Working Paper presented at the

Peer-to-Peer Financial Systems 2017 Workshop

2017

Scoring Models for P2P Lending Platforms: An Evaluation of Predictive Performance

Paolo Giudici
Università di Pavia

Branka Hadji Misheva
Università di Pavia

Powered by

P2P Financial Systems

Exponential Science

Scoring Models for P2P Lending Platforms: An Evaluation of Predictive Performance

by

Paolo Giudici[1], and Branka Hadji Misheva[2]

Abstract

Due to technological advancement, Peer-to-Peer (P2P) platforms have allowed significant cost reduction in lending. However, this improved allocation comes at the price of a higher credit risk. In this paper, the authors investigate the effectiveness of credit scoring models employed by P2P platforms with respect to loan default prediction. We claim that, because of differences in risk ownership with respect to traditional lenders, the rating grades obtained from P2P scoring models may not be the best predictors of loan default.

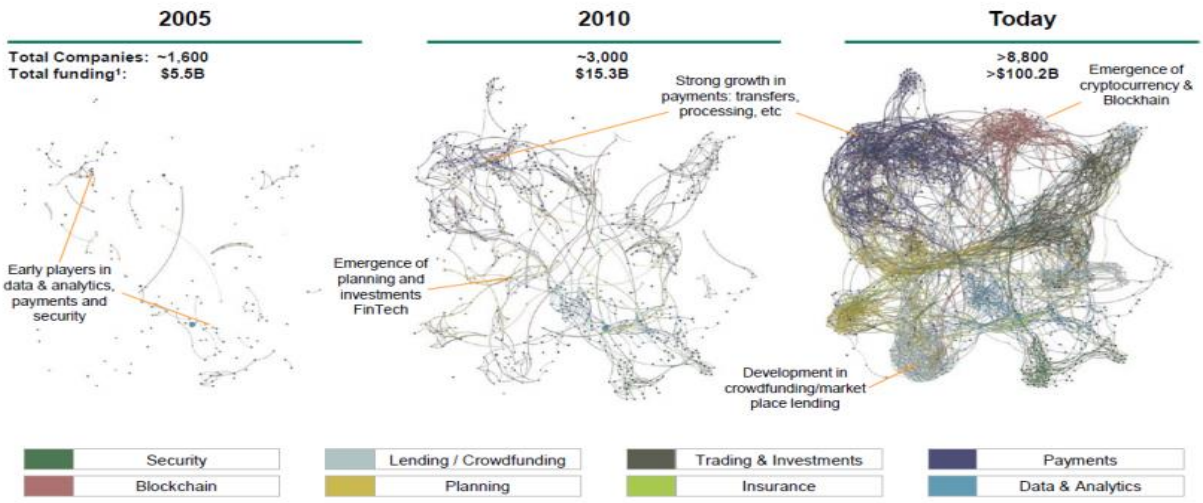
Key words: Credit ratings, Default prediction, Logistic regression models, Consumer Credit.
JEL classification codes: C55, G23

[1] University of Pavia, Department of Economics and Management, Pavia, Italy, E-mail: paolo.giudici@unipv.it
[2] University of Pavia, Department of Economics and Management, Pavia, Italy, E-mail: branka.hadjimisheva01@universitadipavia.it

I. Introduction

Within the past ten years, the emergence of financial technology ventures ('Fintechs') in both the consumer and commercial credit space have introduced many opportunities for both lenders and investors and have redefined the roles of traditional intermediaries. In this paper, we study peer-to-peer (P2P) lending platforms, which allow private individuals to make small, unsecured loans to other private borrowers or small companies. Since 2005, the growth of Fintech investments has been exponential with total funding jumping from around $5.5B in 2005 to more than $100.2B in 2017 (**Figure 1**).

Figure 1. The Growth of Fintech



Source: Bullmann (2017)

The advances in information technology has enabled online markets to provide an alternative to traditional financial intermediaries which is intended to redefine the banking system by making finance more cost efficient, consumer friendly and transparent thus improving overall value and quality of service. With the increasing role of these online lending marketplaces, a key point of interest becomes assessing the risk associated with P2P lending. As a general matter, P2P platforms are less able to deal with asymmetric

information compared to traditional banks and this in turn can lead to adverse selection in which investors cannot distinguish between borrowers belonging to different ranks of credit risk. This problem is made worse by the difference in risk ownership that exists between P2P and traditional banking models. Although both banks and P2P platforms rely on scoring models for the purpose of estimating the probability of default of a loan, the incentive for model accuracy between the two entities may differ significantly as in the context of P2P lending platforms, the credit risk is not born by the platform but by the investors. We claim that because of P2P's inability to solve for asymmetric information as efficiently as traditional banks and the differences in risk ownership, the grading system may not sufficiently reflect the probability of loan default. We test this by assessing the predictive performance of traditional scoring models employed by consumer-focused P2P lending platforms.

The sections are organized as follows. Section II outlines the motivation of analysis and presents the hypothesis development. Section III explains the data and methodology employed for the purpose of testing the predictive performance of P2P scoring models whereas Section IV presents the empirical results. Section V concludes the discussion and outlines opportunities for future contributions in the context of P2P scoring models.

II. Motivation

Many factors explain the increasing role of P2P lending platforms in the global world of finance. As these online marketplaces do not collect deposits, they can avoid many intermediation costs typically associated with traditional financial services. Namely, P2P platforms are not required to respect bank capital requirements nor pay fees associated with state deposit insurance practices and this in turn allows them to operate with lower costs. As argued by Serrano-Cinca et al. (2015) loans approved through P2P platforms are not accounted to the books of the platform thus no particular liability for the credit is required. The benefits associated with disintermediation are ultimately transferred to both borrowers and lenders (Serrano-Cinca et al., 2015). Explicitly, borrowers benefit because they are able to receive credits at lower

interest rates and in some cases with no collateral whereas lenders are incentivized to participate in the market because they can receive higher rate of return on investment due to reduced transaction costs (Jeong et al., 2012 and Emekter et al., 2015).

Additional to this, advancements in information technology have been a key force driving the exponential growth of P2P platforms. Big data analytics has changed how data is collected, processed, and evaluated which in turn has led to significant reductions in search costs for credit information (Yan et al., 2015). In this context, as noted by Yan et al. (2015) many P2P platforms rely not only on "hard" but also "soft" information for the purpose of carrying out credit checks, a practice not typically employed by traditional banks. This further allows P2P platforms to facilitate the identification of credible borrowers and expend credit availability. Prior studies have investigated the impact of several such "soft" information such as the applicants' pictures, descriptions concerning loan's usage as well as social networking. In line with this, Ge et al. (2016) find that two forms of social media information serve as a signal concerning an applicant's creditworthiness and those are: (i) the self-disclosure of social media account and (ii) the social media network and overall engagement.

With the implementation of the new EU Revised Payment Service Directive (PSD2) in 2018, P2P platform further stand to benefit from their inclination to use advanced data analytics. With this directive, the "monopoly" which banks currently have on their clients' account information and payment transactions will disappear as this information will be disclosed through application payment interfaces. The implementation of this directive would further pave the way for P2P platforms to improve their matching efficiency. This been said, it is clear that P2P platforms have the potential to improve allocative efficiency.

The advantages associated with P2P lending platforms notwithstanding, they can also pose significant risks to a financial system. Lending will always be associated with the risk of deterioration in the credit rating of the counterparty and compared to traditional banks, P2P are less able to eliminate asymmetric information thus increasing the risk of bad debt accumulation. Economic theory argues that banks represent an institutional solution to the problem of asymmetric information in the credit market between the borrower and the lender (Akerlof, 1970, Myers and Majluf, 1984 and Roure et al., 2016).

Namely, banks are able to access detailed information on clients' past financial and business transaction which in turn allows them to better discriminate between consumers of different credit risk rank. Additional to this, P2P platforms remain at a disadvantage because banks can better sustain the cost of monitoring the clients once a loan has been assigned. In a recent paper, Roure et al. (2016) claim that banks' expertise in screening and monitoring the activities of borrowers gives them a competitive advantage over P2P lenders, as both ex ante and ex post asymmetric information are mitigated.

Additional to this, further point of concern with respect to P2P platforms concerns the difference in risk ownership. In order to explain this, we offer a brief comparison between the business models and risk management of traditional banking institution and P2P lending platforms, respectively.

Figure 2. Bank risk model

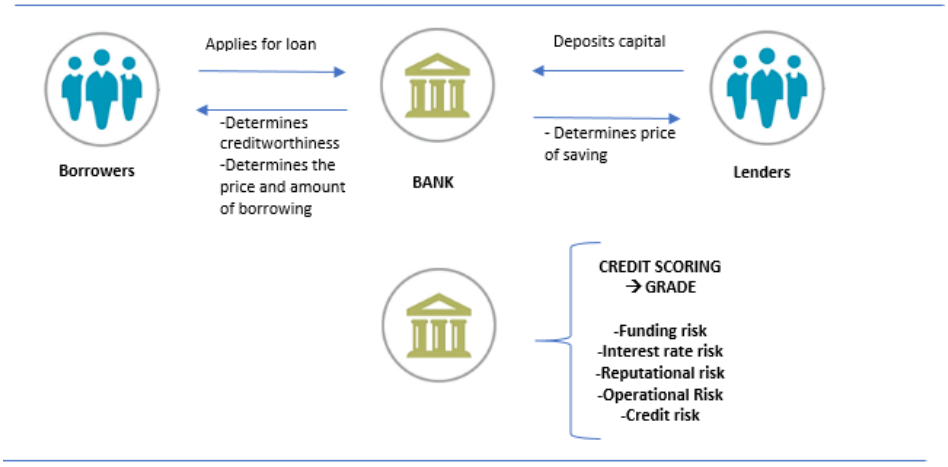
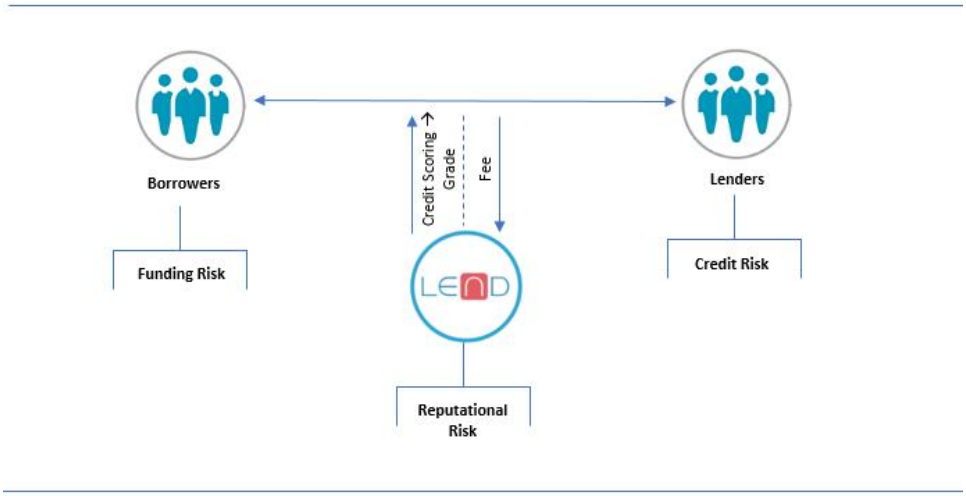


Figure 3. P2P risk model



In the context of traditional banking (**Figure 2**), there is a "many-to-one-to-many" approach, in which the financial intermediary (the bank) collects deposits, from several entities, fixing a borrowing price, and making decisions concerning to whom, among many, to lend the deposits, at what price and in which amount. Such decisions have a degree of transparency, as rating and prices are typically, at least partially disclosed. However, the intermediary's decision is not automatically determined by such information. In other words, the intermediary controls every aspect of the lending process (within a regulated environment).

On the other hand, P2P lending (**Figure 3**) is built on the basis of a "many-to-many" approach in which the financial intermediary empowers each lender to decide to whom borrower to lend and for what amount. To guide the process, the P2P platform provides lenders with information on the potential borrowers, their loan purpose and, more importantly, on their rating. P2P platforms assign a grade for each loan, which is intended to represent a unifying indicator of the overall creditworthiness of each individual loan applicant, on which decisions of the lender could be based. In other words, the intermediary does not really intermediate by making a lending decision but rather it provides the information on which such decision may be based.

A key issue, related with this, is the credit scoring. For both banks and P2P lenders, a rating system has the purpose of estimating the probability of default of a loan, which is then used in the decision process

concerning approval, interest rates and volume specifications. Although both traditional financial institutions and P2P rely on scoring models, the incentives for model accuracy may differ because of the differences in risk ownership. For banks, the grading is conducted by the financial institution itself which is the actual entity that assumes the credit risk. A bank is thus interested to have the most accurate possible model. On the other hand, in a P2P platform, grading is determined by the platform but the risk is fully borne by the lender (Serrano-Cinca et al., 2016). In other words, P2P lenders allow for direct matching between borrowers and lenders without the loans being held on the intermediary's balance sheet (Milne and Parboteeah, 2016). From a risk-return perspective, while in classical banking the financial institution chooses its optimal trade-off between risks and returns (subject to regulation constraints), in P2P lending, the platform maximizes its returns without taking care of the risks which are borne by the lenders. As a combined result of both asymmetric information and difference in risk ownership, the credit rating calculated by a platform may be upward biased. In line with this, the hypothesis to be tested is:

H0: The rating assigned by a P2P lending platform is not a good predictor of default

Loans that are more likely to default should receive a lower rating. However, we expect an upward bias in rating, due to the risk-return "dissociation" described before. It is important to mention that although the literature on P2P is not extensive, some authors have carried out investigations into the credit scoring models of P2P platforms (Serrano-Cinca et al., 2015, Guo et al., 2016 and Serrano-Cinca and Hutiernez-Nieto, 2016). In testing the hypothesis H0, we advance the field of research, contributing with a specific empirical focus on the measurement of possible credit scoring biases of P2P lenders.

III. Methodology and Research Design

3.1. Data Collection

To test the specified hypotheses, data is collected from Lending Club, which is the largest online marketplace connecting borrowers and investors. The analysis relies on loans' data covering the period 2007-2011 obtained from the platform's official webpage (available at: https://www.lendingclub.com/info/download-data.action). The key variables of interest are the ratings ("grades") assigned to each loan applicant and the status of the loan, which allows to identify the portion of those which have defaulted over the period of analysis.

3.2. Traditional Credit Scoring Models

Statistical theory offers a great variety of models for building and estimating the probability of default of lenders. All different approaches can be grouped in two broad categories: (i) parametric and (ii) non-parametric (Genriha and Voronova, 2012). For the purpose of reproducing the P2P grade-decision process and evaluating its performance in predicting loans' default, we employ the logistic regression which is the most known model available. In the context of P2P lending, logistic regression has been used in the studies of Andreeva et al. (2007), Barrios et al. (2013), Emekter et al. (2015) and Serrano-Cinca and Gutierrez-Nieto (2016).

Logistic regression aims to classify the dependent variable in two groups. In our case, two different regressions are carried out. In the first, the dependent variable distinguishes between creditworthy and not creditworthy applicants [1=creditworthy; 0=not creditworthy], a classification derived from the grades assigned by the P2P platform. The purpose of this is to reproduce the grading process of Lending club and identify the variables that the platform considers crucial determinants of the probability of loan default. The second regression, on the other hand, has the purpose of evaluating the performance of the grades assigned

by Lending club in predicting loan default against more advanced statistical models. In this context, the dependent variable distinguishes between defaulted and non-defaulted loans [1=default; 0=not default]. In both cases, logistic regression methods lead to the calculation of the predictive probability of default. Mathematically:

$$ln\left(\frac{p}{(1-p)}\right) = \alpha + \beta x + \varepsilon \quad (1)$$

where p is, respectively: (i) the probability of a loan applicant being ranked creditworthy, in the first regression and (ii) the probability of a loan defaulting, in the second regression. The logistic distribution constrains the estimated probabilities to belong to the range [0,1]. Mathematically, the probability of default will be obtained as:

$$PD = \frac{1}{1+e^{-\alpha+\beta x}} \quad (2)$$

IV. Empirical Findings: Traditional Scoring Models of P2P Lending Platforms

4.1.1. Descriptive Statistics

With respect to the specified hypothesis, we claim that because of P2P's inability to solve for asymmetric information as efficiently as traditional banks and the difference in risk-ownership between P2P and banks' models, the grading system may not sufficiently reflect the probability of loan default. For the purpose of investigating whether there is a basis for such an argument, we first present descriptive statistics.

Table 1 provides an exploratory analysis of the continuous variables which Lending club collects from loan applicants. We investigate the average value of the indicators across different grades and what becomes clear is that for some of the variables there is not a significant variability between the highest and lowest grade (ex. total number of accounts, revolving balance). Looking at the individual indicators, the highest variability is noticed with the revolving utility and loan amount over income variables.

In order to see whether the platform is taking into consideration the right information when assigning the grades, we also consider the variability of the indicators with respect to loan status (**Table 2**). Overall, the averages are not significantly different which in turn can be an indicator that the platform should expand the scope of information necessary to accurately predict default.

(Table 1 about here)

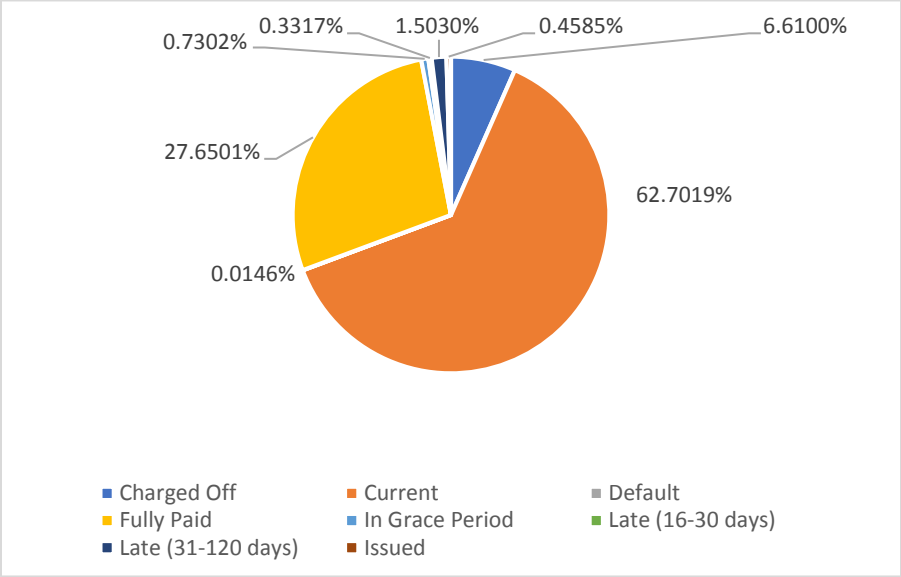
(Table 2 about here)

Table 3 in turn, provides a cross tabulation with respect to the categorical variables. Considering the grade assigned, it is clear that there exists a relationship between the grade and the loan status. The table shows that 93.9% of the loans graded A did not default and the percentage decreases as the grades become lower. This can be considered evidence of the fact that the P2P lending platform does improve allocative efficiency as it supplies credits to consumers who are considered not creditworthy by traditional financial institutions. Similar arguments are also offered by Serrano-Cinca et al. (2015). However, cause for concern does exists when one considers that the majority of defaulted loans were ranked "C". This can be considered an indicator of the upward bias discussed previously. Furthermore, the data presented in Table 3 indicates that a very small proportions of the applicants' self-reported information has been verified which is unexpected. Finally, there is some variability in the percentages of defaulted loans among different loan purposes indicating there might exists a need for a cluster- or network-dependent grading model.

(Table 3 about here)

Before testing H0, the following figure represents the proportion of defaulted loans in the context of Lending club, for the period 2007 - 2016.

Figure 2. Loan Status, Lending Club



From **Figure 2**, defaulted loans together with those classified as charged-off comprise 6.6246% of total loans intermediated by Lending club. Although the proportion is not high, we believe that the frequency of the event is sufficient for developing and testing statistically predictive models. We remark that, in the case of lower default frequencies, and for robustness purposes, logistic regression could be extended with the Generalized Extreme regression scoring model proposed by Calabrese and Giudici (2015).

4.1.2. Predictive Performance of Lending Club's Scoring Model

Before testing the predictive performance of Lending Club's grading system, we want to identify which information among those publicly available are most relevant in determining an applicant's creditworthiness. In order to achieve this, an attempt is made to reproduce Lending Club's grading process. **Table 4** reports the findings from logistic regression.

(Table 4 about here)

In **Table 4**, results from 3 models are presented. Out of the 13 main variables included, twelve (among which ownership, total number of accounts, the FICO score, loan purpose, number of inquiries in the past 6 months, debt-to-income ratio, location, the number of months since borrower's last delinquency, revolving balance, and revolving line utilization rate) were found to have a statistically significant impact on the assigned grade. The results further suggest that annual income has no impact on the assigned grade which is somewhat surprising. Empirical research on the determinants of credit ranking in the context of traditional financial institutions have repeatedly found borrowers' income to be a significant determinant of the assigned rank or grade (Adams et al., 2003 and Jin and Zhu, 2015). The fact that there is not enough evidence to reject the hypothesis that annual income does not influence credit ranking in the context of Lending Club could be an indicator of the biased scoring model employed by this intermediary. Further evidence in support of this argument can be found in the estimated coefficient concerning the verification status. Common economic logic would dictate that the verification of the information provided by borrowers is of crucial importance to the credit ranking. Our empirical findings show that although the variable is found statistically significant, its sign is ambiguous.

In the next step, we proceed towards evaluating the predictive performance of the grades assigned by Lending Club with respect to loan default, in a second regression model. The results are presented in **Table 5.**

(Table 5 about here)

From **Table 5**, the grade variable is a statistically significant predictor of loan default, but its overall predictive power is limited. Namely, if we consider the estimated area under the ROC curve (AUC) as a measure of predictive performance, the results suggest that the assigned grades do not have high predictive utility - the AUC value for the model using only the grade as a predictor of loan default is equal to 0.618. Furthermore, in order to investigate whether there is evidence if favor of the argument that grades are bias upwards we conduct additional diagnostic tests. The error matrix as well as additional statistics are presented in **Table 6**.

(Table 6 about here)

The confusion matrix provides evidence of the argued upward bias inherit in the P2P grading process due to both its inability to solve for asymmetric information and the different risk-ownership compared to traditional financial institutions.

The predictive performance of the default model does not change significantly even if we consider more advanced statistical models. In this respect, **Table 5** also presents the results from two additional models aimed at capturing the determinants of loan default in the context of P2P platforms. An important finding from the conducted estimations is that the predictive performance of the scoring model improves by several percentage points once terms capturing the interaction between purpose and other control variables are included. **Table 5** shows that several interaction terms were found statistically significant thus suggesting that same control variables can differently affect the probability of default dependent on the purpose for which the loan is taken.

Although the predictive power of the scoring model increases as time and space predictors as well as interaction terms are included in the estimation, the improvements can be considered small as AUC values vary within the range 0.618-0.678. This is not to say that such improvements are irrelevant as even a small improvement in accuracy can lead to significant future savings (West, 2000). Still predictive performance below 70% represents a concern and there is a clear need to increase the scoring accuracy of the credit decisions. What these preliminary insights suggest is that loan default in the context of P2P platforms is impacted by factors other than those observed and requested by Lending Club. In order to pursue improvements in the credit scoring models, it is thus necessary to explore other approaches beyond the traditional scoring models.

V. Summary Discussion

This study aims at investigating the predictive performance of P2P credit scoring by analyzing data collected from Lending Club, the largest online marketplace connecting borrowers and lenders. Our empirical findings suggest that although there is a statistically significant relationship between the assigned rating grades and loan default, grades do not have high predictive power. Furthermore, the predictive performance does not change significantly if we apply more elaborate statistical models using the information that Lending Club collects on the borrowers. Such findings indicate a need for exploring other approaches beyond the traditional scoring models.

A possible step forward is employing a network-based scoring model which will take into account the financial relationships between borrowers and lenders. Namely, a key characteristic of P2P platforms is that they are, by construction, globally interconnected. Classical banks have, over the years, segmented their reference markets into specific territorial areas thus increasing their expertise and the accuracy of their ratings. Differently, P2P platforms are based on a "universal" banking model that is fully inclusive, without space and business type limitations, which in turn makes the determination of a correct rating a particularly difficult task. However, these platforms have the advantage of an improved data collection on the network to which a borrower belongs. Using network information can improve the scoring model and in turn default prediction.

VI. References

Adams, M., Burton, B. and Hardwick, P. (2003). The Determinants of Credit Rating in the United Kingdom Insurance Industry, *Journal of Business Finance and Accounting*, 30(3), pp.539-572

Andreeva, G., Ansell, J. and Crook, J. (2007). Modelling Profitability Using Survival Combination Scores, *European Journal of Operational Research*, 183(3), pp. 1537–1549.

Akerlof, G. A. (1970). The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics*, 488–500.

Avdjiev, S., Giudici, P. and Spelta, A. (2017). Measuring Contagion Risk in International Banking. Technical Report, submitted.

Barrios, L.J.S., Andreeva, G. and Ansell, J. (2013). Monetary and relative scorecards to assess profits in consumer revolving credit, *Journal of the Operational Research Society*, 65 (3), pp. 443–453.

Calabrese, R., Giudici, P.S. (2015). Estimating Bank Default with Generalized Extreme Value Regression Models. *Journal of the Operational Research Society* 66 (11), pp. 1783-1792.

Emekter, R., Tu, Y., Jirasakuldeeh, B. and Lu, M. (2015). Evaluating Credit Risk and Loan Performance in Online Peer-to-Peer (P2P) Lending. *Applied Economics*, 47(1), pp.54-70;

Ge, R., Fend, J. and Gu, B. (2016). Borrower's Default and Self-Disclosure of Social Media Information in P2P Lending. *Financial Innovation*, 2(30), pp. 1-6;

Genriha, I. and Voronova, I. (2012). Methods for evaluating the creditworthiness of borrowers. *Economics and Business* 22 (2012), pp. 42-49.

Giudici, P.S., Spelta, A. (2016). Graphical network models for international financial flows. *Journal of Business and Economic Statistics* 34 (1), pp. 126-138;

Guo, Y., Zhou, W., Luo, C., Liu, C. and Xiong, H. (2016). Instance-Based Credit Risk Assessment for Investment Decisions in P2P Lending. *European Journal of Operational Research*, 246, pp.417-426

Hung, J. and Luo, B. (2016). FinTech in Taiwan: A Case of a Bank's Strategic Planning for an Investment in a Fintech Company. *Financial Innovation*, 2(15), pp:1-16;

Jin, Y. and Zhu, Y. (2015). A Data-Driven Approach to Predict Default Risk of Loan for Online Peer-to-Peer (P2P) Lending. *Fifth International Conference on Communication Systems and Network Technologies*.

Jeong, G., Lee, E. and Lee, B. (2012). Does Borrowers' Information Renewal Change Lenders' Decision in P2P Lending? An Empirical Investigation. Proceeding ICEC '12 Proceedings of the 14th Annual International Conference on Electronic Commerce.

Kolaczyk, E.D. and Csardi, G. (2013). Statistical analysis of network data with R. *Springer-Varlag*, Berlin, 2013;

Li, J., Hsu, S., Chen, Z. and Chen, Y. (2016). Risks of P2P Lending Platforms in China: Modeling Failure Using a Cox Hazard Model. *The Chinese Economy*, 49, pp:161-172

Milne, A. and Parboteeah, P. (2016). The Business Models and Economics of Peer-to-Peer Lending. Technical report, European Credit Research Institute.

Myers SC, Majluf NS. (1984). Corporate Financing and Investment Decisions When Firms Have Information that Investors Do Not Have. *Journal of Financial Economics*, 13(2), pp. 187–221.

Rajan, U., Seru, A. and Vig, V. (2014). The Failure of Models that Predict Failure: Distance, Incentives and Default. *Journal of Financial Economics*, 115, pp.237-260;

Roure, C., Pelizzon, L. and Tasca, P. (2016). How Does P2P Lending Fit into the Consumer Credit Market. Deutsche Bundesbank. Discussion Paper, no. 30.2016.

Serrano-Cinca, C. and Hutiernez-Nieto, B. (2016). The Use of Profit Scoring as an Alternative to Credit Scoring Systems in Peer-to-Peer Lending. *Decision Support Systems*, 89, pp.113-122;

Serrano-Cinca, C., Hutiernez-Nieto, B. and Lopez-Palaclos, L. (2015). Determinants of Default in P2P Platforms. *Plos One*

Stiglitz, J. and Weiss, A. (1981) Credit Rationing in Markets with Imperfect Information, *American Economic Review*, 71 pp.393–419.

Yan, J., Yu, W. and Zhao, L. (2015). How Signaling and Search Costs Affect Information Asymmetry in P2P Lending: The Economics of Big Data. *Financial Innovation*, 1(19), pp. 1-11;

West, D. (2000). Neural Network Credit Scoring Models. *Computer & Operations Research*, 27(11-12), pp. 1131-1152

	Grade							
	A	B	C	D	E	F	G	Non-parametric Tests
Loan Amount	13196.8	12868.7	13770.4	14784.2	17837.2	19244.8	20924.9	0.000***
Annual Income	84481.48	73537.10	71158.83	69801.91	73464.61	74569.49	79923.55	0.000***
Debt Ratio	14.71	16.76	18.23	18.93	19.74	19.76	19.53	0.000***
Fico Score	736	703	692	687	686	683	681	0.000***
Credit revolving balance	16827	15639	15452	15156	16398	16175	16787	0.000***
Revolving line utilization rate	38.51%	52.61%	57.33%	60.15%	60.91%	61.82%	61.76%	0.000***
Total Number of Accounts	27	25	25	25	26	26	26	0.000***
Loan Amount over Income	.1786	.1982	.2149	.2333	.2675	.2814	.2934	0.000***

Table 1. Descriptive analysis of categorical variables across grades. The Ho hypothesis of the non-parametric test is that the distribution of the independent variable is the same across categories of grade.
***** - significant at 1% l.s.**

	Loan Status	
	Not Defaulted	Defaulted
	Mean	Mean
Loan Amount	13979.3	15053.5
Annual Income	75478.61	67319.54
Debt Ratio	17.16	19.48
Fico Score	702	692
Credit revolving balance	15877	15337
Revolving line utilization rate	53.16%	58.14%
Total Number of Accounts	25	25
Loan Amount over Income	.2080	.2465

Table 2. Descriptive analysis of continuous variables

		Loan Status					
		Not Defaulted		Defaulted		Total	
		Count	Row N %	Count	Row N %	Count	Row N %
Grade	A	70723	93.9%	4574	6.1%	75297	100.0%
	B	124339	88.3%	16528	11.7%	140867	100.0%
	C	107626	81.1%	25158	18.9%	132784	100.0%
	D	61134	74.8%	20618	25.2%	81752	100.0%
	E	28411	68.4%	13132	31.6%	41543	100.0%
	F	9927	63.4%	5731	36.6%	15658	100.0%
	G	2452	59.9%	1644	40.1%	4096	100.0%
Term	36 months	316832	85.5%	53859	14.5%	370691	100.0%
	60 months	87780	72.4%	33526	27.6%	121306	100.0%
Home Ownership	ANY	8	100.0%	0	0.0%	8	100.0%
	MORTGAGE	205686	84.4%	37931	15.6%	243617	100.0%
	NONE	36	83.7%	7	16.3%	43	100.0%
	OTHER	114	80.9%	27	19.1%	141	100.0%
	OWN	38663	81.8%	8630	18.2%	47293	100.0%
	RENT	160105	79.7%	40790	20.3%	200895	100.0%
Verification Status	Not Verified	132397	86.3%	20945	13.7%	153342	100.0%
	Source Verified	132806	81.0%	31091	19.0%	163897	100.0%
	Verified	139409	79.8%	35349	20.2%	174758	100.0%
Purpose of Loan	car	5126	87.7%	717	12.3%	5843	100.0%
	credit_card	85837	84.7%	15565	15.3%	101402	100.0%
	debt_consolidation	237721	81.3%	54604	18.7%	292325	100.0%
	educational	270	82.8%	56	17.2%	326	100.0%
	home_improvement	25295	84.8%	4529	15.2%	29824	100.0%
	house	2296	82.3%	494	17.7%	2790	100.0%
	major_purchase	9449	85.4%	1618	14.6%	11067	100.0%
	medical	4191	80.1%	1043	19.9%	5234	100.0%
	moving	2850	79.1%	754	20.9%	3604	100.0%
	other	21478	80.8%	5118	19.2%	26596	100.0%
	renewable_energy	338	79.5%	87	20.5%	425	100.0%
	small_business	5329	72.8%	1988	27.2%	7317	100.0%
	vacation	2475	82.2%	536	17.8%	3011	100.0%
	wedding	1957	87.6%	276	12.4%	2233	100.0%

Table 3. Descriptive analysis of categorical variables across loan status

	(1)			(2)			(3)		
	Estimate	**Std. Error**	**Sig.**	**Estimate**	**Std. Error**	**Sig.**	**Estimate**	**Std. Error**	**Sig.**
(Intercept)	-2.02E+01	1.22E-01		-2.09E+01	1.25E-01	***	-2.34E+01	1.34E-01	***
Annual income	-6.39E-08	7.02E-08							
Loan amount									
Loan over income				-3.72E+00	3.79E-02	***	-4.28E+00	3.78E-02	***
Ownership	9.33E-02	8.07E-03	***	9.12E-02	8.18E-03	***	2.75E-02	8.22E-03	***
Total number of accounts	1.57E-02	3.61E-04	***	1.30E-02	3.61E-04	***	1.17E-02	3.61E-04	***
Fico score	3.04E-02	1.69E-04	***	3.21E-02	1.74E-04	***	3.33E-02	1.76E-04	***
Inquiries	-4.52E-01	4.18E-03	***	-4.89E-01	4.26E-03	***	-5.04E-01	4.28E-03	***
Address_group1	6.95E-02	8.29E-03	***	7.82E-02	8.42E-03	***	8.28E-02	8.45E-03	***
Address_group2	-3.85E-01	1.11E-01	***	-3.94E-01	1.12E-01	***	-4.61E-01	1.13E-01	***
Address_group3	-5.83E-01	1.56E-01	***	-5.33E-01	1.57E-01	**	-5.68E-01	1.58E-01	***
Months since the borrower's last delinquency	2.17E-03	1.71E-04	***	1.48E-03	1.73E-04	***	1.55E-03	1.74E-04	***
Revolving line utilization rate	-5.31E-01	1.89E-02	***	-5.22E-01	1.91E-02	***	-5.57E-01	1.91E-02	***
Credit revolving balance	2.97E-06	2.24E-07	***	2.47E-06	2.21E-07	***	7.79E-07	2.11E-07	***
Verification status	-7.74E-01	8.25E-03	***	-6.20E-01	8.45E-03	***			
Debt ratio	-3.81E-02	5.24E-04	***	-2.65E-02	5.21E-04	***	-2.62E-02	5.23E-04	***
Purpose_1							2.00E+00	3.82E-02	***
Purpose_2							1.39E+00	3.78E-02	***
Purpose_3							4.90E-02	6.18E-02	
AUC	**0.7848775**			**0.7983443**			**0.8014684**		

Table 4. Models (1), (2) and (3) are logistic regression models with grade as a dependent variable. Codes = * α=0.01; * α=0.1. Explanations of variables are presented in Appendix 1.**

	(1)			(2)			(3)		
	Estimate	**Std. Error**	**Sig.**	**Estimate**	**Std. Error**	**Sig.**	**Estimate**	**Std. Error**	**Sig.**
(Intercept)	-1.14726	0.01	***	4.72E+00	1.41E-01	***	3.54E+00	7.75E-01	***
Grade	-1.07895	0.01	***						
Annual income							1.14E-06	5.56E-07	*
Loan amount							1.04E-05	3.91E-06	**
Loan over income ratio				2.58E+00	4.12E-02	***	2.33E+00	7.03E-02	***
Ownership				-1.94E-01	9.23E-03	***	-1.78E-01	9.39E-03	***
Total number of accounts				-6.21E-03	4.20E-04	***	-8.10E-03	2.93E-03	**
Fico score				-1.05E-02	1.93E-04	***	-8.03E-03	1.05E-03	***
Inquiries				1.66E-01	4.17E-03	***	1.58E-01	4.21E-03	***
Address_group1				-1.70E-01	9.85E-03	***	-1.73E-01	9.91E-03	***
Address_group2				-5.09E-01	1.33E-01	***	-3.38E-01	1.36E-01	*
Address_group3				-1.65E+00	3.01E-01	***	-1.29E+00	3.01E-01	***
Months since the borrower's last delinquency				-1.89E-03	2.01E-04	***	-1.97E-03	2.02E-04	***
Revolving line utilization rate				2.15E-01	2.24E-02	***	1.41E-01	1.40E-01	
Revolving line utilization rate,				-2.58E-06	3.03E-07	***	3.75E-07	1.25E-06	
Verification Status				2.37E-01	1.05E-02	***	2.26E-01	1.08E-02	***
Debt ratio				2.30E-02	5.80E-04	***	2.48E-02	6.23E-04	***
Purpose1							1.49E+00	8.17E-01	.
Purpose2							1.44E+00	7.92E-01	.
Purpose3							6.63E-01	1.49E+00	
income_group1							-2.27E-06	6.03E-07	***
income_group2							-1.24E-06	5.67E-07	*
income_group3							-5.05E-06	1.51E-06	***
loan_group1							5.45E-07	4.03E-06	
loan_group2							-8.23E-06	3.92E-06	*
loan_group3							1.00E-05	8.76E-06	
fico_group1							-2.98E-03	1.11E-03	**
fico_group2							-2.74E-03	1.07E-03	*
fico_group3							-8.41E-04	2.04E-03	
balance_group1							-3.06E-06	1.38E-06	*
balance_group2							-3.10E-06	1.31E-06	*

balance_group3			1.51E-06	3.26E-06	
util_group1			-3.42E-02	1.46E-01	
util_group2			7.94E-02	1.42E-01	
util_group3			5.68E-02	2.44E-01	
total_group1			1.38E-04	3.03E-03	
total_group2			1.02E-03	2.96E-03	
total_group3			-3.84E-03	5.21E-03	
issue_year07			2.78E-01	2.32E-01	
issue_year08			1.57E-01	1.19E-01	
issue_year09			-7.06E-04	6.33E-02	
issue_year10			1.55E-02	3.92E-02	
issue_year11			1.24E-01	2.81E-02	***
issue_year12			-4.37E-02	1.85E-02	*
issue_year13			-1.02E-01	1.37E-02	***
issue_year14			2.61E-01	1.26E-02	***
issue_year15			1.08E-01	1.31E-02	***
issue_year16			-1.81E+00	4.15E-02	***
AUC	0.6184062	0.6595419	0.6787838		

Table 5. Models (4), (5) and (6) are logistic regression models with loan status as a dependent variable. Codes = * α=0.01; ** α=0.05; * α=0.1. Explanations of variables are presented in Appendix 1.**

		Reference	
		Not Defaulted	Defaulted
Prediction	Not defaulted	101247	21752
	Defaulted	0	0

Accuracy	0.8232
95% CI	(0.821, 0.8253)
Kappa	0
Mcnemar's Test P-value	0
Sensitivity	
Specificity	0
Pos Pred Value	0.8232
Neg Pred Value	NaN
Prevalence	0.8232
Detection Rate	0.8232
Detection Prevalence	1
Balanced Accuracy	0.5
Positive Class	0

Table 6. Error Matrix and Additional Statistics; 50% threshold

Appendix 1 – Explanations of Variables (Lending Club)

Variable	Explanations	Variable	Explanation
Grade	Assigned loan grade	balance_group1	Interaction between revolving balance and purpose group
Annual income	The self-reported annual income provided by the borrower	balance_group2	
Loan amount	The listed amount of the loan applied for by the borrower	balance_group3	
Loan over income ratio	Loan amount divided by annual income	util_group1	Interaction between revolving line utilization rate and purpose group
Ownership	Binary variable taking value [1] if the borrower owns or morgages his home	util_group2	
Total number of accounts	The total number of credit lines currently in the borrower's credit file	util_group3	
Fico score	The upper boundary range the borrower's FICO at loan origination belongs to.	total_group1	Interaction between total number of accounts and purpose group
Inquiries	The number of inquiries in past 6 months	total_group2	
Address_group1	Locations were classified into four groups based on the default odds ratios	total_group3	
Address_group2		issue_year07	Dummies referring to the year
Address_group3		issue_year08	
Months since the borrower's last delinquency	The number of months since the borrower's last delinquency.	issue_year09	
Revolving line utilization rate	Total credit revolving balance	issue_year10	
Revolving line utilization rate,	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.	issue_year11	
Verification Status	Indicates if income was verified by Lending club	issue_year12	
Debt ratio	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations	issue_year13	
Purpose1	Debt purposes were classified into four groups based on the default odds ratios	issue_year14	
Purpose2		issue_year15	
Purpose3		issue_year16	
income_group1	Interaction between annual income and purpose group	fico_group1	Interaction between fico score and purpose group
income_group2		fico_group2	
income_group3		fico_group3	
loan_group1	Interaction between loan over income ratio and purpose group		
loan_group2			
loan_group3			