Emerging Markets Finance & Trade, 55:118–132, 2019

ISSN: 1540-496X print/1558-0938 online
DOI: https://doi.org/10.1080/1540496X.2018.1479251

Routledge
Taylor & Francis Group

Predicting Prepayment and Default Risks of Unsecured Consumer Loans in Online Lending

Zhiyong Li[1,2], Ke Li[2,3], Xiao Yao[4], and Qing Wen[1]

[1]*School of Finance, Southwestern University of Finance and Economics, Chengdu, China;* [2]*Collaborative Innovation Center of Financial Security, Southwestern University of Finance and Economics, Chengdu, China;* [3]*School of Statistics, Southwestern University of Finance and Economics, Chengdu, China;* [4]*Business School, Central University of Finance and Economics, Beijing, China*

ABSTRACT: Online lending provides a means of fast financing for borrowers based on their creditworthiness. However, borrowers may undermine this agreement due to early repayment or default, which are two major concerns for the platform and lenders, since both affect the profitability of a loan. While default risk is frequently focused on credit scoring literature, prepayment has received much less attention, despite a higher prepayment rate being observed in online lending when compared with default. This article uses multivariate logistic regression to predict the probability of both the underlying prepayment and default risks. Real consumer lending data of 140,605 unsecured loans provides evidence that these two events have their own distinct patterns. We consider systemic risk by incorporating macroeconomic factors in modeling and address the influence of economic conditions, which are lessons learnt from the last financial crisis. The out-of-sample validation has shown that both prepayment and default can be accurately predicted. This article highlights the necessity of regulations on prepayment given the fast growing online lending market.

KEY WORDS: credit, default, online lending, prediction, prepayment

JEL CLASSIFICATION: H81, C53, D81

With the development of financial technology (commonly known as FinTech), the online lending market has quickly become a popular source of credit for internet users. Borrowers of different grades of creditworthiness can conveniently get access to credit via online lending platforms. Innovative peer-to-peer (P2P) lending has altered the mechanism of lending and borrowing. When borrowers submit loan requests online, loans are funded by a number of interested lenders who contribute a partial portion, as opposed to being fully funded by a single financial institution such as commercial banks. P2P companies act more like matchmakers to provide information and services for users.

Unlike conventional banking, where creditors usually have powerful risk assessment techniques to evaluate the creditworthiness of borrowers, lenders on these platforms largely rely on their experience and personal risk tolerance to make decisions on loan selection. Therefore, information asymmetry presents a problem in P2P lending, and in general default risk is reduced due to the increased access to information (Miller 2015). Lenders are able to make judgements on borrowers based on their friendship networks (Lin, Prabhala, and Viswanathan 2013), photographs borrowers upload (Duarte, Siegel, and Young 2012), or soft information such as loan purpose and text description, in evaluating a borrower's creditworthiness (Iyer et al. 2016). Zhang and Liu (2012) noticed that herding behavior among lenders is evident, wherein lenders tend to follow others in selecting loans for investment. The price of loans, namely the interest charged, is to a large extent based on the underlying credit quality of

Address correspondence to Ke Li, School of Statistics, Southwestern University of Finance and Economics, 555 Liutai Avenue, Wenjiang, Chengdu 611130, Sichuan, China. E-mail: likec@swufe.edu.cn

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/mree.

borrowers. Rigbi (2012) for instance found that default probability is not affected by the interest rate cap on the platform, even though borrowers of low grades are charged at high rates.

The studies mentioned above are just a few examples from the increasing amount of literature addressing the issues in this alternative lending which have emerged in recent years. However, as P2P lending is essentially another form of lending, the resultant default probability and its accurate prediction is therefore of importance to both the lending platform and lenders making use of it. There have been a fast increasing number of articles focused on credit risk prediction using Cox proportional hazard models (Emekter et al. 2015), discrete hazard model (Hwang and Chu 2014), random forest models (Malekipirbazari and Aksakalli 2015), and instance-based models (Guo et al. 2016).

In the meantime, as online lending provides easy access to credit, people are continuing to use it as an alternative to mainstream banking. Particularly, borrowers of low credit scores usually have to seek for finance at high-cost lending institutions such as payday loan companies (Bhutta, Skiba, and Tobacman 2015). Payday loans have features of small amounts and short terms, and receive strict regulations in the US because of its nature of high risks (Melzer and Morgan 2015). However, given the fact that P2P platforms do not charge penalties on early repayment or prepayment, borrowers have strong incentives to replace payday loans by online P2P loans as the prices including interest rates and fees are significantly lower than payday loans. We have not seen regulations of P2P loans have focused on this. Although prepayment and default are two of the main causes affecting the profitability of a loan (Stepanova and Thomas 2002), so far the remarkably high prepayment rate in online lending has not been addressed.

We add to the literature by using a large sample of 140,605 unsecured consumer loans from a P2P lending platform and build a multinomial logistic model to predict the probability of prepayment and default. We will first give a review on the literature of both default and prepayment models in consumer credit modeling, though the latter is taken mainly from the mortgage market. We then introduce the method and data employed in the analysis. Loans are segmented by grade in the preliminary analysis and the results of out-of-sample predictions are discussed in the analysis section. Following final conclusions, possible implications are highlighted to P2P companies, its users and regulators.

Literature Review

In consumer credit, it is argued that the objective of lenders represented by banks is to transfer from risk-based scoring to profit-based scoring, and that profitability is sensitive to events such as prepayment, default, attrition, and cross-selling of other products (Thomas 2009). While attribution and cross-selling is common in revolving credit, prepayment and default is frequently noticeable in fixed term loans, which have become two of the major concerns for creditors. The Basel Committee on Bank Supervision defines default as a borrower failing to meet its obligations in accordance with agreed terms (BCBS 2000). Prepayment therefore represents to some extent a breakage with agreed terms as early settlement is unfavorable for lenders. Whilst default brings losses of the principle plus interest to lenders, early repayment means lenders suffer only part-accrual of the interest for the remaining payments. Nevertheless, their occurrence would still have a great impact on the overall profitability of a loan.

Given the high importance of prepayment and default to lenders, academics have thus far not accorded enough importance to prepayment risk as compared to default risk. In particular, most empirical studies regarding prepayment risk are based on secured loans where collaterals are involved in the lending process. For example, Heitfield and Sabarwal (2004) documented a study of subprime automobile loans in the US, and their results revealed that while prepayment risk increases dramatically with loan age, it is not affected by interest rates; and during its maturity period, prepayment risk is twice as high as default risk. Not surprisingly, most literature is concentrated on mortgage loans, e.g. Ciochetti et al. (2002), Ciochetti et al. (2003) and Pennington-Cross (2010). Given the large volume of

mortgage loans and mortgage-backed securities, accurate pricing of mortgages is the key to controlling the risks in mortgage portfolios. Prepayment and default on mortgages reduces portfolio profitability, which may cause further serious problems of asset pricing for those investing in mortgage-backed securities – one of the reasons behind the recent subprime crisis which started in 2007. Deng, Quigley, and van Order (2000) tested the option theory in the mortgage market, where the property holder can excise the call option by early repayment and refinancing the mortgage, if the market value of the property exceeds its original value; and further, that a put option can be excised by defaulting on the mortgage, if the market value goes below the original value. They also commented that transaction costs would be a significant factor in the excising decision. In commercial mortgages, creditors usually implement penalty clauses to prevent refinancing in the case of prepayment (Steinbuks 2015; Varli and Yildirim 2015). We have noticed, however, that in worldwide online lending markets, prepayment generates no fee, unlike a borrower's transaction cost, so the prepayment rate is relatively very high, as our data shows.

For consumer loans other than mortgages, various default risk-based credit scoring techniques are frequently compared and discussed in academic surveys (Crook, Edelman, and Thomas 2007; Hand and Henley 1997; Thomas 2000), and in practice are preferred by banks, because the control of default risk and the strategy of classification is closely related to a bank's profits (Blöchlinger and Leippold 2006; Stein 2005). Although the prepayment risk of unsecured consumer loans also has a bearing on a bank's profits, it has been far less studied. One attempt is from Agarwal and Taffler (2008), who argue that a decline in borrower credit quality, indicated by a Fair Isaac COmpany (FICO) score, would lower the probability of prepayment, and that similar effects exist in the increase of interest rates and the unemployment rate. In another scenario, Carling, Jacobson, and Roszbach (2001) predicted the dormancy risk of consumer loans defined as moving from active into termination, or where the balance remains below a very low level of revolving credit. Like prepayment, low usage indicates a low expected return from an account, which is unfavorable for a bank.

The method used to differentiate causes of termination is rather vividly called 'competing risks', whereby all possible events which could lead to the exit of an account compete against one another to happen first. Whichever happens first may halt or indeed stop other events from occurring. Competing risks are common in studying mortality rates. In the credit scoring context, default, attrition, prepayment and closure will all lead to a borrower stopping using the existing account. Agarwal, Ambrose, and Liu (2006) used a competing risk model to investigate the causes of the credit utilization change. The examples discussed above show how competing risk models are useful in studying the effects on the borrower of the eventual outcome in consumer credit, but they are also crucial in studying the final termination of a company. For instance, Esteve-Pérez, Sanchis-Llopis, and Sanchis-Llopis (2010) studied Spanish manufacturing firm exits through either liquidation/bankruptcy or acquisition/merger, which are similarly affected by discrete factors.

Survival analysis is not only able to predict the probability of an event occurring given that it has not occurred before, but can also estimate the time span of that event, as suggested by Banasik, Crook, and Thomas (1999). In the context of credit scoring, the events of interest typically include default, early repayment, closure of an account and purchases of other services and products. All of these events influence the overall profitability of a loan portfolio. Andreeva, Ansell, and Crook (2007) have considered the time to both default and to second-purchase using survival analysis, in order to calculate the net present value of involving credit store cards.

The incorporation of time-varying covariates in survival analysis makes it possible to analyze the influence of those systemic factors which may initially have a common effect on all accounts, but which may change overtime. Incorporating systemic factors in credit risk modeling is also the requirement of the Basel Committee for Internal Ratings-Based approaches and stress testing. For example, in the analysis of credit card accounts opened from 1997 to 2005, Bellotti and Crook (2008) included interest rates, FTSE index, GDP index, house price index and consumer confidence index in the hazard function, and both the model fitting and prediction accuracy was improved with them. Leow and Crook (2016) built dynamic default prediction models with macroeconomic variables,

covering the period of the financial crisis. In this way, they found that the estimated parameters were not stable, given the same changes in macroeconomic variables. As clearly stated in the BASEL II Accord, credit risk measurement is affected by macroeconomy, thus should be taken into account in the modeling process (Bonfim 2009; Carling et al. 2007; Nam et al. 2008). This article also considers a range of those consumer-related macroeconomic factors which have appeared in previous literature.

As P2P lending is a new trend in finance, there are a small but growing number of studies which have started to pay attention to the credit risk inherent in peer-matched borrowing and lending. Emekter et al. (2015) firstly used logistic regression to assess how a borrower's credit grade, debt-to-income ratio, FICO score and revolving credit utilization ratio are associated with the default risk; and secondly used Cox proportional hazard models with these variables to show that default risk increases over the loan age or duration. Malekipirbazari and Aksakalli (2015) employ a random forest methodology in classifying good and bad loans, and compare the results with those from Support Vector Machines, Logistic Regression and K-Nearest Neighbour. However, neither of these two studies have validated their predictions in an out-of-sample format. Guo et al. (2016) used cross validation to test the performance of their instance-based model in credit assessment of P2P loans.

It is obvious that the behaviors of online loans are not the same as mortgages, though both of them mark the importance of prepayment and default risks. In this article, we use real unsecured consumer loan data, observed on a monthly basis, to predict remarkably high prepayment and default risk in online lending. As indicated by the previous literature, multinomial logistic regression is capable of modeling both, so we have delineated the methodology in the following section.

Methodology

Literature related to estimating loan default and prepayment competing risks includes Deng, Quigley, and van Order (2000), Pavlov (2001) and Ambrose and Sanders (2003). These studies all adopted proportional hazard models where time to event is incorporated as a function of duration time. Hazard models are dynamic analysis while cross-sectional analysis such as multinomial logistic regression is also popular in literature. Clapp, Deng, and An (2006) presented evidence that the multinomial logistic regression model is an attractive alternative to proportional hazard models in a case of mortgage termination, by either refinancing, moving or defaulting. Clapp, Deng, and An (2006) further incorporated a random effect into the multinomial logistic regression model to better understand the unobserved heterogeneity of borrower characteristics. In this study, we are interested in predicting the occurrence of the events rather than when they will take place. Therefore, we make use of the outcome of loans by considering a multinomial logistic regression model, which takes multiple states of results into account throughout the modeling process.

In our case, there are two events at each observation point: default and prepayment. Multinomial logistic regression treats the outcome as a discrete choice variable. It assumes mutual independence of choices for a given record during an observation period. In contrast to the proportional hazard model, which first considers the joint survival probabilities and then estimates the conditional probability of each choice, multinomial logistic regression directly estimates the probabilities of each outcome which sum up to one. The default and prepayment risk can be estimated directly. The dependent variable $d_i = \{0, 1, 2\}$ has three states of loan i at period t, where $d_i = 0$ indicates that the loan is fully paid, $d_i = 1$ if the loan is in default, and $d_i = 2$ if it is prepaid. A vector of observable independent variables is given as $\mathbf{x}_i = (\mathbf{w}_i, \boldsymbol{\gamma}_i, \mathbf{z}_i)$, including borrower characteristics $\mathbf{w}_i$, loan features $\boldsymbol{\gamma}_i$ and macroeconomic factors $\mathbf{z}_i$. This setting is in line with Kelly and O'Malley (2016). A lag of three periods is applied to the model, so we actually use the information of three periods prior to the occurrence of interested events. State '0' represents the reference category of fully paid loans for the response variable d_i, and then the multinomial logistic regression model is to fit the ratio of the expected proportion for each response category over the expected proportion of the reference category, where the logit functions are defined as

$$\ln\left(\frac{P(d_i=j)|\mathbf{x}_i)}{P(d_i=0)|\mathbf{x}_i)}\right)=\boldsymbol{\beta}_j^T\mathbf{x}_i,\quad j=1,2 \tag{1}$$

where j is the index of the outcome events and $\boldsymbol{\beta}_j$ represents the vector of the parameter estimates of the corresponding logit functions defined in. The estimated probabilities of each outcome category conditioning on the covariates are given as

$$\begin{aligned}P(d_i=0)|\mathbf{x}_i)&=\frac{1}{1+\sum_{j=1}^{2}\exp(\boldsymbol{\beta}_j^T\mathbf{x}_i)}\\ P(d_i=j)|\mathbf{x}_i)&=\frac{\exp(\boldsymbol{\beta}_j^T\mathbf{x}_i)}{1+\sum_{j=1}^{2}\exp(\boldsymbol{\beta}_j^T\mathbf{x}_i)}\end{aligned} \tag{2}$$

The parameters can be estimated using maximum likelihood where the log-likelihood function is

$$\ln L=\sum_i \ln(P(d_i|\mathbf{x}_i)) \tag{3}$$

Finally, for a given record the predicted category is found to be

$$\hat{j}=\arg\max\left\{\hat{P}(d_i=j),\ \ j=0,1,2\right\} \tag{4}$$

Data and Variables

We collected data from the market-leading P2P company Lending Club at the beginning of 2017. The original data publicly available on its website contains both 36-month and 60-month unsecured loans, issued between 2007Q3 and 2016Q4. In this article, we have only focused on the 36-month loans, because the 60-month term loan issued in 2012 have not yet fully matured. For each loan, all monthly repayment information is included in the data. In total, there are 865,724 loans with 14,021,777 loan-month records for 36-month loans. However, due to the development of Lending Club, it has made significant policy change in 2012. We notice a dramatic increase of loan volume since 2012, and the number of loans in 2012 and 2013 account for 81% of all sample (right axis in Figure 1). So, we exclude all samples prior to 2012 and those censored loans with unobserved outcomes. Finally, a total of 140,605 loans issued in 2012 and 2013 with definite outcomes (fully paid, prepaid, charged off) have been retained.

At Lending Club, the base interest rate is set to be 5.05%, regardless of the loan amount or term. Each loan is graded by Lending Club's internal credit scoring model from A to G, taking into account not only the issued amount, but also information submitted by the borrowers, and then each grade is further broken into five subgrades, based on its internal credit ranking. The interest premium, which factors in credit risk and market volatility, is mapped into each subgrade. The interest rate is given as a sum of the base rate and the premium, and is fixed over the whole term. Payment records such as the monthly payments and total payments received to date are included in the data. Lending Club updates the status of each loan on a regular basis, based on its performance. For example, a 'current' loan indicates that the loan repayments are up to date. If no further payments are expected to be made, a loan is considered to be in default and will be charged off. Therefore, a loan is terminated subject to it being either fully paid off or charged off. Prepayment is defined as a loan being paid off before its maturity, namely the 36th month.

For convenience of analysis we group all the loans into three sections, in terms of their final outcomes: Default, Fully Paid and Prepayment. All the charged-off loans are grouped into the

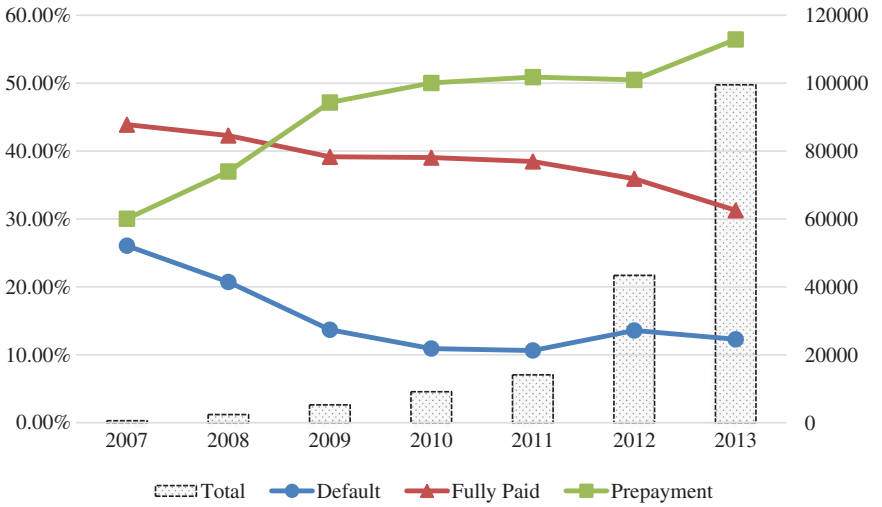

Figure 1. Performance and number of loans across all years.

'Default' category. Loans that are fully paid off before maturity are grouped into the 'Prepayment' basket, and the remaining loans are fully paid off. To observe the historical change of loan performance across all years, we draw a figure according to the issue year in Figure 1. There were obviously increasing numbers of prepaid loans, accounting for over a half in recent years. In the early years, due to the financial crisis the default rate kept above 20% and in 2012 and 2013, it stayed around 13%.

Table 1 shows the cross-tabulation between the loan sections and their grades. Small numbers of highly risky loans, namely Grades E, F and G, are grouped together (marked as E-), and they contribute to 3% of the total sample. It is observed that the default rate increases rapidly from 5.32% to 23.59% as the loan grade deteriorates, which confirms that Lending Club's rating system is reasonable. On the contrary, the prepayment rates of each grade show an opposite pattern, in that it decreases along with the decrease of the quality of the loan. 58.31% of the Grade A loans were repaid early, i.e. before the maturity, while that figure for Grade E- is 49.78%. This motivates us to make further investigations into the prepayment behaviors, namely into which factors are driving the prepayment of P2P loan borrowers.

To facilitate closer scrutiny on default and prepayment behaviors, we exhibit the density distributions of Months On Book (MOB) for each grade in Figure 2. In terms of the MOB of those loans entering into default, Figure 2 shows that the defaults are accumulated rapidly after having been issued for 5 months, and in general the default risk of a loan peaks at the 18th month on book and gradually declines after that. There is still a small portion of 1.8% of charged-offs which occurred in the 36th month.

The second histogram in Figure 2 shows a totally difference pattern for prepayment as its probability gradually increases over time. What draws our attention is that in the 1st and 35th months, the probability density is outstanding. In the latter case, borrowers show no patience to wait until the last month to clear the balance but prefer to pay it off when the loan is approaching maturity. In the

Table 1. Loan performance by grade.

	A	B	C	D	E-	Total
Default	5.32%	10.28%	15.68%	20.39%	23.59%	17,449
Fully paid	36.36%	33.26%	29.60%	27.95%	26.63%	45,074
Prepayment	58.31%	56.45%	54.71%	51.66%	49.78%	78,082
Total	27,414	55,969	33,615	18,995	4,612	140,605

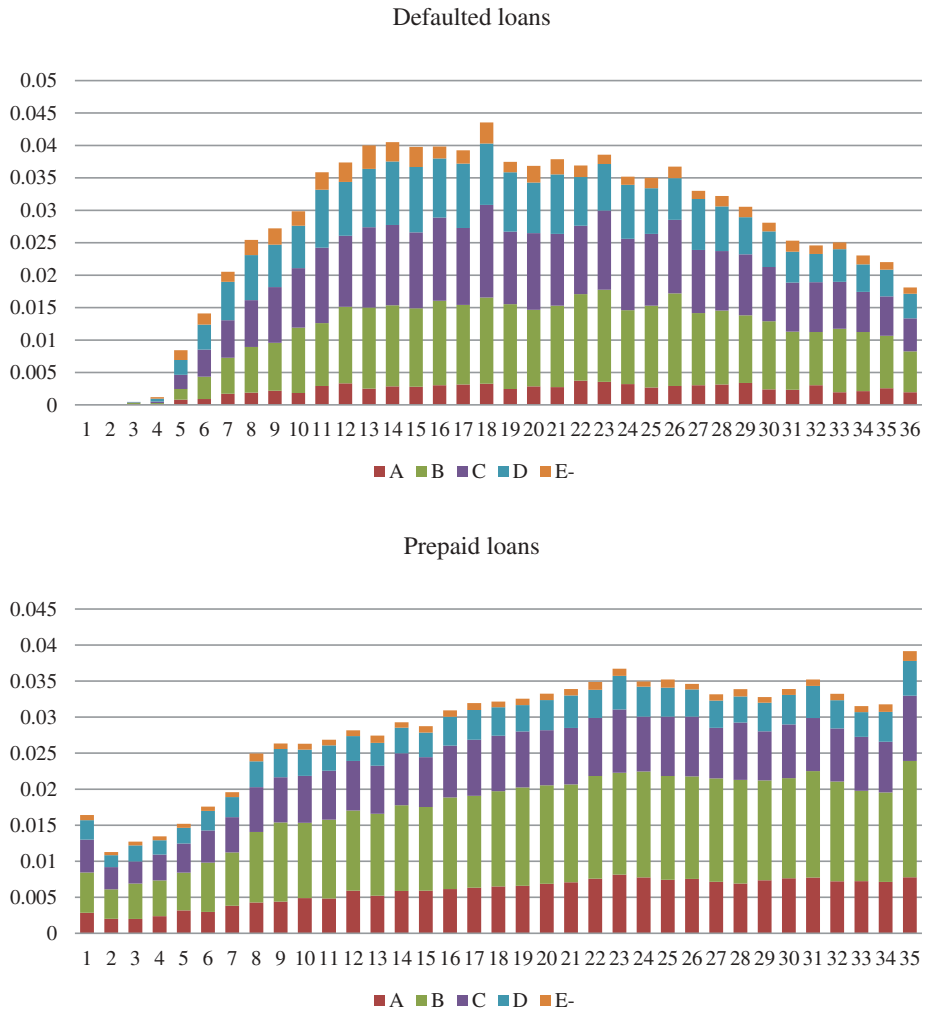

Figure 2. Probability density distribution of month-on-book.

former case, there is a noticeable amount of prepayments taking place within the first month where MOB is equal to 1. It seems as if borrowers return the money shortly after they receive it, either changing their minds or having no further demands for credit. A reasonable explanation for this phenomenon is that these loans work as emergency funds for the borrowers, which is similar to how payday loans are used. Payday loans are a type of short-term unsecured loans as cash advances which are popular in the US. That no penalty is charged for prepayment by the lending platform probably attracts borrowers of emergent financial needs. In general, the volume of prepayments of 36-month loans is significant compared to those for personal loans with traditional banks where the latter has to keep profitability of their businesses. Because Lending Club does not charge any additional fee for prepayment, borrowers have an incentive to pay off the loans early, to save the payments from accruing interest if their personal circumstances change. Different grades of loans also accounted for different portions in the frequency bars in the charts.

Independent variables of predictive potential are included in the modeling, which are grouped into three types: borrower characteristics, loan features, and macroeconomic factors. Debt to Income, Employment Length and Home Ownership, and a range of other credit related variables describe a borrower's characteristics. We also include the loan specific information and macroeconomic factors

Table 2. List of variables.

Variables	Definition
Borrower characteristics	
Debt to income	A ratio of the borrower's total monthly debt other than the loans with Lending Club to the borrower's self-reported monthly income;
Employment length	The borrower's length of employment, ranging from 0 to 10, where 0 indicates a length of less than 1 year and 10 indicates a length longer than or equal to 10 years;
Home ownership	A categorical variable indicating the borrower's residential status: values own, rent, or mortgage; two dummies are used to represent them and the null group is "Own".
Open credit lines	The number of open credit lines in the borrower's credit file;
Total credit lines	The total number of credit lines currently in the borrower's credit file;
Revolving credit utilization	A ratio of the amount of credit the borrower is using relative to all available revolving credit;
Public record	The number of derogatory public records;
Last FICO	The last FICO score of the borrower;
Tax lines	Number of tax lines;
Charge-off within 12 months	Number of charge-offs in the last 12 months;
Public record bankruptcies	Number of public record bankruptcies;
Inquiries 6 months	The number of inquiries in past 6 months;
Loan features	
Interest rate	The interest rate of the loan in percentage;
Payment due to income	A ratio of the monthly due payment to monthly income;
Grade	Loan grade (A, B, C, D, E-) assigned by Lending Club. Four dummies are used to represent them and the null group is "A".
Macroeconomic factors	
GDP growth rate	The GDP growth rate of the US;
Federal funds	The Federal funds rate;
Bankruptcy fillings	The log-ratio of monthly consumer bankruptcy in bank filings (personal, family, or household purpose).

which have a direct impact on consumer behavior. Macroeconomic (systemic) factors are extracted from the CEIC Global Database, which provides economic statistics for many countries. The macroeconomic values are matched to the monthly record by date. The detailed definitions of all independent variables are listed in Table 2.

The descriptive statistics of variables above are presented in Tables 3 and 4 shows the frequency of Home Ownership. In the sample, mortgaged and rented properties share the largest portion of all home ownership types. We also notice a smaller Debt-to-Income ratio in prepaid and fully paid loans than that in defaulted loans. Defaulters use more revolving credit compared to other people, and they have a significantly lower FICO score (on average 618) than others. Other characteristics have no marginal differences except for inquiries as we see defaulted borrowers have a greater number of inquiries in the last 6 months.

Analysis

The whole sample of 140,605 loans is randomly split up into a training set and a test set with a ratio 2:1. The coefficients of the predictive model are constructed on the training set and predictions are made on the test set as out-of-sample validation. The multinomial logistic regression model is estimated with an iterative maximum likelihood method. The variables left in the equation are presented in Table 5.

We now examine the determinants that impact the likelihood of loan default and prepayment compared to fully paid. In the basic characteristics of borrowers, it is evident that Debt to Income

Table 3. Descriptive statistics of variables.

	Default		Fully paid		Prepayment	
	Mean	Std	Mean	Std	Mean	Std
Borrower characteristics						
Debt to income	18.020	7.659	16.885	7.620	16.281	7.539
Employment length	5.657	3.607	6.013	3.615	5.740	3.615
Open credit lines	10.885	4.618	10.770	4.540	10.890	4.576
Total credit lines	22.945	10.853	22.876	10.650	24.716	11.171
Revolving credit utilization	0.609	0.222	0.588	0.227	0.562	0.233
Public record	0.107	0.371	0.094	0.352	0.115	0.381
Last FICO	618.188	61.117	697.914	59.337	709.882	46.156
Tax lines	0.013	0.176	0.012	0.172	0.014	0.173
Charge-off within 12 months	0.005	0.082	0.003	0.063	0.006	0.089
Public record bankruptcies	0.087	0.292	0.075	0.272	0.093	0.305
Inquiries 6 months	0.933	1.082	0.678	0.951	0.800	1.022
Loan features						
Interest rate	14.803	3.713	12.800	3.867	12.916	3.868
Payment due to income	0.175	0.107	0.082	0.044	0.079	0.041
Macroeconomic factors						
GDP growth rate	2.278	0.652	2.054	0.670	2.317	0.638
Federal funds	9.732	6.383	19.635	10.573	9.709	6.350
Bankruptcy fillings	–0.029	0.067	–0.011	0.055	–0.029	0.066

Table 4. Home ownership.

	Default	Fully paid	Prepayment	Total
Mortgage	40.08%	48.14%	50.01%	67,737
Own	9.07%	9.34%	7.97%	12,016
Rent	50.85%	42.52%	42.02%	60,852
Total	17,449	45,074	78,082	140,605

will have a positive significance for default behavior, since a borrower that has a higher Debt-to-Income ratio is more likely to default. Employment Length shows negative impact on prepayment only. Loans with mortgage or rented properties are also more likely to be defaulted. These results are generally consistent with the expectations in credit scoring.

The numbers of open credit lines, all credit lines and derogatory public records for a borrower have similar impacts on both logit models. Open credit lines refer to those of installment and revolving features such as credits with utility companies. If the number of open credit or total credit lines that a borrower holds is large, the borrower may either default or choose to prepay. This can be explained by the notion that the borrower may either carry too heavy a debt burden, or that the ability to pay back credit is good enough. If the revolving line utilization rate is low, the loan is more likely to be fully paid in both models. Low revolving utilization indicates that borrowers used to pay in full without incurring finance charges. The FICO score is also strongly correlated with both response outcomes though with different signs. A higher FICO score implies better credit quality, leading to a higher likelihood to fully pay off or prepay, and the borrower with a lower FICO score tends to default, which is consistent with Agarwal and Taffler (2008). However, in further analysis on the purpose of loan in the loan request statement, we find the primary reason for taking a P2P loan is to pay back credit cards or other debts. These borrowers of

Table 5. Model output.

	Fully paid vs. default				Fully paid vs. prepayment			
	Coefficient	**Error**	***p*-Value**	**Ave. marginal effect**	**Coefficient**	**Error**	***p*-Value**	**Ave. marginal effect**
Intercept	18.043***	0.285	<0.0001	−0.7032	0.069	0.179	0.6999	−2.9932
Borrower characteristics								
Debt to income	−0.022***	0.002	<0.0001	−2.9569	−0.005***	0.001	0.0001	−2.9650
Employment length	−0.001	0.005	0.8169	−2.9905	−0.027***	0.003	<0.0001	−2.9938
Home ownership mortgage	0.264***	0.061	<0.0001	−2.9860	0.123***	0.034	0.0002	−3.0814
Home ownership rent	0.200***	0.060	0.0008	−2.9657	0.013	0.034	0.7015	−2.9936
Open credit lines	−0.027***	0.005	<0.0001	−2.9780	−0.030***	0.003	<0.0001	−2.9299
Total credit lines	0.035***	0.002	<0.0001	−3.1575	0.027***	0.001	<0.0001	−3.0136
Revolving credit utilization	−0.718***	0.083	<0.0001	−3.2817	−0.515***	0.046	<0.0001	−2.9746
Public record	0.195	0.156	0.2108	−2.9614	0.042	0.091	0.6457	−2.9742
Last FICO	−0.025***	0.000	<0.0001	−2.9616	0.006***	0.000	<0.0001	−2.9474
Tax lines	0.098	0.177	0.5799	−2.9266	0.064	0.107	0.5511	−2.9010
Charge-off within 12 months	0.315*	0.125	0.0116	−0.2653	0.449***	0.131	0.0006	−0.2464
Public record bankruptcies	0.477**	0.166	0.0041	−0.4261	0.384***	0.097	<0.0001	−0.2414
Inquiries 6 months	0.128***	0.017	<0.0001	−0.1280	0.006	0.010	0.5283	−0.5571
Loan features								
Interest rate	0.229***	0.015	<0.0001	−0.2533	0.298***	0.010	0.0001	−0.2341
Payment due to income	24.200***	0.318	<0.0001	−0.0739	−1.015***	0.215	<0.0001	−0.5919
Grade B	0.227**	0.082	0.0058	−0.3094	0.317***	0.041	<0.0001	−1.0104
Grade C	0.339**	0.124	0.0061	−0.1064	0.410***	0.066	<0.0001	−0.1565
Grade D	0.503**	0.167	0.0027	−0.2533	0.439***	0.092	<0.0001	−0.1502
Grade E	0.705**	0.222	0.0015	−0.1194	0.547***	0.126	<0.0001	−0.1097
Macroeconomic factors								
GDP growth rate	−1.318***	0.031	<0.0001	0.0772	−0.910***	0.019	<0.0001	−0.2324
Bankruptcy fillings	−2.298***	0.258	<0.0001	−0.2643	−2.277***	0.144	<0.0001	−0.2145
Federal funds	−0.183***	0.003	<0.0001	−0.2510	−0.163***	0.001	<0.0001	−0.2570

*, **, and *** indicate that the coefficients are significant at 5%, 1%, and 0.1% level, respectively.

high FICO scores are at the same time more likely to prepay the loan from other loan sources. Other credit bureau information such as public record bankruptcies and inquiries is positively related to default, which is in line with the common sense.

The loan features are all significant in identifying the default and prepayment risk. It is noted that interest rates are positively related to the default and prepayment risk. When the interest rate increases borrowers prefer to prepay loans, to avoid higher interest payments in future; lower interest implies that borrowers are more likely to keep the loan for longer time period. For default risk, a higher interest rate increases the rate of default, as it increases the burden of the loan and vice versa. The monthly due payment to income has a positive impact on the probability of default, suggesting that a higher ratio indicates a greater chance to default. On contrast, if the due payment only takes a small portion of income, i.e. the smaller the ratio is, the greater change the borrower will prepay the loan, which means the borrower has better affordability. Affordability is usually a measure of a borrower's ability to repay the loan burden compared to the borrower's income and it is frequently used in the mortgage assessment (Kelly and O'Malley 2016). In here, we find affordability has not only impact on default but also on prepayment behavior.

Finally, we find all the macroeconomic factors are positively significant in all the logit models. The GDP growth rate is negatively correlated with both default and prepayment. A developing economy reduces the chance of default as well as the chance of prepayment. The Federal funds base rate is positively correlated with default and prepayment, as it increases the cost of finance. The bankruptcy rate shows the average default rate of all bank consumer loans and so by definition is positively correlated with default. Compared to full payment as scheduled, at times of smaller personal bankruptcy rates, people tend to repay a loan in advance.

The predicted frequencies across all grades are clear compared to what are observed in Figure 3. The predicted numbers of fully paid loans are consistently lower than the true values. This is similar for default while for prepaid loans, there are more predicted loans than observed loans. In the confusion matric in Table 6, we found that on average classifications are shifted to prepayment, probably due to the unbalanced proportions of three outcomes, which is common in credit scoring applications (Crone and Finlay 2012).

Table 6 also shows the predictive accuracy. We define the accuracy of each category as the number of correctly classified records divided by the relevant total number of each row. Overall, the model produces good predictions, 75.12% cases in the training set and 74.93% cases in the test set being accurately classified. Upon closer inspection, the prepayment group accounts for the majority of correctness, but only half of fully paid loans are discovered. This is reasonable as both behaviors indicate the loans are technically paid off. The predictions of default are relatively good.

Further breaking down the predictive accuracy of the test set in Table 7, we find the performance varies slightly across all grades. The prediction of default in Grade A loans significantly underperformed the average with just above 50% correctness. Overall 76.63% of Grade D loans can be grouped into the true baskets. Generally, multinomial logistic regression works fine in distinguish three possible outcomes of online leading loans.

Conclusion

Prepayment and default are two of the main events leading to the termination of loan repayment and also loss of profit for creditors. However, whether in practice or in academics, credit scoring models have to a large extent only been used to evaluate the default risk of applicants, but have neglected the importance of prepayment. The new trend currently taking place in the banking and lending industry is turning risk scoring to profit scoring, where scorecards are built to maximize the profitability of customers and accounts (Crook, Edelman, and Thomas 2007). Prepayment obviously has a great impact on the lifetime value of customers, as early repayment accrues no further payable interest. This has been a particularly notable phenomenon in the online lending market, as borrowers in urgent need of money seek instant

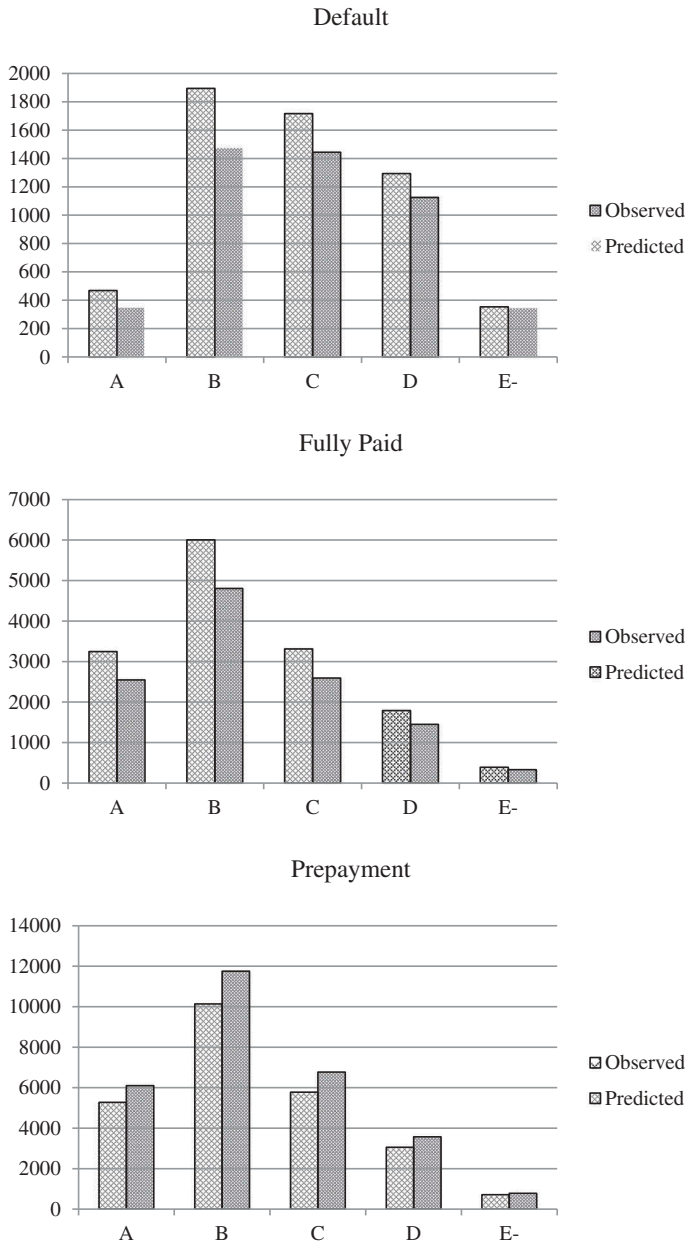

Figure 3. Predicted frequencies of three outcomes.

finance online. Various lending companies including P2P platforms currently charge no fees for this type of behavior, which stimulates a much higher rate of prepayment than default.

Unlike previous literature, which has mostly concentrated on mortgage loans, this article addresses this problem in the context of unsecure consumer loans matched online between a borrower and a number of lenders. We use multinomial logistic regression to model the three levels of outcomes of a loan: fully paid, prepayment and default. Given the observable information of borrower characteristics and loan features, combined with the influence of macroeconomy, both default and prepayment can be accurately predicted, although the predictive performance for default is slightly poorer than that for

Table 6. Prediction results.

		Actual				
		Default	**Fully paid**	**Prepayment**	**Total**	**Accuracy**
Training set						
Predicted	Default	7,376	820	3,025	11,221	65.73%
	Fully paid	1,062	17,137	11,069	29,268	58.55%
	Prepayment	838	5,162	41,849	47,849	87.46%
	Total	9,276	23,119	55,943	88,338	75.12%
Test set						
Predicted	Default	3,756	427	1,553	5,736	65.48%
	Fully paid	514	8,729	5,850	15,093	57.83%
	Prepayment	410	2,636	21,558	24,604	87.62%
	Total	4,680	11,792	28,961	45,433	74.93%

Table 7. Predictive accuracy across all grades.

	A	**B**	**C**	**D**	**E-**
Default	53.07%	63.24%	66.74%	71.32%	67.71%
Fully paid	54.86%	57.74%	58.19%	61.47%	64.78%
Prepayment	86.06%	88.12%	88.05%	87.98%	86.50%
Total	72.59%	75.19%	75.32%	76.63%	75.98%

prepayment. We found that high interest rates not only indicate large probability of default but also increase the probability of prepayment as borrowers do not wish to bear high interests. The borrower characteristics such as the debt-to-income ratio and the FICO score have significant impact on both outcomes, where a large FICO score implies that the borrower has a large chance to repay the loan early. Macroeconomic factors such as GDP growth, Federal fund rates and personal bankruptcy rates can influence the occurrence of two events. These systemic factors matter in maintaining the stability of financial systems.

Online P2P lending as a fast and low-cost source of finance is becoming popular among borrowers. In the reasons listed in the loan request form, repaying credit cards is one of the main purposes for a P2P loan, accounting for 21% of the total. This implies that many borrowers do not use the funds in consumption such as buying a product but as a remedy to mitigate their financial stress. Different financial instruments are interlinked to potentially affect the stability of all stakeholders. Considering that the volume of payday loans has grown rapidly in recent years, with expensive annualized percentage rates (Bhutta 2014), borrowers seem to be using P2P lending as an alternative to short-term and high-cost payday loans. Without penalties, P2P lending is typically used as a short-term but low-cost loan, even though it is designed for 36-month or longer terms. Prepayment is therefore much more likely to happen and there may be an arbitrage opportunity of abusing P2P loans. Regulators have to pay attention to it, as interlinked credit lines by early settlement and refinancing will cause problems for the financial system. It is suggested that lenders should address this and the platform may consider charging a penalty for prepayment, in order to compensate potential losses to their underlying loan portfolios. P2P companies such as Lending Club now repackage the loan portfolios and resell them to other financial institutions. As disclosed on Lending Club's website, there are increasing volumes of online loans funded by financial institutions. Its appropriate pricing is important for investors and regulators who want to avoid any more disasters like the subprime crisis.

This study provides the foundation of accurate pricing of online loan portfolios. The data is for both 36-month and 60-month term loans, though we only include the 36-month term ones in analysis because the early 60-month loans have not yet reached maturity. However, we do notice that the prepayment rate in the 60-month loans is even higher than for 36-month loans, which means the term of maturity may make an impact. It is of interest to use survival analysis to model the duration time on book, or from the rather novel perspective, as suggested by Stepanova and Thomas (2002), to model the remaining time to maturity. Stepanova and Thomas (2002) found the remaining time to maturity has an impact on the probability of prepayment. Survival analysis dealing with multi-period data will further extend the scope of this research.

Highlights

- Large portions of prepaid and defaulted loans are observed in online lending.
- Prepayment and default are found to have different patterns.
- A multinomial logistic regression model is used to predict both events.
- Both prepayment and default can be accurately predicted by a range of variables.
- Out-of-sample validation is given for unsecure consumer loan data.

References

Agarwal, S., B. W. Ambrose, and C. Liu. 2006. Credit lines and credit utilization. *Journal of Money, Credit and Banking* 38 (1):1–22. doi:10.1353/mcb.2006.0010.

Agarwal, V., and R. Taffler. 2008. Comparing the performance of market-based and accounting-based bankruptcy prediction models. *Journal of Banking & Finance* 32(8):1541–51. doi:10.1016/j.jbankfin.2007.07.014.

Ambrose, B. W., and A. B. Sanders. 2003. Commercial mortgage-backed securities: Prepayment and default. *The Journal of Real Estate Finance and Economics* 26(2–3):179–96. doi:10.1023/A:1022978708728.

Andreeva, G., J. Ansell, and J. Crook. 2007. Modelling profitability using survival combination scores. *European Journal of Operational Research* 183(3):1537–49. doi:10.1016/j.ejor.2006.10.064.

Banasik, J., J. N. Crook, and L. C. Thomas. 1999. Not if but when will borrowers default. *Journal of the Operational Research Society* 50(12):1185–90. doi:10.1057/palgrave.jors.2600851.

Basel Committee on Banking Supervision (BCBS). Principles for the Management of Credit Risk.2000.

Bellotti, T., and J. Crook. 2008. Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society* 60(12):1699–707. doi:10.1057/jors.2008.130.

Bhutta, N. 2014. Payday loans and consumer financial health. *Journal of Banking & Finance* 47:230–42. doi:10.1016/j.jbankfin.2014.04.024.

Bhutta, N., P. M. Skiba, and J. Tobacman. 2015. Payday loan choices and consequences. *Journal of Money, Credit and Banking* 47(2–3):223–60. doi:10.1111/jmcb.12175.

Blöchlinger, A., and M. Leippold. 2006. Economic benefit of powerful credit scoring. *Journal of Banking & Finance* 30(3):851–73. doi:10.1016/j.jbankfin.2005.07.014.

Bonfim, D. 2009. Credit risk drivers: Evaluating the contribution of firm level information and of macroeconomic dynamics. *Journal of Banking & Finance* 33(2):281–99. doi:10.1016/j.jbankfin.2008.08.006.

Carling, K., T. Jacobson, J. Linde, and K. Roszbach. 2007. Corporate credit risk modeling and the macroeconomy. *Journal of Banking & Finance* 31(3):845–68. doi:10.1016/j.jbankfin.2006.06.012.

Carling, K., T. Jacobson, and K. Roszbach. 2001. Dormancy risk and expected profits of consumer loans. *Journal of Banking & Finance* 25(4):717–39. doi:10.1016/S0378-4266(00)00093-5.

Ciochetti, B. A., Y. Deng, B. Gao, and R. Yao. 2002. The termination of commercial mortgage contracts through prepayment and default: A proportional hazard approach with competing risks. *Real Estate Economics* 30(4):595–633. doi:10.1111/1540-6229.t01-1-00053.

Ciochetti, B. A., Y. Deng, G. Lee, J. D. Shilling, and R. Yao. 2003. A proportional hazards model of commercial mortgage default with originator bias. *The Journal of Real Estate Finance and Economics* 27(1):5–23. doi:10.1023/A:1023694912018.

Clapp, J. M., Y. Deng, and X. An. 2006. Unobserved heterogeneity in models of competing mortgage termination risks. *Real Estate Economics* 34(2):243–73. doi:10.1111/j.1540-6229.2006.00166.x.

Crone, S. F., and S. Finlay. 2012. Instance sampling in credit scoring: An empirical study of sample size and balancing. *International Journal of Forecasting* 28(1):224–38. doi:10.1016/j.ijforecast.2011.07.006.

Crook, J. N., D. B. Edelman, and L. C. Thomas. 2007. Recent developments in consumer credit risk assessment. *European Journal of Operational Research* 183(3):1447–65. doi:10.1016/j.ejor.2006.09.100.

Deng, Y., J. M. Quigley, and R. van Order. 2000. Mortgage terminations, heterogeneity and the exercise of mortgage options. *Econometrica* 68(2):275–307. doi:10.1111/1468-0262.00110.

Duarte, J., S. Siegel, and L. Young. 2012. Trust and Credit: The role of appearance in peer-to-peer lending. *Review of Financial Studies* 25(8):2455–83. doi:10.1093/rfs/hhs071.
Emekter, R., Y. Tu, B. Jirasakuldech, and M. Lu. 2015. Evaluating credit risk and loan performance in online peer-to-peer (P2P) lending. *Applied Economics* 47(1):54–70. doi:10.1080/00036846.2014.962222.
Esteve-Pérez, S., A. Sanchis-Llopis, and J. A. Sanchis-Llopis. 2010. A competing risks analysis of firms' exit. *Empirical Economics* 38(2):281–304. doi:10.1007/s00181-009-0266-x.
Guo, Y., W. Zhou, C. Luo, C. Liu, and H. Xiong. 2016. Instance-based credit risk assessment for investment decisions in P2P lending. *European Journal of Operational Research* 249(2):417–26. doi:10.1016/j.ejor.2015.05.050.
Hand, D. J., and W. E. Henley. 1997. Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 160(3):523–41. doi:10.1111/j.1467-985X.1997.00078.x.
Heitfield, E., and T. Sabarwal. 2004. What drives default and prepayment on subprime auto loans? *The Journal of Real Estate Finance and Economics* 29(4):457–77. doi:10.1023/B:REAL.0000044023.02636.e6.
Hwang, R.-C., and C.-K. Chu. 2014. Forecasting forward defaults with the discrete-time hazard model. *Journal of Forecasting* 33(2):108–23. doi:10.1002/for.2278.
Iyer, R., A. I. Khwaja, E. F. P. Luttmer, and K. Shue. 2016. Screening peers softly: Inferring the quality of small borrowers. *Management Science* 62(6):1554–77. doi:10.1287/mnsc.2015.2181.
Kelly, R., and T. O'Malley. 2016. The good, the bad and the impaired: A credit risk model of the Irish mortgage market. *Journal of Financial Stability* 22:1–9. doi:10.1016/j.jfs.2015.09.005.
Leow, M., and J. Crook. 2016. The stability of survival model parameter estimates for predicting the probability of default: Empirical evidence over the credit crisis. *European Journal of Operational Research* 249(2):457–64. doi:10.1016/j.ejor.2014.09.005.
Lin, M. F., N. R. Prabhala, and S. Viswanathan. 2013. Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending. *Management Science* 59(1):17–35. doi:10.1287/mnsc.1120.1560.
Malekipirbazari, M., and V. Aksakalli. 2015. Risk assessment in social lending via random forests. *Expert Systems with Applications* 42(10):4621–31. doi:10.1016/j.eswa.2015.02.001.
Melzer, B. T., and D. P. Morgan. 2015. Competition in a consumer loan market: Payday loans and overdraft credit. *Journal of Financial Intermediation* 24(1):25–44. doi:10.1016/j.jfi.2014.07.001.
Miller, S. 2015. Information and default in consumer credit markets: Evidence from a natural experiment. *Journal of Financial Intermediation* 24(1):45–70. doi:10.1016/j.jfi.2014.06.003.
Nam, C. W., T. S. Kim, N. J. Park, and H. K. Lee. 2008. Bankruptcy prediction using a discrete-time duration model incorporating temporal and macroeconomic dependencies. *Journal of Forecasting* 27(6):493–506. doi:10.1002/for.985.
Pavlov, A. D. 2001. Competing risks of mortgage termination: Who refinances, who moves, and who defaults? *The Journal of Real Estate Finance and Economics* 23(2):185–211. doi:10.1023/A:1011158400165.
Pennington-Cross, A. 2010. The duration of foreclosures in the subprime mortgage market: A competing risks model with mixing. *The Journal of Real Estate Finance and Economics* 40(2):109–29. doi:10.1007/s11146-008-9124-4.
Rigbi, O. 2012. The effects of usury laws: Evidence from the online loan market. *Review of Economics and Statistics* 95 (4):1238–48. doi:10.1162/REST_a_00310.
Stein, R. M. 2005. The relationship between default prediction and lending profits: Integrating ROC analysis and loan pricing. *Journal of Banking & Finance* 29(5):1213–36. doi:10.1016/j.jbankfin.2004.04.008.
Steinbuks, J. 2015. Effects of prepayment regulations on termination of subprime mortgages. *Journal of Banking & Finance* 59:445–56. doi:10.1016/j.jbankfin.2015.07.012.
Stepanova, M., and L. Thomas. 2002. Survival analysis methods for personal loan data. *Operations Research* 50(2):277–89. doi:10.1287/opre.50.2.277.426.
Thomas, L. 2009. *Consumer credit models: Pricing, profit, and portfolios*. New York: Oxford University Press.
Thomas, L. C. 2000. A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. *International Journal of Forecasting* 16(2):149–72. doi:10.1016/S0169-2070(00)00034-0.
Varli, Y., and Y. Yildirim. 2015. Default and prepayment modelling in participating mortgages. *Journal of Banking & Finance* 61:81–88. doi:10.1016/j.jbankfin.2015.09.003.
Zhang, J., and P. Liu. 2012. Rational herding in microloan markets. *Management Science* 58(5):892–912. doi:10.1287/mnsc.1110.1459.