International Core Journal of Engineering Volume 7 Issue 11, 2021
ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0051

Income Stability and Default Risk Prediction on Online Credit Market

Yi Li

The Chinese University of Hong Kong, China

Abstract

On the rise of fintech development on online credit, there are numerous efforts to enhance the performance of the platforms' default risk model. One of the important criteria is the income from borrowers. Relying on detailed data from a major European marketplace lending platform, we infer borrowers' income stability and use this information to improve credit risk model performance. Applying machine learning techniques, we find that the income stability measure is negatively related to default probability. It serves a significant role to improve the performance of the benchmark model used by the platform: the AUC increased from 0.69 to 0.73 after including the income stability measure. The results hold in out-of-sample test and when using more precise sub-sample.

Keywords

Machine learning, credit risk model, income from borrowers.

1. Introduction

With the emerging of web 2.0, a new credit source become available: called online market place lending platforms. Up to now, there exist more than a dozen commercial operators in the P2P lending marker in Europe. (Ashta & Assadi, 2009) The collapse of the financial system in 2008 and the high interests issued by traditional banks have driven numerous small business and individuals to find alternative finance support. In this convoluted context, P2P lending companies have emerged in. Although P2P lending has been around for a lot less time than traditional bank loans, it has been growing dramatically since 2010: from its began in 2010, P2P transactions reached $80 billion by 2015 and continue to grow exponentially. (Aveni et al., 2015) Peer-to-peer lending is used to describe online marketplaces where lenders can lend to individuals or small businesses. Since it started in 2005, there seem dozens of companies over the world, especially the well-known Lending club and Prosper. (Anshari, Almunawar, Masri, & Hrdy, 2021) These days, marketplace lending is a more fitting term for evolving business, so marketplace lenders have focused on unsecured consumer credit as the main predictor. ("Can P2P Lending Reinvent Banking?," 2015)

Therefore, the issue that evaluating credit risk is material should be concerned due to the attention on unsecured consumer credit by lenders. One of the explanations to finance crisis on 2008 is large wave of credit default. So, credit risk management is relevant to main finance stability. Moreover, it is comparable on marketplace lending platform, which is growing and becoming very relevant as a new source of financing for individuals and startups. These credit loans are always lack of collateral, and since individuals do not have professional skills to screen the bad lone out of them, it will be damaged if such management is absent. Furthermore, when going over the previous models of predicting the default, such as KMV and Credit Risk+, one could conclude that credit risk is the inevitable variable to predict default.(Crouhy, Galai, & Mark, 2000)

Literature has pointed out that total income is conducive to help enterprises and online platform set different credit risk level for borrowers, for example, 'A-G' credit grades. (Möllenkamp, 2017) In

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0051

some cases, researchers document a negative relationship between aggregate income and default. (A. Rampini, 2005) However, total income is the sum of income from multiple income sources. Aggregating all the income can mask away important information imbedded in different sources. Different components of income such as income from principal employers, income from child support, income from family allowance, income from leave pay, income from pension, and income from social welfare, contain information indicating the stability of borrowers' cash flow. As the previous paper statement, income stability is the main factor to predict default probability. (McCollum & Pace, 2017) Therefore, income sources could provide additional useful information to improve the performance of platforms or enterprise model, when we decompose the total income into different subgroups.

Following the above discussion, in this paper, we conjecture that ratio of income principal and ratio of income alternative can improve default risk prediction. We use the capability of individuals' income coming from different income sources to measure income stability. By introducing such new concept, we conjecture that income stability could function as a supplement role to help online platform and enterprise to better evaluate credit risk of borrowers. We use the data coming from Bondora, a P2P lending platform, which is on the rise and has become a prominent platform in Europe, in order to prove our view.(Bondora, 2021b)

According to the criterion that whether income is direct or not, we transform different kinds of income into two categories: income from principal and income from alternative. Then, we proceed by discussing the key economic outcomes and implications of our findings. We take advantage of data from Bondora which provides information on detailed income sources of borrowers. We use income from principal and income from alternative to measure income stability and use them to predict credit default risk. To start with, we introduce one mainstream machine learning algorithm and a supplement one (logistic regression and random forest) and one criterion (Area under the curve) (The AUC ranges from 50% (pure random prediction) to 100% (perfect prediction). It is widely use to measure discriminatory power of credit scores. (Berg, et al 2019)) to assess model performance. First, results from logistic regression show that there is a significant relation between income stability and default dummy. Including income stability improves the prediction of default risk by raising the ROC from 0.6905 to 0.7323. Similar results are also found in the out-of-sample test: The ROC area increases from 0.6906 to 0.7332. In addition, only the verified data to fit logistic regression model, I find that model performance improves: the ROC area is 0.7506. (The verified data come from Bondora platform, which mean that the authenticity is verified by Bondora platform.) This is probably due to the increased precision of the data. The main results are robust also when we include different sets of control variables. Finally, resulting in above examination, the Roc scores increase from 0.76 to almost 0.79 if we use random forest, although the algorithm is assembled and some parameters of the algorithm are fixed.

Above the presentation, the main contribution made by this paper is that income stability as a set of variables transformed by original data complements rather than substitutes for the prediction default.

Our thought originated from some comparative views that by evaluating individuals' certain information can help banks lower their exposure to borrowers credit risk Traditionally, banks relied on seasoned senior to tune both crediting criteria and standers. And after the finance crisis in 2008, the evaluation method has changed into more digital and internet-related ways. (Imbierowicz & Rauch, 2014; Treacy & Carey, 2000)

Furthermore, this paper is relevant to credit risk studies based on digital platform, where researchers take advantages of the comprehensive data on borrowers to improve prediction. Researchers have investigated the increased predictive power using digital footprint, checking account activities, dividing into different short period and some commonly used further control variables, such as gender, age, etc. (Iyer, Khwaja, Luttmer, & Shue, 2016; Khandani, Kim, & Lo, 2010; Puri, Gombović, Burg, Berg, & Karolyi, 2020) We contribute to this strand of literature by focusing on income sources, which are the representativeness of income stability.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0051

Some traditional statistical methods such as discriminant analysis and linear regression model are not appliable to many distributions, say, Gamble, normal, Poisson and so on. Whereas, logistic regression model is more suitable for credit scoring problems and binary problem, which is suitable for prediction default. (Ong, Huang, & Tzeng, 2005) In the end, referring to the concept of income stability, we combine this concept into our specific model to transform income total into a set of variables related to income. The goal of us is further improving the performance of prediction default model in case of previous fix control variables. (Peng Hongfeng, & Ye Yonggang, 2011)

The remainder of this paper is structured as follows. The second part illustrates the descriptive statistic and the basic methodology regarding data selection. The third part introduces the algorithms that we use in this paper. In part four, the completed model and conclusion are discussed and we also do some external validation. Then in part five, we draw some conclusions based on part four and give some expectations.

2. Data

2.1 Introduction of Bondora platform

Bondora is an online P2P lending platform, which has established in 2008 with 45 employees, has become one of the major lending markets in Europe. Being an Estonian platform, it focus on Finland, Spain, and Estonia markets, now, it has processed more than EUR 1.4 billion of loan applications and issued over EUR 74 million in loans. In comparison to its peers, Bondora has a solid track record in generating returns for investors. What's more, Investors can generate 400 to 1000 basis points of additional yield, depending on borrower credit score grades, without noticeably increasing risk. (Bondora, 2021a)

2.2 Data selection

We select the dataset from Bondora platform, which includes 167,513 observations all came from EU-countries over February 2009 till April 2021. Then, the illogical observations are dropped: age is less than 18 years old or greater than 70, income is negative or some attributes' value that can not be accessible by scrutinizing data dictionary. After the initial screening process, we contain 164,389 observations.

Although Bondora provides over one hundred attributes for lenders to consider, we have only two variables as main variables and five variables (further we are going to transform two of them into dummy variables) as additional control. Besides, we are going to use some attributes not considered to be variables but as part of our evidence to explain the result. We regard as benchmark the probability that Bondora has had and transform the Default date into binary due to the requirement of logistic regression. The methodology is that if the original observation of Default data is not absent, we will regard it as default, vice versa. Finally, for robustness, we introduce verification type variable, which symbolize the data are verified by Bondora platform.

For main variables, considering that we should not identify what the exact income values are but utilize the ratio of them to represent the magnitude of both principal income and alternative income, we divide the two by income total so that get the number between zero and one. The relationship between the two variables and income stability conform to our common sense: the higher proportion of income principal means one has a legitimate and stable income and the higher proportion of income alternative means one has multiple finance supports. However, the value of total income should not be dismissed, sine the total income to some extent reflects individual's overview finance performance that ratio income is not able to reveal it. Thus, taking the weight into consideration, we use log function to transform the total income variable as one of additional control. For example, assuming one has high ratio of income principal, but it does not mean the total income he earns is enough to afford the debt, only by having a high total income at the same time can the debt is affordable for him. Just to brief summarize the above, although income total is not able to represent individual's income

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0051

stability, it can be a complement variable to income stability that includes ratio of income principal and income alternative.

In addition to age and gender is commonly deemed basic to model performance, we introduce home ownership type and employment duration current employer for further control. On the one hand, home ownership type straightforward shows the individual's fix asset to lenders on Bondora platform, a good home ownership type record makes borrower more convincing comparing to those who do not have an estate. It also signals that people with estate do not want to be default only in emergency. On the other hand, one with a stable work station usually symbolizes that he or she has a stable income. Because there are different kinds of home ownership type and employment duration current employer type, we divide these into dummy variables in order to get classified variables. Taking into account that the nature of the home ownership type and the personal property it represents, we transformed such types as home ownership type with finance risk including homeless, mortgage and owner with encumbrance, home ownership type with other including minority others, home ownership type with tenant including pre-furnished property, unfurnished property, council house and joint tenant property, home ownership type with joint property including living with parents, and home ownership type with property including owner and joint owner. So does Employment Duration Current Employer. It is transformed into trail, other retire and the time from up to 1 years to more than 5 years.

2.3 Descriptive statistic

Through the Appendix A: Table 1, we can draw some empirical conclusions. The average age is about 41 years old, which is consider solvent and able to work. Moreover, the mean of income total is €1,627, and it is lower than the average level of EU, which is about €2,200. ("Eurostat," 2021) The phenomenon is self-evident since P2P platforms service underbanked clients (Tang, 2019) and those who tentative have to take out online loans for certain justifications. And because of the low average probability of default, 0.24, it shows that most of people do not want to default.

The Table 2 shows the correlation table for all variables we used. After the transformation, all of them illustrate that there is no heavy correlation between any two variables: all correlations are lower than 0.4 means there are no variables should be dropped according to the table.

3. Methodology

We introduce two machine learning methods to explore the predicted results. To begin with, we introduce logistic regression, since the dependent variable, default, is binary variable, and logistic regression is obvious robustness when it comes to binary rather than continues variable. Then, we use random forest as a supplement test to justify our conclusion is suitable for other test methods. The reason is that it is difficult to say which one is superior to the other, considering numerous situations. Using logistic regression should follow some preconditions, it just like a prior assumption. Whereas, random forest test is like a black-box test, because we cannot clearly to explain the specific meaning of the coefficient of variables. By contrast, although logistic regression is more complicated comparing to random forest, we can simply justify the meaning of such coefficients. Therefore, we decide to use logistic as main justification and random forest as supplement one.

3.1 Logistic regression

The central mathematical concept that underlies logistic regression is the logit—the natural logarithm of an odds ratio. generally, logistic regression is well suited for describing and testing hypotheses about relationships between a categorical outcome variable and one or more categorical or continuous predictor variables. (Peng, Lee, & Ingersoll, 2002)

In this paper, we use four-stage logit model to verify our assumption. After following the prerequisite of logistic model, the data are fitted by the logistic regression model we set before. The principle behind this algorithm is maximum likelihood estimation. According to the criteria of confusion matrix (Each row of the matrix represents the instances in an actual class while each column represents the instances in a predicted class, or vice versa – both variants are found in the literature) and the results

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0051

came from our model, we summarize the predicted true and false value, and the actual default or not value, then judge whether the prediction results are up to our expectation.

The first model is used for benchmark, which only includes probability of default calculated by Bondora. The second model contains our main variables, which symbolize the income stability. The third model is the combination of benchmark and main variables, the goal of this method is to justify that income stability can play a role of improving the model's performance. The fourth model adds further control variables on the basis of model 3 so as to help verify our main variables can be a supplement role to improve the model performance. The extended versions of the last 3 models can be founded in Appendices: equations 7, 8 and 9.

$$(1)\ \boldsymbol{y}(\boldsymbol{default}) = \beta_0 + \beta_1(Probability) + \varepsilon_i$$

$$(2)\ \boldsymbol{y}(\boldsymbol{default}) = \beta_0 + \beta_1(Income\ stability) + \varepsilon_i$$

$$(3)\ \boldsymbol{y}(\boldsymbol{default}) = \beta_0 + \beta_1(Probability) + \beta_2(Income\ stability) + \varepsilon_i$$

$$(4)\ \boldsymbol{y}(\boldsymbol{default}) = \beta_0 + \beta_1(Probability) + \beta_2(Income\ stability) + \beta_3(Further\ control) + \varepsilon_i$$

3.2 Random Forest classification

The concept of random forest raised in 2001 is an ensemble algorithm based on decision tree. The generalization error for forests converges to a limit as the number of trees in the forest becomes large. Because of this nature of random forest, random forest not only is robust to noise, but also popular in lost of areas such as industry and research.(Breiman, 2001)

To begin with, decision tree has a tree-like structure and it has node to let tree split accord dichotomy criterion. What's more, users also are able to prune the decision tree such as restrict it to certain layers and circumscribe it to limit number of nodes, etc. Through the concept of impurity, we can find how to branch the decision tree. The methodologies to measure the impurity are Gini and Entropy.

$$(5)\boldsymbol{Entropy} = -\sum_{j=1}^{j} p_j \log_2 p_j$$

$$(6)\boldsymbol{Gini} = 1 - \sum_{j=1}^{j} p_j^2$$

In these two formulas, pj is the probability of class j. In each branching, the decision tree calculates the impurity of all the features and selects the feature with the lowest impurity for branching. After branching, it calculates the impurity of each feature for different values of branching and continues to select the feature with the lowest impurity for branching. However, decision trees are very easy to overfit, in order to bypass this shortcoming, random forest is introduced.

Random forest is a representative bagging algorithm, whose classifiers are all decision trees. Like the decision tree, which use homogeneity to judge the category of variables, random forest also use this methodology not only in single trees, but also represent the category of variables based on all trees. It means that we would refer the majority of trees classification as the final results, which ensure the robustness to bias sample. Therefore, random forest has the criterion n_estimators to help users to control the number of trees they set. Furthermore, in order to make the base classifier as different as possible, it is easy to understand that different training sets are used for training. The bagging method is to form different training data through random sampling technique with back. Bootstrap is used to control the parameters of the sampling technique. (Biau & Scornet, 2016; Shi & Horvath, 2006) But, although random forest has such advantages to use, it is hard to translate the meaning of coefficients of variables, so we just use this algorithm as a supplement way to verify our assumption.

4. Results

4.1 Income stability

We summarize the four regression results, which separately include the benchmark, our main control variables, the combination and further control containing both categorical and continuous variables.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0051

We use the logistic regression algorithm and AUC for the four specifications. AUC (Area under the curve) is commonly criterion for evaluating the discriminatory power of credit risk.(Stein, 2002)

In the use of experience of AUC, it has poor performance or, say, purely random prediction, if the AUC score is 50 percent; by contrast, if the AUC score is 100 percent, it has perfect prediction, but this situation is idealization. In practice, when the AUC is above 70 percent, it means that there is 70 percent chance that the model will be able to distinguish between positive class and negative class (Narkhede, 2018), and this model keeps in information rich circumstance.

Starting with Column 1 of table 3, it shows that we just use the probability of default as the only independent variable. It is because such particular attribute calculated by Bondora platform to predict the default borrowers. Therefore, we regard the result of Column 1 as benchmark. The result is straightforward, the coefficient of probability of default is positive, and this result is verified by Z-test (Since most random variables in practical problems obey or nearly obey normal distribution, we call this test method using statistics obeying standard normal distribution z-test, or u-test.), which means this single variable is a highly significant predictor of default. The more probability of default the value is, the more default the borrower will be. This explanation is followed our common understanding. The AUC using probability of score alone is 0.6905 and is significantly different from purely random prediction. So, this suggests that we could use this model as the benchmark of Bondora platform.

The rest of the models are based on a comparison of the first model. Column 2 shows the main variables of income stability alone; based on Column 2, Column 3 combines the probability of default in order to justify the main variables of income stability could play as a supplement role to improve the model performance. Column 4 adds some further control variables, which are popular in default prediction trend, mentioned before in order to enhance the usefulness and credibility of our specification. We report AUCs in the bottom of Table 3 and test the differences in AUCs using method from this paper. (DeLong, DeLong, & Clarke-Pearson, 1988). What's more, all the coefficient value are kept two decimal places.

In column 2, borrower who has ratio income alternative, which is exp(0.79)=2.20, means he or she has about 2 times higher to be default than who does not have it. Since people with more ratio income taking their income show they do not have ability to afford themselves. Besides, the Z-tests of column 2 are all significant, but the AUC is 0.5629, which is less than the AUC of Column1. Thus, we could draw that using the stability of income could only play a supplement role to improve the prediction ability. In Column 3, we use the combination of probability of default, ratio of principal and ratio of alternative. Not only do the coefficient of the main variables related income stability not change too much, but also the coefficient of probability does not so. And the AUC score is 0.7266, which overshadow the Column1 and Column 2. Therefore, the combination one also stresses the supplement role that our main variable play in. Finally, in Column 4, we add gender, age, income total transformation, home ownership type and employment current duration as our further control variables. The coefficients of variables remain almost unchanged, and the AUC improve a bit. So, this specification which is regarded as a robustness test endorses our conclusion that income stability variables are valuable supplement one as well.

4.2 Out-of-sample test

In-sample test is used for Table 3 to draw these above conclusions, but it maybe elicit overfitting problem, which has perfect performance in fitting process, but has bad performance in practice. Thus, we use out-of-sample test to verify our model. First, we set a random seed and give each of observations a specific random number. Second, we split them up by picking up even numbers, then, we randomly get 82,195 samples as our test dataset, and the remaining part as our training dataset.

Therefore, we estimate a predictive logistic regression using remaining part and use the coefficients to create predicted values for the 82,195 observations. We use this methodology for our four models in table 3 and get the corresponding specifications.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0051

Table 4 provides the results. It shows that all of the results of out-of-sample tests are almost the same as Table 3, even the differences are even a thousandth of a percent. These results show our sole purpose of showing that the specifications are not overfitting so that support our view of point.

4.3 Random Forest classification

Following the above discussion, we use random forest algorithm to conduct our experiment. As illustrates from Table 5, it summarizes the feature importance and AUCs for both all samples and verified samples by Bondora platform. The range of feature importance is from 0.00 to 1.00. according to the Table 5, the feature importance of our main variables are about 0.1, although they are less than the feature importance of probability of default, they are still useful to predict the default. What's more, the income total transformation variable, age and gender also important, which correspond the above statement that these variables are widespread used for further control. For home ownership type and employment duration, they take some places because they represent income stability to some extent even though the proportion is not high enough. However, the AUCs are higher than the AUCs of Table 3 and Table 4. Nevertheless, since the goal of random forest is feature selection and the full interpretation is kind of hard to enunciate due to the character of randomness. Thus, we just use it as a robustness justification to endorse our point that the main income stability variables can play a supplement role to improve prediction default.

5. Conclusion

Prediction default is essential for P2P lending platform. Although each P2P lending platform has its own set of criteria and algorithm to predict default, every single refinement in criteria or algorithm can contribute nonnegligible economic benefits. This study targets at income stability as the supplement role to the original algorithm of platform to more accurately predict borrower's default. In other words, platform that uses information from both the probability of default and the income stability variables can make superior lending judgement to provide for lenders. Processing more than 160,000 observations, we use two different classification methods and compare them in order to get the most accessible conclusion.

Our goal is justified by empirical results which show not only high significant test value, but also intuitive coefficient of variables. In this paper, the income stability of borrowers is mainly judged by ratio of income principal and ratio of income alternative. The coefficient of ratio of income alternative is positive and is not affected by model changes. It proxies for the individual's financial independence. If one's ratio of income alternative is very large, it means that he or she most likely does not have the ability to pay for debt alone. Therefore, such individuals likely seem to have bad loan. The coefficient of income principal is positive maybe due to the borrower's willingness to repay. Moreover, the random forest variable importance matrix confirms this by finding that the feature importance to measure income stability is high. In sum, through empirical analysis, our theory is consistent with the specifications.

Overall, our research introduces the concept of income stability to help improve the model of platform. For setting regulations, it is important to evaluate a person's credit status quo by measuring one's ratio of income principal and ratio of income alternative. regulators can depend on these two criteria to formulate a set of credit evaluation system, which can refrain from financial crisis to some extent. For online lending platforms, since the two variables were seldom mentioned in the previous studies, so if the enterprises reasonable and in-depth conduct to use them, they will enhance the accuracy of default prediction, so as to improve the applicability of their platforms. Eventually, this effect will attract more lenders to come by. For researchers, they can define different criteria to divide individual income through research and apply them to statistical models for further optimization.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0051

6. Appendices

This table show the summary statistics for the whole sample except categorical variables. We select the dataset from Bondora platform and after screening out this sample, we present 164,389 observations all came from EU-countries over February 2009 till April 2021.

Table 1. Descriptive statistics

Variable	Unit	N	Mean	SD	P25	Median	P75
Amount	1 = 1 Euro	164,389	2,571.30	2,177.69	744	2,125	3,825
Gender	Dummy(0=male,1=female)	164,389	0.47	0.63	0	0	1
Age	Number	164,389	40.62	12.32	31	39	50
Probability of default	Number	164,389	0.24	0.14	0.13	0.22	0.33
Ratio principal	Number	164,389	0.18	0.37	0	0	0
Ratio alternative	Number	164,389	0.03	0.13	0	0	0
Income total	1 = 1 Euro	164,389	1,627	7,614	860	1,200	1,816

This table shows the correlation of all variables including categorial and continuous variables. These numbers range from minus one to one: minus one means the two variables are perfect negative correlation and one means the two variables are perfect positive correlation.

Table 2. Correlation

	ProbabilityOf Default	Ratio Principal	Ratio alternative	Age	Gender	Income Total trans	Empdur morethan5	Empdur other	Empdur retiree	Empdur trial
ProbabilityOf Default	1.00	-0.04	-0.03	-0.07	0.09	0.15	-0.01	0.02	0.04	0.00
Ratio Principal	-0.04	1.00	0.10	-0.13	0.03	-0.06	0.00	-0.10	-0.12	0.07
Ratio alternative	-0.03	0.10	1	0.05	0.07	-0.05	0.01	-0.05	-0.05	0.07
Control variables										
Age	-0.07	-0.13	0.05	1.00	0.03	0.13	0.25	-0.05	0.34	-0.04
Gender	0.09	0.03	0.07	0.03	1.00	-0.06	0.03	0.04	0.02	0.00
IncomeTotal trans	0.15	-0.06	-0.05	0.13	-0.06	1.00	0.19	-0.07	-0.09	-0.04
Empdur morethan5	-0.01	0.00	0.01	0.25	0.03	0.19	1.00	-0.17	-0.20	-0.05
Empdur other	0.02	-0.10	-0.05	-0.05	0.04	-0.07	-0.17	1.00	-0.05	-0.01
Empdur retiree	0.04	-0.12	-0.05	0.34	0.02	-0.09	-0.20	-0.05	1.00	-0.02
Empdur trial	0	0.07	0.07	-0.04	0.00	-0.04	-0.05	-0.01	-0.02	1.00
Emdur1	0.02	0.01	-0.02	-0.24	-0.03	-0.08	-0.37	-0.10	-0.12	-0.03
Emdur2	0.02	0.27	0.10	-0.09	0.00	-0.05	-0.16	-0.04	-0.05	-0.01
Emdur3	0	0.23	0.10	-0.07	-0.01	-0.04	-0.14	-0.04	-0.04	-0.01
Emdur4	0	0.20	0.08	-0.04	0.00	-0.02	-0.12	-0.03	-0.04	-0.01
Emdur5	0.04	-0.18	-0.07	-0.14	-0.03	-0.01	-0.44	-0.12	-0.14	-0.04
ho_finacerisk	0.01	0	-0.01	0.06	-0.01	0.16	0.13	-0.03	-0.04	-0.01
ho_jointproperty	0.10	0.05	-0.02	-0.30	0.03	-0.17	-0.12	0.00	-0.06	0.02
ho_other	0.01	-0.13	-0.06	0.01	0.04	-0.04	-0.03	0.08	0.04	-0.02
ho_property	0.16	-0.01	0.03	0.25	-0.03	-0.02	0.12	-0.05	0.03	-0.02
ho_tenant	0.10	0.05	0.02	-0.07	0.00	0.06	-0.11	0.03	0.03	0.03

Table 3 Correlation — continue

	Emdur1	Emdur2	Emdur3	Emdur4	Emdur5	ho_finacerisk	ho_jointproperty	ho_other	ho_property	ho_tenant
ProbabilityOf Default	0.02	0.02	0.00	0.00	-0.04	0.01	0.10	-0.01	-0.16	0.10
Ratio Principal	0.01	0.27	0.23	0.20	-0.18	0.00	0.05	-0.13	-0.01	0.05
Ratio alternative	-0.02	0.10	0.10	0.08	-0.07	-0.01	-0.02	-0.06	0.03	0.02
Control variables										
Age	-0.24	-0.09	-0.07	-0.04	-0.14	0.06	-0.30	0.01	0.25	-0.07
Gender	-0.03	0.00	-0.01	0.00	-0.03	-0.01	0.03	0.04	-0.03	0.00
IncomeTotal trans	-0.08	-0.05	-0.04	-0.02	-0.01	0.16	-0.17	-0.04	-0.02	0.06
Empdur morethan5	-0.37	-0.16	-0.14	-0.12	-0.44	0.13	-0.12	-0.03	0.12	-0.11
Empdur other	-0.10	-0.04	-0.04	-0.03	-0.12	-0.03	0.00	0.08	-0.05	0.03
Empdur retiree	-0.12	-0.05	-0.04	-0.04	-0.14	-0.04	-0.06	0.04	0.03	0.03
Empdur trial	-0.03	-0.01	-0.01	-0.01	-0.04	-0.01	0.02	-0.02	-0.02	0.03
Emdur1	1.00	-0.09	-0.09	-0.07	-0.26	-0.07	0.11	0.01	-0.10	0.06
Emdur2	-0.09	1.00	-0.04	-0.03	-0.11	-0.02	0.04	-0.06	-0.02	0.03

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0051

Emdur3	-0.09	-0.04	1.00	-0.03	-0.10	-0.01	0.03	-0.05	0.00	0.02
Emdur4	-0.07	-0.03	-0.03	1.00	-0.08	0.00	0.02	-0.04	-0.01	0.02
Emdur5	-0.26	-0.11	-0.10	-0.08	1.00	-0.04	0.02	0.03	-0.01	0.00
ho_finacerisk	-0.07	-0.02	-0.01	0.00	-0.04	1.00	-0.15	-0.10	-0.28	-0.22
ho_jointproperty	0.11	0.04	0.03	0.02	0.02	-0.15	1.00	-0.12	-0.34	-0.26
ho_other	0.01	-0.06	-0.05	-0.04	0.03	-0.10	-0.12	1.00	-0.22	-0.17
ho_property	-0.10	-0.02	0.00	-0.01	-0.01	-0.28	-0.34	-0.22	1.00	-0.49
ho_tenant	0.06	0.03	0.02	0.02	0.00	-0.22	-0.26	-0.17	-0.49	1.00

we estimate default rate regressions where the dependent variable (Default(0/1)) is equal to one if such variable has specific value, which means the borrower has default. Column 1 provides results using the probability provided by Bondora platform; Column 2 provides results using transformed income variables as main variables; Column 3 provides results using the combination between Column 1 and Column 2; resulting the Column 3, Column 4 adds further control variables.

Table 4. Default regressions

	(1) Probability of default		(2) Ratio of income		(3) Probability of default & Ratio of income		(4) Probability of default & Ratio of income, further controls	
Variables	Coef.	z-stat	Coef.	z-stat	Coef.	z-stat	Coef.	z-stat
Probability of default	5.20***	116.39			5.70***	114.20	5.46***	106.54
Ratio of income principal			0.77***	56.23	1.00***	63.26	1.01***	56.66
Ratio of income alternative			0.79***	20.95	1.13***	26.45	1.06***	24.60
Constant	-1.63***	-135.55	-0.52***	-91.67	-1.95***	-142.25	-2.89***	-40.13
Control for age, gender, income, homeownership type, and employment current duration type	No		No		No		Yes	
Observations	164,389		164,389		164,389		164,389	
Pseudo R2	**0.0819**		**0.0175**		**0.1074**		**0.1151**	
AUC	**0.6905**		**0.5629**		**0.7266**		**0.7323**	
(SE)	**(0.0013)**		**(0.0010)**		**(0.0013)**		**(0.0012)**	
Difference to AUC=50%	**0.1905*** **		**0.0629*** **		**0.2266*** **		**0.2323*** **	
Difference AUC to (1)			**-0.1276*** **		**0.0361*** **		**0.0418*** **	

This table shows robustness test for results from Table 3 by presenting AUC scores. Column 1 shows the base line whereas Column 2 represents out-of-sample estimation.

Table 5. Out-of-sample estimates

	(1) Baseline(in-sample)	(2) Out-of-sample
AUC Probability of default	0.6905	0.6906
N	164,389	82,195
AUC Ratio of income alternative	0.5629	0.5653
N	164,389	82,195
AUC Probability of default& Ratio of income alternative	0.7266	0.7276
N	164,389	82,195
AUC Probability of default& Ratio of income alternative, further controls	0.7323	0.7332
N	164,389	82,195

This table shows the feature importance of verified and all samples, besides, all numbers are reserved for two decimal places. And at the last row, it represents the AUC score, which is the valuation of the model performance.

Table 6. Random forest summary

Variable name	Feature importance	Feature importance (verified sample)
ProbabilityOfDefault	0.45	0.46
Ratio_Principal	0.07	0.06
Ratio_Alternative	0.02	0.02
IncomeTotal_trans	0.18	0.17
Age	0.13	0.13

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0051

Gender	0.06	0.06
HomeOwnershipType_used_FinaceRisk	0.01	0.01
HomeOwnershipType_used_JointProperty	0.01	0.01
HomeOwnershipType_used_Other	0.00	0.00
HomeOwnershipType_used_Property	0.01	0.01
HomeOwnershipType_used_Tenant	0.01	0.01
EmploymentDurationCurrentEmployer_used_MoreThan5Years	0.01	0.01
EmploymentDurationCurrentEmployer_used_Other	0.00	0.00
EmploymentDurationCurrentEmployer_used_Retiree	0.00	0.01
EmploymentDurationCurrentEmployer_used_TrialPeriod	0.00	0.00
EmploymentDurationCurrentEmployer_used_UpTo1Year	0.01	0.01
EmploymentDurationCurrentEmployer_used_UpTo2Years	0.00	0.00
EmploymentDurationCurrentEmployer_used_UpTo3Years	0.00	0.00
EmploymentDurationCurrentEmployer_used_UpTo4Years	0.00	0.00
EmploymentDurationCurrentEmployer_used_UpTo5Years	0.01	0.01
AUC	**0.78**	**0.79**

This figure shows the number of each category from employment duration current employer. The range of the period is from less than one year to more than 5 years and also contains minority groups.

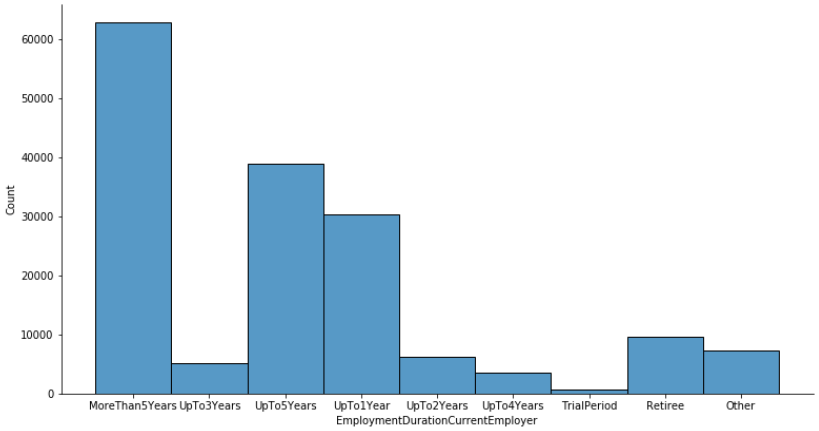

Figure 1. The number of each category from employment duration current employer.

This figure shows the number of each category from home ownership type. The criterion of this figure refers to the nature of the properties.

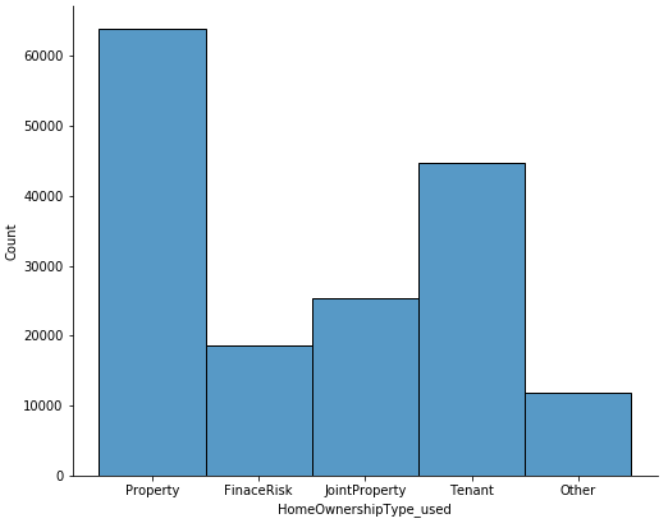

Figure 2. The number of each category from home ownership type.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0051

This picture shows the AUC for various model specifications. Xb_1 represents probability of default. Xb_2 represents ratio income. Xb_3 represents combination. Xb_4 represents combination and further control.

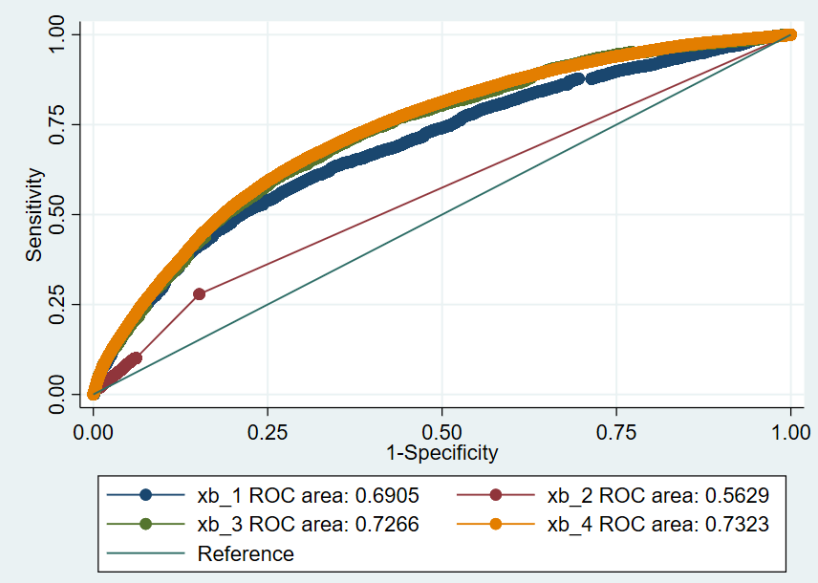

Figure 3. The AUC for various model specifications.

Additional equations

$$(7)\ \boldsymbol{y}(\boldsymbol{default}) = \beta_0 + \beta_1(Ratio\ of\ \ income\ principal) + \beta_2(Ratio\ of\ \ income\ alternative) + \varepsilon_i$$

$$(8)\ \boldsymbol{y}(\boldsymbol{default}) = \beta_0 + \beta_1(Probability\ of\ \ default) + \beta_2(Ratio\ of\ \ income\ principal) + \beta_3(Ratio\ of\ \ income\ alternative) + \varepsilon_i$$

$$\begin{aligned}(9)\ \boldsymbol{y}(\boldsymbol{default}) = {} & \beta_0 + \beta_1(Probability\ of\ \ default) + \beta_2(Ratio\ of\ \ income\ principal) + \\ & \beta_3(Ratio\ of\ \ income\ alternative) + \beta_4(IncomeTotal\_trans) + \\ & \beta_5(Age) + \beta_6(Gender) + \\ & \beta_7(EmploymentDurationCurrentEmployer\_used\_Other) + \\ & \beta_8(EmploymentDurationCurrentEmployer\_used\_Retiree) + \\ & \beta_9(EmploymentDurationCurrentEmployer\_used\_TrialPeriod) + \\ & \beta_{10}(EmploymentDurationCurrentEmployer\_used\_UpTo1Year) + \\ & \beta_{11}(EmploymentDurationCurrentEmployer\_used\_UpTo2Years) + \\ & \beta_{12}(EmploymentDurationCurrentEmployer\_used\_UpTo3Years) + \\ & \beta_{13}(EmploymentDurationCurrentEmployer\_used\_UpTo4Years) + \\ & \beta_{14}(EmploymentDurationCurrentEmployer\_used\_UpTo5Years) + \\ & \beta_{15}(HomeOwnershipType\_used\_FinaceRisk) + \\ & \beta_{16}(HomeOwnershipType\_used\_JointProperty) + \\ & \beta_{17}(HomeOwnershipType\_used\_Other) + \\ & \beta_{18}(HomeOwnershipType\_used\_Property) + \\ & \beta_{19}(HomeOwnershipType\_used\_Tenant) + \varepsilon i\end{aligned}$$

References

[1] A. Rampini, A. (2005). Default and aggregate income. Journal of Economic Theory, 122(2), 225-253. doi:https://doi.org/10.1016/j.jet.2004.04.004

[2] Anshari, M., Almunawar, M. N., Masri, M., & Hrdy, M. (2021). Financial Technology with AI-Enabled and Ethical Challenges. Society, 1-7.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0051

[3] Ashta, A., & Assadi, D. (2009). An analysis of European online micro-lending websites. Cahiers du CEREN, 29, 147-160.

[4] Aveni, T., Qu, C., Hsu, K., Zhang, A., Lei, X., & Hemrika, L. (2015). New Insights Into An Evolving P2P Lending Industry: how shifts in roles and risk are shaping the industry.

[5] Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25(2), 197-227.

[6] Bondora. (2021a). Retrieved from https://www.bondora.com/blog/wp-content/uploads/Bondora_ Presentation. pdf

[7] Bondora. (2021b). How reliable is the credit model? Retrieved from https://support.bondora.com/en/how-reliable-is-the-credit-model

[8] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[9] Can P2P Lending Reinvent Banking?. (2015). Retrieved from https://www.morganstanley.com/ideas/p2p-marketplace-lending

[10] Crouhy, M., Galai, D., & Mark, R. (2000). A comparative analysis of current credit risk models. Journal of Banking & Finance, 24(1), 59-117. doi:https://doi.org/10.1016/S0378-4266(99)00053-9

[11] DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics, 837-845.

[12] Eurostat. (2021). Retrieved from https://ec.europa.eu/eurostat/databrowser/view/ sdg_08_10/default/ table?lang=en

[13] Imbierowicz, B., & Rauch, C. (2014). The relationship between liquidity risk and credit risk in banks. Journal of Banking & Finance, 40, 242-256.

[14] Iyer, R., Khwaja, A. I., Luttmer, E. F. P., & Shue, K. (2016). Screening Peers Softly: Inferring the Quality of Small Borrowers. Management Science, 62(6), 1554-1577. doi:10.1287/mnsc.2015.2181

[15] Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance, 34(11), 2767-2787. doi:10.1016/j.jbankfin.2010.06.001

[16] McCollum, M., & Pace, R. K. (2017). Income Stability and Mortgage Default. Available at SSRN 3001218.

[17] Möllenkamp, N. (2017). Determinants of loan performance in P2P lending. University of Twente,

[18] Narkhede, S. (2018). Understanding auc-roc curve. Towards Data Science, 26, 220-227.

[19] Ong, C.-S., Huang, J.-J., & Tzeng, G.-H. (2005). Building credit scoring models using genetic programming. Expert systems with applications, 29(1), 41-47.

[20] Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. The journal of educational research, 96(1), 3-14.

[21] Puri, M., Gombović, A., Burg, V., Berg, T., & Karolyi, A. (2020). On the Rise of FinTechs: Credit Scoring Using Digital Footprints. The Review of Financial Studies, 33(7), 2845-2897. doi:10.1093/rfs/hhz099

[22] Shi, T., & Horvath, S. (2006). Unsupervised learning with random forest predictors. Journal of Computational and Graphical Statistics, 15(1), 118-138.

[23] Stein, R. M. (2002). Benchmarking default prediction models: Pitfalls and remedies in model validation. Moody's KMV, New York, 20305.

[24] Tang, H. (2019). Peer-to-peer lenders versus banks: substitutes or complements? The Review of Financial Studies, 32(5), 1900-1938.

[25] Treacy, W. F., & Carey, M. (2000). Credit risk rating systems at large US banks. Journal of Banking & Finance, 24(1-2), 167-201.

[26] Peng Hongfeng, & Ye Yonggang. (2011). Research on loan pricing model based on repayment ability and willingness to repay. Chinese Management Science, 19(06), 40-47.