding success in an online P2P lending community. The result of research provide support, that demographic attributes such as race and gender do affect likelihood of funding success, their effects are very small in comparison to effects of borrowers' financial strength (credit grade, debt to income ratio, and homeownership) and their effort when listing and publicizing the loan.

Freedman and Jin (2008), examine the functioning of online lending based on Prosper's transaction data. They study the effect of social network in identifying risks and find evidence both for and against it. In some years later, the authors found, that borrowers with social ties are consistently more likely to have their loans funded and receive lower interest rates; however, most borrowers with social ties are more likely to pay late or default. Freedman and Jin (2016) provide evidence that these findings are driven by lenders not fully understanding the relationship between social ties and unobserved borrower quality. Overall, the findings suggest caution for using online social networks as a signal of quality in anonymous transactions.

Interesting hypothesis was tested by Lu et al. (2012). The analysis suggests that online borrowers are significantly influenced by defaults in their social networks. The authors find that not all friends have equal influences. The social influence is highly significant among online friends made through the peer-to-peer lending site. Social influence is much weaker in magnitude among offline friendships that were carried over to the peer-to-peer lending site.

Zhang et al. (2017) analyse the factors that determine the probability of obtaining the loan in online P2P lending. The result of the model is interesting. It indicates that annual interest rate, repayment period, description, credit grade, successful loan number, failed loan number, gender, and borrowed credit score are significant factors to determine the availability of loan funding on Paipaidai platform. Annual interest rate, credit grade, successful loan number, gender, and borrowed credit score have a positive impact on loan success, while repayment period, description and failed loan number have a negative impact.

Jagtiani and Lemieux (2017) explore the advantages/disadvantages of loans made by a large fintech lender – Lending Club and similar loans that were originated through traditional banking channels. The authors found high correlation with interest rate spreads, Lending Club rating grades, and loan performance.

II. Dataset

For our econometric model we use dataset from Estonian online peer-to-peer platform, Bondora. It is one of the biggest platforms in Europe, with 125 millions EUR of loans issued and net return of 11.9%. We have chosen this platform because of its popularity, a lot of loans available to investors, and reach database contains much information about borrowers and their characteristics. Another reason to analyse Bondora is that this dataset is not as popular is scientific researches as other peer-to-peer lending platforms such as Lending Club, Prosper or Zopa.

What is more, Bondora operates within unique set of rules. It enables borrowers to requests chosen amount of loans and specified maximum interest rate. Then, Bondora, based on its scoring models decide whether to accept one's offer, or reject it. When accepted, the bidding begins and investors submit their lending offers with specified interest rates. After bidding is over, borrower is connected to investor who offered the lowest interest rates, but it happens only when the offered rate is lower than the maximum rate defined by borrower. This business model is close to Pareto optimum in the way that each borrower receive the lowest interest rate as possible given his characteristics and previous loan history, and investors have almost perfect information about each borrower to set the premium accurately.

At the beginning, our dataset contained of 54178 observations and 112 variables. However, our goal is to predict whether investing in given loans will result in gain or loss, so we kept only variables that are available to investor just before deciding to invest in loan or not. What is more, we focused only on closed loans. It left us with 39 variables and 12047 observations.

We constructed class variables in following manner: firstly, for each loans we computed internal rate of return for irregular cash flows (Xirr). When the loan's status was 'Default', we included debt collected by debt collection company. Given lack of a date when amount of collected debt was given to investor, we assumed that it happened one year after default date.

Then, after binarizing Xirr, we received GainLoss class variables that took 1 when Xirr was greater or equal to 0, and 0 otherwise. What is more, we created new variables based on

previous ones and include ‘Other’ level to categorical variables with many levels and created ‘Undefined’ level for observation, when the value for categorical variable is missing. For numerical variables, we simply imputed missing values with mean within two groups based on IsHarju variable. Then, to choose appropriate subset of variables we used filter feature selection method, Joint Mutual Information (JMI), that is said to be the state-of-the-art method for selecting relevant and not redundant features [Brown et al. (2012)]. Finally, after analysing, cleaning and features engineering that included creating new variables based on previous ones and features selection process, we constructed database contained of 12047 observation, one class variable and 24 independent ones. Descriptive statistics of explanatory variables are presented in the Table 1 and Table 2.

Table 1. Descriptive statistic of continuous variables

Statistics / Abbreviated Name	Mean	St. Dev.	Min	Median	Max
Free Cash	413.60	444.848	-2332.00	329.00	11508.11
Amount	1686.33	1610.002	6.39	1165.00	10630.00
Income Total	2885.0	5070.215	0.0	1161.0	133000.0
Liabilities Total	667.5	1799.079	0.0	451.0	172510.0
Stable Income	2485	4479.545	0	1065	133000
Applied Amount	1766.15	1701.306	31.96	1200.00	10630.00
Amount Of Previous Loans Before Loan	625	1618.487	0.0	0.00	29590
Debt To Income	21.38	17.82289	0.00	18.93	198.02
Real Interest Rate	0.4506	0.2476088	0.0630	0.3795	3.5935
Previous Repayments Before Loan	360.33	1178.297	0.00	0.00	29535.43
Age	36.2	11.50433	0.0	34.0	77.0
Loan Duration	20.41	12.06336	1.00	18.00	60.00
Unstable Income	400.4	1381.347	0.0	0.0	34000.0
Existing Liabilities	3.842	3.359637	0.000	3.000	27.000

Table 2. Discrete Variables

Variable		n
		403
Is Harju:	Yes	9
		800
	No	8
		214
Country:	ES	8
		117
	FI	1
	SK	95
		863
	EE	3
		108
Language Code:	Finnish	6
		775
	Estonians	6
		212
	Spanish	9
		107
	Other	6
		672
Gender:	Man	6
		490
	Woman	7
	Undefined	414
Home Ownership Type:	Homeless	45
	Council house	141
	Joint ownership	846
	Joint tenant	355
		203
	Living with parents	2
		100
	Mortgage	2
		308
	Owner	8
	Owner encumb	159
		187
	Tenant prefur	2
	Tenant unfur	855
		165
	Undefined	2
		138
Education:	Basic education	6
		323
	Higher education	4
	Primary education	121
		540
	Secondary education	6
	Undefined	53
	Vocational education	184

		7
Work Experience:	Less than 2 years	964
		194
	2 to 5 years	6
		276
	5 to 10 years	5
		210
	10 to 15 years	4
		231
	15 to 25 years	2
		186
	More than 25 years	0
	Undefined	96
Marital Status:	Divorced	949
		330
	Married	1
		404
	Single	4
		344
	Cohabitant	3
	Widow	218
	Undefined	92
		657
Nr Of Dependants:	0	4
		281
	1	2
		167
	2	7
	3 and more	745
	Undefined	239
Use Of Loan:	Business	596
	Educations	410
	Health	515
		281
	Home	0
		271
	Loan consolidation	5
		306
	Other	9
	Travel	620
	Unknown	39
		127
	Vehicle	3

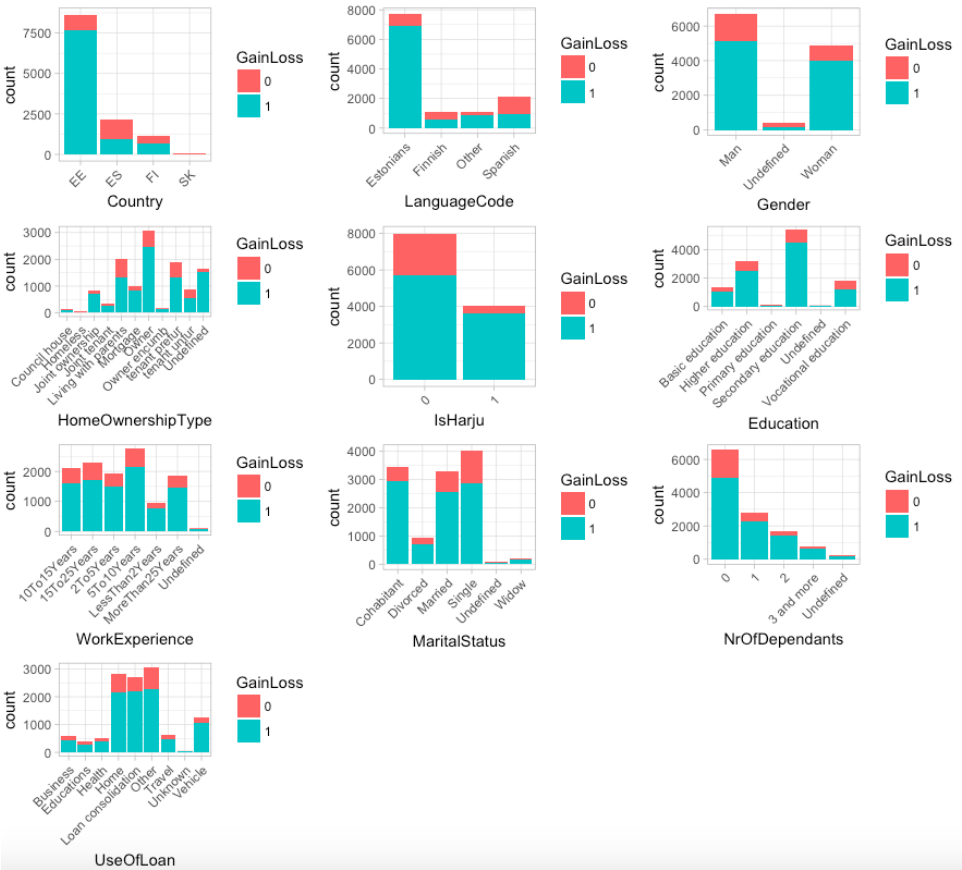
III. Exploratory Data Analysis

Before modelling and testing hypothesis we perform exploratory data analysis. In our dataset, in 22.89% cases investors lose their money and in 77.11% cases they make a profit. It means that class variable is imbalanced and such evaluation statistics as Count R2 (Accuracy) would not be helpful in evaluation of our model. We must be aware that imbalanced data can result in the model that assign all observations to one, bigger class, giving accuracy equals to its ratio. In our case, it would result in accuracy equals to 77.11%.

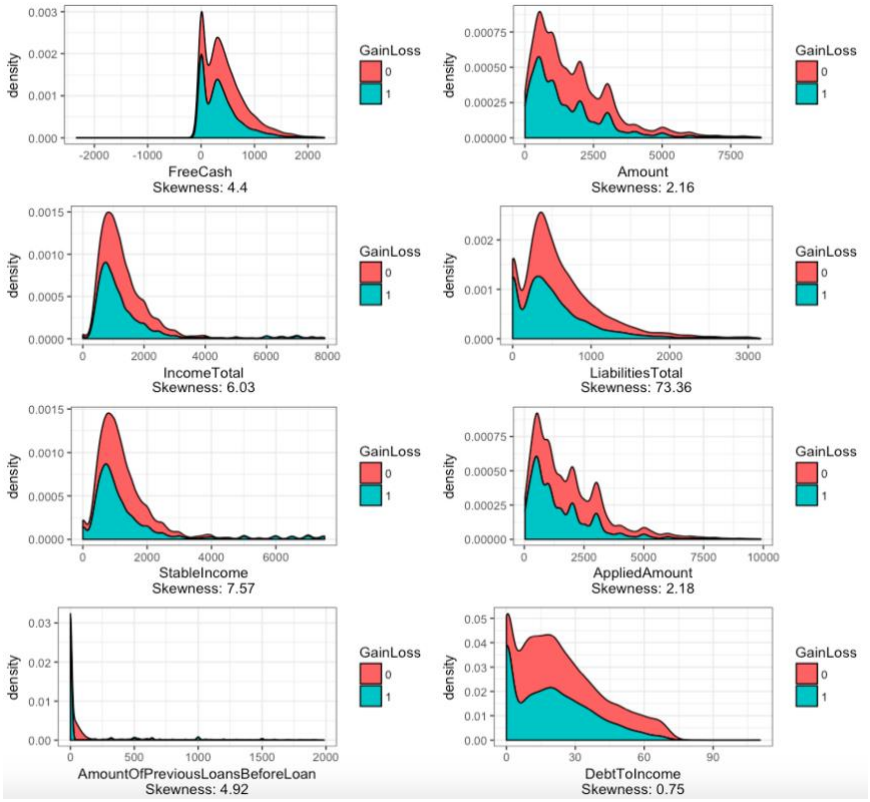
On the Picture 1 we can see histograms for categorical variables. Based on the visualisation we can see that borrowers from Harju region may have significant impact on the loss. Another interesting observation is that investors usually lose money on loans where the borrower is single and has no dependants. What is more, we can see that investing in Estonians loans lower probability of loss when compared with borrowers from other countries, especially from Spain.

Based on the Picture 2 and Picture 3 we can conclude that the most of variables are positive (right) skew, what means that mean is bigger than median of the variables.

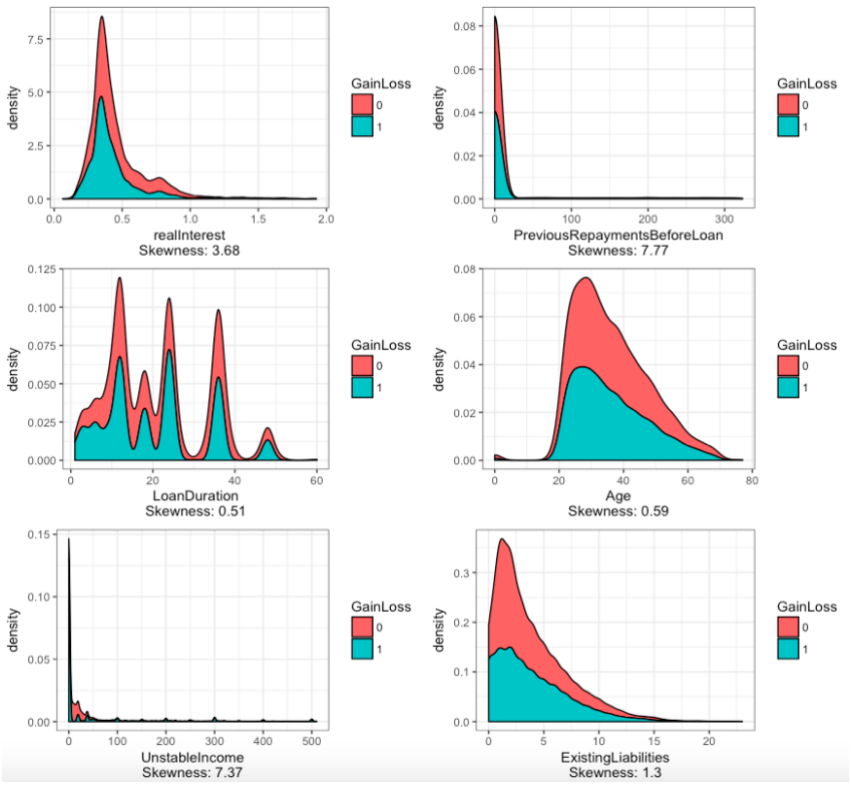
Picture 1. Categorical variables



Picture 2.

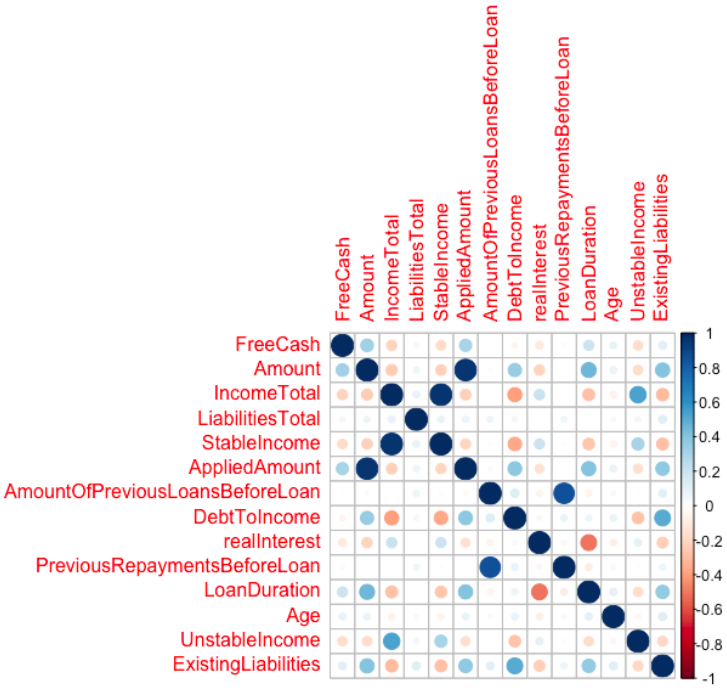


Picture 3.



Using Pearson method we computed the correlation between numeric variables. The results are presented on the Picture 4.

Picture 4. Pearson's correlation



We found, that groups of (Stable Income and Total Income) and (Amount and Applied Amount) were highly correlated, that's why we excluded previously created within features engineering process StableIncome and AppliedAmount variables.

IV. Methods

Because primary goal of investors is not to lose money very often, our optimization will focus on severely discriminating against borrowers with potential for default. In our research, we use three methods: linear probability model, logit model and probit model to analyse the dependence between investor's gain or loss and early presented independent variables. Based on information criteria and general to specific approach, we found that probit model performs better than logit. The results of final estimation are presented in the Table 3.

Based on visualizations and AIC information criteria, we added to model non-linear variables such Amount2, AmountOfPreviousLoansBeforeLoan2, DebtToIncome2, realInterest2, LoanDuration2, Existing Liabilities2 and two interactions – Harju region both with Existing Liabilities and Vocational Education. What is more, we used binary one hot encoder method

to binarize all categorical variables. It left as with base level of male borrower who is native Estonian, has a basic education level, owns the council house, has cohabitant, no dependants, has the work experience of 10 to 15 years and need money for business purpose.

Table 3. Results of Logit and Probit models

	Logit Model			Probit Model		
Coefficients:	**Estimate**	**Std. Err.**	**Pr(>|z|)**	**Estimate**	**Std. Err.**	**Pr(>|z|)**
Constant	3,101e+00	4,502e-05	< 2e-16 ***	1,79E+00	1,27E-01	< 2e-16 ***
Amount	-3,16E-04	4,50E-05	2,34e-12 ***	-1,803e-04	2,58E-05	2,64e-12 ***
IncomeTotal	6,02E-05	1,044e-05	8,03e-09 ***	2,97E-05	5,18E-06	9,54e-09 ***
AmountOfPreviousLoansBeforeLoan	-2,17E-04	3,202e-05	1,21e-11 ***	-1,313e-04	1,86E-05	1,56e-12 ***
realInterest	-2,15E+00	3,298e-01	7,46e-11 ***	-1,23E+00	1,89E-01	7,93e-11 ***
PreviousRepaymentsBeforeLoan	4,52E-04	5,12E-05	< 2e-16 ***	2,45E-04	2,77E-05	< 2e-16 ***
LoanDuration	-3,933e-02	1,02E-02	0,000119 ***	-2,163e-02	5,74E-03	0,000166 ***
ExistingLiabilities	1,31E-01	2,33E-02	1,74e-08 ***	7,65E-02	1,320e-02	6,86e-09 ***
CountryES	-1,88E+00	7,44E-02	< 2e-16 ***	-1,115e+00	4,33E-02	< 2e-16 ***
CountrySK	-2,43E+00	2,54E-01	< 2e-16 ***	-1,463e+00	1,500e-01	< 2e-16 ***
LanguageCodeFinnish	-1,43E+00	8,57E-02	< 2e-16 ***	-8,338e-01	5,06E-02	< 2e-16 ***
LanguageCodeOther	-3,66E-01	9,72E-02	0,000165 ***	-2,14E-01	5,40E-02	7,39e-05 ***
GenderUndefined	-4,09E-01	1,226e-01	0,000859 ***	-2,501e-01	7,42E-02	0,000747 ***
HomeOwnershipTypeHomeless	-3,21E+00	3,56E-01	< 2e-16 ***	2,306e-01	5,84E-02	7,79e-05 ***
HomeOwnershipTypeMortgage	3,89E-01	1,05E-01	0,000196 ***	2,306e-01	5,84E-02	7,79e-05 ***
EducationHigher_education	4,641e-01	6,75E-02	5,99e-12 ***	2,90E-01	3,941e-02	1,72e-13 ***
EducationSecondary_education	3,693e-01	6,08E-02	1,28e-09 ***	2,28E-01	3,54E-02	1,12e-10 ***
WorkExperience15To25Years	-2,18E-01	6,60E-02	0,000957 ***	-1,21E-01	3,76E-02	0,001261 **
WorkExperience2To5Years	-3,229e-01	7,15E-02	6,26e-06 ***	-1,827e-01	4,03E-02	5,67e-06 ***
WorkExperienceLessThan2Years	-3,55E-01	9,85E-02	0,000321 ***	-1,96E-01	5,48E-02	0,000342 ***
NrOfDependants2	1,93E-01	7,91E-02	0,014745 *	1,10E-01	4,40E-02	0,012133 *
UseOfLoanLoan_consolidation	2,34E-01	7,58E-02	0,002042 **	1,38E-01	4,25E-02	0,001148 **
UseOfLoanOther	1,30E-01	6,30E-02	0,039265 *	7,41E-02	3,60E-02	0,039318 *
UseOfLoanVehicle	2,61E-01	9,20E-02	0,004547 **	1,431e-01	5,12E-02	0,005225 **
IsHarjuXExistingLiabilities	3,47E-02	1,30E-02	0,007604 **	1,57E-02	6,94E-03	0,024077 *
Amount2	2,085e-08	5,219e-09	6,46e-05 ***	1,18E-08	3,01E-09	9,30e-05 ***
AmountOfPreviousLoansBeforeLoan2	-1,05E-08	2,60E-09	5,09e-05 ***	-5,10E-09	1,47E-09	0,000540 ***
DebtToIncome2	-1,21E-04	2,63E-05	4,19e-06 ***	-7,423e-05	1,52E-05	9,65e-07 ***
realInterest2	6,26E-01	1,31E-01	1,70e-06 ***	3,59E-01	7,60E-02	2,34e-06 ***
LoanDuration2	4,742e-04	1,871e-04	0,011257 *	2,546e-04	1,051e-04	0,015422 *
ExistingLiabilities2	-6,98E-03	1,52E-03	4,06e-06 ***	-4,01E-03	8,670e-04	3,70e-06 ***
EducationVocational_educationalXIsHarju	_	_	_	2,06E-01	1,013e-01	0,041744 *
AIC		**10121**			**10098**	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Model evaluation

Linktest

We perform linktest to check whether final model has appropriate specification. We regress fitted values of probit model raised both to first and second power on dependent variable, GainLoss. The result of linktest can be found in Table 6.

Table 6. Linktest

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)
Constant	0,002742	0,019691	-0,139	0,889
yhat	0,550138	0,022839	24,088	<2e-16 ***
yhat2	0,013194	0,009612	1,373	0,17

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

With Z-value equals to 1.373 and p-value > 0.5, we reject null hypothesis that fitted values raised to second power have any explanatory power. Thus, we conclude that the final model is specified correctly.

Table 7. Hosmer and Lemeshow goodness of fit (GOF) test

X-squared	df	p-value
24,884	8	0,00164

From Hosmer and Lemeshow Test, we receive X-squared statistic equals to 24.89 and corresponding p-value equals to 0.00162. We reject null hypothesis that model is well fitted.

However, statisticians argue whether Hosmer and Lemeshow Test can be treated as state-of-the-art test in case of goodness of fit (Allison, Brown et al). Instead, they propose an alternative to Hosmer and Lemeshow test - Osius and Rojek test, the more stable one, that we performed as well.

Table 8. Osius and Rojek goodness of fit (GOF) test

Z	p-value
-0,2204812	0,825

In that case, we cannot reject the null hypothesis that data is well fitted in our model.

R2 Statistics

Table 9. R2 statistics

McKelvey. Zavoina	Count R2	Adj. Count R2
0,33	0,80	0,14

McKelvey-Zavoina statistic indicate how much variation of class variable is explained by the model. In our case, if latent variable is observed, descriptive variables explain 30% of variation.

What is more, based on Count R2 (the accuracy) we can say that the model correctly assigned 80% of observations to deliver gain or loss. However, as stated previously, in case of imbalanced data, accuracy can be misleading. Instead, we use Adjusted Count R2 that is more robust than accuracy. It shows the ratio of correctly assigned observation resulting from model's explanatory power - in our case Adjusted R2 = 12%. It indicates rather poor performance. We can conclude that investor's gain or loss on Bondora peer-to-peer lending platform cannot be fully explained by borrowers characteristics. There is a random noise in data that can include early repayments resulted from unpredictable events, current situation on labour market, performance of domestic economy and other.

V. RESULTS

Marginal effects

For providing a good approximation to the amount of change in GainLoss that is produced by a 1-unit change of explanatory variable, we calculate Marginal effects (Table 5). Because of specific characteristic of probit model, that the marginal effects depend on initial value of variable, we have chosen to explain marginal effects based on mean of each numerical variable (Marginal Effect at Mean method). All the conclusion are presented for average borrower in dataset.

Table 10. Marginal effect

	dF/dx	Std. Err.	z	P>|z|
Amount	-4.7097e-05	6.7243e-06	-7.0039	2.489e-12 ***
IncomeTotal	7.7597e-06	1.3387e-06	5.7966	6.767e-09 ***
AmountOfPreviousLoansBeforeLoan	-3.4279e-05	4.8486e-06	-7.0699	1.550e-12 ***
realInterest	-3.2125e-01	4.9350e-02	-6.5097	7.531e-11 ***
PreviousRepaymentsBeforeLoan	6.4056e-05	7.1923e-06	8.9062	< 2.2e-16 ***
LoanDuration	-5.6479e-03	1.4997e-03	-3.7660	0.0001659 ***
ExistingLiabilities	1.9970e-02	3.4404e-03	5.8046	6.453e-09 ***
CountryES	-3.6665e-01	1.5777e-02	-23.2389	< 2.2e-16 ***
CountrySK	-5.2674e-01	5.1887e-02	-10.1516	< 2.2e-16 ***
LanguageCodeFinnish	-2.7598e-01	1.9130e-02	-14.4267	< 2.2e-16 ***
LanguageCodeOther	-6.0350e-02	1.6293e-02	-3.7041	0.0002122 ***
GenderUndefined	-7.2141e-02	2.3376e-02	-3.0861	0.0020281 **
HomeOwnershipTypeHomeless	-6.4518e-01	5.4191e-02	-11.9057	< 2.2e-16 ***
HomeOwnershipTypeMortgage	5.4870e-02	1.2524e-02	4.3811	1.181e-05 ***
EducationHigher_education	7.1116e-02	9.0210e-03	7.8834	3.186e-15 ***
EducationSecondary_education	5.8992e-02	9.0460e-03	6.5214	6.965e-11 ***
WorkExperience15To25Years	-3.2766e-02	1.0488e-02	-3.1241	0.0017833 **
WorkExperience2To5Years	-5.0406e-02	1.1674e-02	-4.3180	1.575e-05 ***
WorkExperienceLessThan2Years	-5.5078e-02	1.6413e-02	-3.3558	0.0007914 ***
NrOfDependants2	2.7761e-02	1.0640e-02	2.6091	0.0090792 **
UseOfLoanLoan_consolidation	3.4824e-02	1.0317e-02	3.3755	0.0007369 ***
UseOfLoanOther	1.9032e-02	9.0765e-03	2.0968	0.0360076 *
UseOfLoanVehicle	3.5413e-02	1.1976e-02	2.9569	0.0031072 **
EducationVocational_educationXIsHarju	4.9006e-02	2.1694e-02	2.2590	0.0238852 *
IsHarjuXExistingLiabilities	4.0884e-03	1.8102e-03	2.2585	0.0239146 *
Amount2	3.0739e-09	7.8631e-10	3.9093	9.256e-05 ***
AmountOfPreviousLoansBeforeLoan2	-1.3305e-09	3.8382e-10	-3.4663	0.0005276 ***
DebtToIncome2	-1.9384e-05	3.9609e-06	-4.8939	9.886e-07 ***
realInterest2	9.3740e-02	1.9840e-02	4.7249	2.303e-06 ***
LoanDuration2	6.6501e-05	2.7455e-05	2.4222	0.0154264 *
ExistingLiabilities2	-1.0477e-03	2.2610e-04	-4.6338	3.590e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

FINDINGS

First of all, we will check whether results met hypothesis stated previously.
However, before interpreting model, it is important to state base levels of binary variables.

Table 11. Base levels

Variable	Base level
Country	EE
LanguageCode	Estonians
Gender	Man
HomeOwnershipType	Council house
Education	Basic education
WorkExperience	10To15Years
MaritalStatus	Cohabitant
NrOfDependants	0
UseOfLoan	Business

Based on the final results, we can indeed confirm that the higher amount of issued loan, the lower possibility of gain - an increase of amount of average loan by one euro lowers the probability of investor's gain by 0.005 percentage points. It is very intuitive finding and in fact can support the economics 'no free lunch' theorem. What is more, the finding can push investors towards greater diversification of loans portfolio, given that investing in one bigger loan is riskier than investing in multiple smaller ones.

However, we reject our second hypothesis that borrowers from relatively rich region has greater probability of paying off the debt. The binary variable IsHarju is not significant in our model. Though, the interaction between region variable and amount of existing liabilities of borrower is statistically significant. One Euro increase of existing liabilities for borrower living in Harju increases the probability of gain for investor by 0.41 p.p.

Furthermore, we present the most interesting findings about variables that were not part of main or second hypothesis. An increase in average total income by one euro increases the probability of investor's gain by approximately 0.0008 p.p. Importantly, an increase of average interest rate by one percent decreases the probability of investor's gain by 32.13 p.p. Moreover, we see that relationship between class variable and AmountOfPreviousLoansBeforeLoan is not linear - both AmountOfPreviousLoansBeforeLoan and AmountOfPreviousLoansBeforeLoan ^ 2 are

statistically significant . When change in amount of previous loans before loans is positive, the sign of relationship is negative. However, in the case when the change is negative, the less loans borrower had, the greater probability of gain investor has, until the change in amount of previous loans of average borrower is greater than 102. Then, the relationship changes its sign.

What is more, on average, investing in loan gotten by borrower form Spain decreases the probability of investor's gain by 36.67 p.p. compared with investing in Estonian's loan. Simillary, investing in loan gotten by borrower form Slovakia decreases the probability of investor's gain by 52.67 p.p. Moreover, those borrowers who did not state their gender are less likely to provide a profit for investor by 7.21 p.p. compared with male borrowers. On average, lending money to someone who speaks Finnish as his / her main language decreases the probability of gain by 27.6 p.p. compared with borrower who speaks Estonian. What is more, lending money to borrower who is homeless decreases probability of gain by 64.52 p.p. on average. The interesting issue is how the Bondora's scoring model let homeless people borrow money- probably, the homeless state in that case means that someone does not own house and lives with family, friends, or rents a flat. That thesis can be supported by homeless borrowers relatively high values of total income. What is more, investing in loan of borrower with mortgage increases the probability of gain by 5.49 p.p. on average compared with borrower with council house. It is not surprising, that the work experience in most cases has negative impact on investor's profit. Lending money to borrower with work experience less than 2 years decreases the probability of investor's gain by 5.51 p.p. Lending money to borrower with work experience from 2 to 5 years decreases the probability of investor's gain by 5.04 p.p. Surprisingly, lending money to someone with work experience from 15 to 25 years decreases the probability of investor's gain by 3.28 p.p. compared with the one with 10 to 15 years of experience. Finally, the borrower with secondary level of education delivers gain to investor with probability greater by 5.9 p.p. than the borrower with basic level of education. Lending money to borrower with higher level of education comparing to the borrower, who has primary level of education, increases investor's probability of gain by 7.11 p.p. Again, latter findings are intuitive.

We do not provide odd ratios for probit model because they cannot be interpreted.

Summary

We analysed peer-to-peer lending market based on example of Bondora, leading lending marketplace in Estonia. We performed logit and probit regressions and based on AIC information criterion we selected probit model as better fitted. Then, we evaluated model's goodness of fit using statistic tests such as Linktest, Hosmer and Lemeshow test and Osius and Rojek test. In two of three cases we concluded that model is well specified. Only Hosmer and Lemeshow test, said to be not stable, showed that our model should be improved. What is more, we have found that many variables can significantly affect borrower's ability to pay the loan. Among obvious factors as Amount of loan or interest rate, we have found that investing in loans gotten by borrowers from Spain and Slovakia in fact lower investor's probability of gain. It can be possibly explain by the fact that majority of transactions in Bondora platform are made by people from Estonia, who could prefer investing in local loans.

Future improvements of our works may include collecting of economy factors that can result in improvement or deterioration of borrower's situation. What is more, one can adapt regularization methods such as Lasso or L2 to simplify model.

References

D. W. Hosmer and N.L. Hjort. Goodness-of-fit processes for logistic regression: Simulation results. *Statistics in Medicine 21*:2723–2738. 2002. https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.1200

G. Brown, A. Pocock, M. Zhao, M. Lujan. Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection. *The Journal of Machine Learning Research*. 13. 27-66. 2012. http://www.jmlr.org/papers/volume13/brown12a/brown12a.pd

G. Osius, G. and D. Rojek. Normal goodness-of-fit tests for multinomial models with large degrees-of-freedom. *Journal of the American Statistical Association 87*: 1145–1152. 1992. https://www.tandfonline.com/doi/abs/10.1080/01621459.1992.10476271

I. Galloway. Peer-to-Peer Lending and Community Development Finance. Community *Development Investment Center Working Paper*. San Francisco: Federal Reserve Bank of San Francisco. 2009. https://www.frbsf.org/community-development/files/galloway_ian.pdf

J. Jagtiani and C. Lemieux. Fintech Lending: Financial Inclusion, risk pricing, and alternative information. *Federal Reserve Bank of Philadelphia Working Paper No. 17-17*. Philadelphia. 2017

M. Herzenstein, R. Andrews, U. M. Dholakia, E. Lyandres. The democratization of personal consumer loans? Determinants of success in online peer-to-peer lending communities. 2008. https://pdfs.semanticscholar.org/127f/3b1106835ccee15d2cdf1e1099cf699247cd.pdf?_ga=2.151429945.14009074.1524596125-1759071529.1524596125

P. Allison. Alternatives to the Hosmer-Lemeshow Test. 2014. https://statisticalhorizons.com/alternatives-to-the-hosmer-lemeshow-test

S. Ceyhan, X. Shi, J. Leskovec. Dynamics of bidding in a P2P lending service: effects of herding and predicting loan success. 2011. https://www-cs-faculty.stanford.edu/people/jure/pubs/prosper-www11.pdf

S. Berger and F. Gleisner. Emergence of financial intermediaries on electronic markets: The case of online p2p lending. *Business Research*, 2(1), 39-65 2009.

https://link.springer.com/content/pdf/10.1007%2FBF03343528.pdf

M. Hulme and C. Wright. Internet based social lending: Past, present and future. *Social Futures Observatory*, 2006.

R. Tibshirani. Regression Shrinkage and Selection via the lasso. *Journal of the Royal Statistical Society*. Series B (methodological). Wiley. **58** (1): 267–88.1996.

https://www.jstor.org/stable/pdf/2346178.pdf?refreqid=excelsior%3Ab876afcbe907722716f78160f291072d

S. Freedman and G. Jin. Do social networks solve information problems for peer-to-peer lending? Evidence from prosper.com. *Working Paper*. 2008. http://www.fas.nus.edu.sg/ecs/events/seminar/seminar-papers/12Mar09.pdf

S. Freedman and G. Jin. The information value of online social networks: lessons from peer-to-peer lending. *NBER Working Paper No. 19820*. 2016. http://kuafu.umd.edu/~ginger/research/freedman_jin_06222016.pdf

S. Garman, R. Hampshire, R. Krishnan. A theoretic model of person-to-person lending. 2008. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.574.1877&rep=rep1&type=pdf

Y. Zhang, H. Li, M. Hai, J. Li, A. Li. Determinants of loan funded successful in online P2P Lendig. *Elsevier B.V.* 2017. https://ac.els-cdn.com/S187705091732700X/1-s2.0-S187705091732700X-main.pdf?_tid=50849aa7-f23a-4fc6-8bfa-5d8cc827acc4&acdnat=1524836638_ef403f01858f067f2240c623a97facf3

Y. Lu, B. Gu, Q. Ye, Z. Sheng. Social Influence and Defaults in Peer-to-Peer Lending Networks. 2012. https://pdfs.semanticscholar.org/6ae6/2cf3c2270b5f3f50e58a805f6ca202fed7bf.pdf?_ga=2.215443223.14009074.1524596125-1759071529.1524596125