The value of personal credit history in risk screening of entrepreneurs

Evidence from marketplace lending

OPEN ACCESS
Journal of Financial Management, Markets and Institutions
(2021) 2150004 (25 pages)

DOI: 10.1142/S2282717X21500043

THE VALUE OF PERSONAL CREDIT HISTORY IN RISK SCREENING OF ENTREPRENEURS: EVIDENCE FROM MARKETPLACE LENDING

GALINA ANDREEVA
University of Edinburgh Business School
29 Buccleuch Place, Edinburgh, EH8 9JS, UK
Galina.Andreeva@ed.ac.uk

EDWARD I. ALTMAN
New York University Stern Business School
44 W 4th Street, New York, 10002, USA
eia1@stern.nyu.edu

Received 14 December 2019
Accepted 27 May 2021
Published 3 July 2021

We explore the quality of risk assessment for entrepreneurs/small business borrowers as compared to consumers, when the same information on previous credit history is used for both segments in marketplace lending. By building several cross-sectional logistic regression and machine-learning models and applying them separately to small business loans (SBL) and consumers we can measure models' predictive accuracy for different segments, and thus, make observations about the value of the information used for screening. We find the differences in profiles between SBL and consumers, hence they should be assessed by separate models. Yet separate SBL models do not perform well when applied to a future time period. We attribute this to the relatively low predictive value of personal credit history for entrepreneurs as compared to the consumers. We advocate the use of additional information for risk assessment of entrepreneurs, in order to improve the quality of credit screening. This should lead to improved access of small business borrowers to credit in situations when they have to compete with consumers for funding.

Keywords: Small business finance; marketplace lending; risk of default; machine-learning.

JEL Classification: D80, G21, G32

1. Introduction

In recent years much needed external funding for small business ventures and entrepreneurs has become available through novel routes, such as marketplace finance or peer-to-peer (P2P) lending (Ahmed *et al.* 2016, Bruton *et al.* 2015). The difficulties of small businesses in getting funding from traditional banks are well documented, often they have to compete with large corporates (Mills & McCarthy 2014, 2016, Jagtiani & Lemieux 2016). In marketplace lending entrepreneurs in many cases have to compete with consumers, since P2P platforms may serve both types of customers. Furthermore, entrepreneurs often fund their businesses through retail credit products, especially at the early stages of business life.

Whether any credit application is successful, largely depends on risk assessment — the procedure known as screening. Nowadays, and especially with large-scale lending, screening is done by scoring models that rely on historic data (Altman *et al.* 2018, Thomas 2009). One of the main determinants of screening quality is predictive value of the information that goes into the model. The current paper explores this particular topic—the predictive value of personal credit history of entrepreneurs in evaluating their potential creditworthiness in the context of marketplace or peer-to-peer (P2P) lending.

There are numerous studies and well-established models for credit scoring of large corporates (e.g. Altman 1968) and consumers (e.g. Thomas 2009), these two groups of models differ in the type of information that is used for prediction—business models rely mainly on financial accounts and market information, whilst consumer models use mostly personal credit history (such as FICO score) and socio-demographics. For consumers, there also have been studies into determinants of credit risk/default in marketplace lending (e.g. Emekter *et al.* 2015). There is also literature, albeit less voluminous, on risk of small firms (SMEs) using business information (e.g. Altman *et al.* 2010).

However, SMEs are known to be particularly challenging for credit risk assessment because of the scarcity of relevant business information that can be used as an input into risk models (Berger & Frame 2007). SMEs may not have stock prices or detailed financial accounts that can be used by lenders for screening potentially bad loans. The problem is especially acute for start-ups that may not have any business-related information at all. That is why in marketplace lending they can be assessed as consumers using personal information. This practice is also common in traditional lending. Dun & Bradstreet (D&B), the largest credit bureau for businesses, mention that lenders normally request personal credit history for start-up and micro-businesses, although it may be not an accurate indication of business performance (Dun & Bradstreet 2021). However, D&B do not provide any further details and to the best of our knowledge, there are no studies about the information value of personal credit history for risk screening of entrepreneurs, especially in the context of new alternative lenders, and this is the main research question that this paper addresses.

This question is important because with low information value and resulting poor predictive performance of risk models lenders cannot rate credit risks accurately, they grant credit to borrowers that cannot re-pay it, and reject potentially creditworthy applicants. This increases the price of loans, because lenders have to cover the increasing costs from non-performing loans, and more and more good borrowers are excluded from access to credit. This situation has been extensively analyzed in the literature on information asymmetry, adverse selection and moral hazard (Jaffee & Russell 1976, Stiglitz & Weiss 1981).

Our research question becomes even more important in the current situation, when the need for external funds has increased enormously due to the effects of Covid-19 pandemic and lockdown. At the same time traditional bank lending to small businesses has been declining for over a decade at least in the USA (Cole 2018), and during 2020 non-bank lenders and financial-technology companies have also scaled down (Rudegeair 2020). Small businesses have to compete for funding with large businesses and consumers even more.

Against this background, our specific research objectives are as follows:

- What is the predictive value of the personal credit history for risk assessment of entrepreneurs? Can it provide the same level of accuracy as compared to consumers?
- How different are risk profiles of entrepreneurs as compared to consumers? Should these two segments be treated as one and screened with a single model?
- How effective is credit screening of entrepreneurs in marketplace lending, when they are assessed with the same model as consumers? And when their risk is estimated separately?

Our exploration is based on the data from Lending Club (LC), the US largest marketplace lender and one of the biggest in the world. LC provides an on-line platform for individuals and entrepreneurs to post a request for loans and for lenders/investors to offer the money and earn some interest. The personal loan portfolio consists predominantly of individual borrowers seeking a relief from existing debts, including credit cards (Refinance). However, it also includes a small proportion of business/entrepreneurial loans. This allows us to compare small business loans (SBL) to consumers using the same scope of information.

The rest of the paper is structured as follows. Section 2 provides a brief literature review of research on small business and marketplace lending risk. Section 3 presents the data and some descriptive statistics. This is followed by Probability of Default (PD) logistic regression models and their predictive accuracy in Sec. 4. The exploration of differences between SBL and Refinance customers is presented in Sec. 5. Section 6 provides robustness checks, including the comparison of different machine-learning algorithms, as alternative modeling techniques to logistic regression. A detailed discussion of the results is given in Sec. 7, with the final section (Sec. 8) providing conclusions.

2. Literature Review

There are two main strands of literature that our study builds on and contributes to: small business lending/bankruptcy prediction and marketplace lending.

Despite the importance of small business for the economy, they often experience difficulties in obtaining the external finance, which is essential to their survival and growth. Banks prefer lending to large corporates, since small loans are less profitable and more costly to process (Rudegeair 2020). The lack of information on SMEs is often regarded as a major difficulty in evaluating their risk profile. Berger *et al.* (2005), Berger & Frame (2007) comment on the difficulties that lenders face when deciding which small businesses are creditworthy because of scarcity of the relevant information that is indicative of a potential failure or default. The majority of small companies do not have publicly traded equity which is the main source of information used for assessing the corporate credit risk. SMEs financial accounts are often incomplete and non-audited. Berger *et al.* (2005) suggest that personal credit scores of entrepreneurs/business owners can improve risk assessment for informationally opaque firms, especially at the start-up stage. However, they do not investigate the value of such information.

For corporate credit there have been many models previously proposed, their detailed overview is given in Altman & Saunders (1997), Allen *et al.* (2004).

However, a number of studies noted that credit risk of SMEs is different from large corporates and should be assessed by separate models. Dietsch & Petey (2004) show that small businesses are riskier as compared to corporates. Altman & Sabato (2005, 2007) prove that risk assessment models for large corporates do not perform well when applied to SMEs. They demonstrate that banks/lenders should develop models and risk assessment processes for SME portfolios separate from large corporates. Vallini *et al.* (2009), Ciampi & Gordini (2013) arrive to the same conclusion that predicting default for SMEs is more difficult and predictive accuracy is lower as compared to corporates.

Furthermore, Altman *et al.* (2010) show that non-financial and non-accounting information about a business (such as age, industry, previous missed payments) improves the quality of prediction for SME default. Later work by Altman *et al.* (2017) prove that financial and non-financial information remain predictive of default over longer time horizons — up to 10 years. However, this strand of literature does not consider the value of personal credit history of business owners for predicting the non-payment or default.

Our paper also builds on the emerging strand of literature on marketplace lending. Most widely researched questions include what factors are associated with chances of a borrower to be funded; and what factors are associated with the probability of default. Lin *et al.* (2013), Freedman & Jin (2014), Liu *et al.* (2015) find that borrowers' online friendships and closer relationship with lenders have significant influence on the probability of funding success and lower default risk. Duarte *et al.* (2012), Gonzalez & Loureiro (2014) analyze borrowers' pictures and find that

"trustworthy" or "attractive" appearance improves the chances of being funded and decreases default. Larrimore *et al.* (2011) show the effect of the quality and quantity of loan's textual description on funding success. Iyer *et al.* (2016) find that soft or non-standard information, e.g. appearance, acceptable maximum interest rate and textual description, are significantly associated with default and improve prediction of default probability as compared to predictions based only on borrower's credit rating.

As for research based on Lending Club data, Emekter *et al.* (2015) discover that FICO scores and credit grades estimated by LC play the most important role in predicting default. Similarly, Serrano-Cinca *et al.* (2015) find that LC grades are the most significant factors to predict default rates but the accuracy of the model is improved by additional variables, such as loan purpose, annual income, housing ownership, credit history, and borrower's indebtedness.

The above studies analyze samples of consumer loan portfolios. In terms of small business loans (SBL), the research on LC data is scarce. Even if SBLs are included into the sample, they are not analyzed separately. The exception is Mach *et al.* (2014) who provide a detailed investigation of SBLs in LC portfolio from 2007–2012. They find that small businesses are more likely to be funded compared to other loan purposes after controlling for borrower's FICO score, employment, loan amount, year of application, state of residence and House Price Index (HPI) in the state. Small businesses are also more likely to be charged higher interest rates, but this is explained by higher probability of default.

However, the above studies have not investigated how well their models predict, and how predictive accuracy differs between consumers and entrepreneurs. Our research closes this gap.

3. Data Description

The development of internet technologies has led to the emergence of on-line lending platforms, such as Lending Club, Prosper, Zopa that serve as meeting places for those who need some money and those who are willing to invest it. Typically, the marketplace lenders offer small fixed-term loans to individuals and businesses that cannot get loans from traditional sources or cannot obtain them on equally attractive terms.

Lending Club (LC) was among the world's largest marketplace lenders, as of September 30, 2020 the company issued 60,188,236,052 USD in loans (Lending Club 2010).[a] It enabled borrowers to obtain unsecured fixed-term loans with interest rates that they found attractive, and investors to fund loans with credit characteristics, interest rates and other terms the investors found attractive. The platform charged borrowers an origination fee and investors a service fee.

[a]In 2020 Lending Club acquired Radius bank and announced that it would be closing the marketplace platform. https://www.lendacademy.com/lendingclub-closing-down-their-platform-for-retail-investors/.

LC verified the identity of borrowers, obtained their credit profiles from consumer reporting agencies, such as TransUnion, Experian or Equifax, and screened borrowers for eligibility to participate in the platform. The screening was based on the prospective borrower's FICO score,[b] a debt-to-income ratio, a credit profile (as reported by a consumer reporting agency). Lending Club provided their own risk grades, and set the interest rates based on the risk level. In addition, LC made available anonymized detailed information on loans granted and their subsequent credit performance, that sophisticated investors could use to build their own screening models (Lending Club 2010).

LC is ideally suited for our investigation because of the large volume of loans, a business model which is typical to many other marketplace lenders (Jagtiani & Lemieux 2016) and the fact that it lent to both consumers and small businesses. LC had a dedicated platform for SMEs, but businesses needed to be at least 12 months old and to have at least 50,000 USD in annual sales, in order to apply there. Those businesses/entrepreneurs that did not meet these criteria could still apply to the consumer loan platform, and it is these businesses that we are interested in.

In this paper, we concentrate on year 2012. The choice of the year was prompted by the desire to include loans with the duration of 60 months, the latest year when such loans reach maturity was 2017 (at the time of data download), which meant that they had applied in 2012.

Small business loans (SBL) constitute a very small proportion in LC consumer credit portfolio as can be seen from Fig. 1. Figure 1 shows that LC consumer portfolio is dominated by "debt consolidation" and "credit card payoff" categories, i.e. individuals that seek restructuring of their existing debt in hope of getting better terms. For subsequent analysis we combine these two categories together into one large group of "Refinance" that constitute 77% of the consumer portfolio, since both categories have the same goal of seeking debt relief.

The share of SBLs in 2012 is 2.6% or 1386 loans, and their default rate is the highest one — 25.47%. The default is defined as Charged Off or 16 + Days Past Due (DPD).

There are two loan maturities/terms: 36 and 60 months. It is known that longer durations attract riskier borrowers (Hertzberg *et al.* 2018) and this is evident from Table 1. For SBLs the share of "60 months" loans in 2012 is higher (21.14% as compared to 18.55%) and the Default Rate is higher too (43.69% versus 27.53% — Refinance 60 months).

There are several possible explanations why entrepreneurial loans demonstrate such a high default rate. In this paper, we do not aim at answering the question what are the reasons leading to default, instead we focus on exploring one possible explanation: the screening of loans is less effective for SBL given the information available for risk assessment.

[b] A FICO score by Fair Isaac Corporation is a numeric credit risk rating of consumers that ranges between 300 and 850. Higher scores correspond to a lower risk of defaulting on credit obligations.

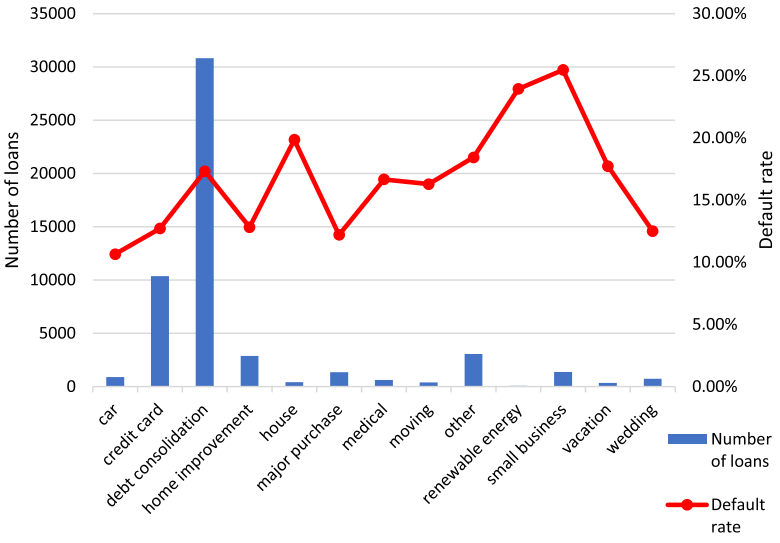

Fig. 1. Number of loans and default rate by loan purpose, 2012. The left-hand side axis and bars refer to the numbers of funded loans for different purposes. The right-hand side axis and lines show the percentage of defaulted loans within each category, or default rate.

Table 1. Default Rates by maturity/loan term (all loans, refinance and SBL), 2012.

Term/maturity	No default	Default	Total	Share in total # of loans, %	Default rate, default/total, %
All loans					
36 months	37567	5903	43470	81.45	13.58
60 months	7172	2725	9897	18.55	27.53
Refinance (Debt consolidation + credit card payoff)					
36 months	28952	4502	33454	83.84	13.46
60 months	5568	2154	7722	16.16	27.89
Small Business Loans (SBL)					
36 months	868	225	1093	78.86	**20.59**
60 months	165	128	293	21.14	**43.69**

To answer this question, one can estimate Probability of Default using the information/variables available to LC at the point of application and measure the accuracy of prediction. To put this measure of predictive accuracy into the context, one can compare it to the equivalent measure for other loan purposes. If the predictive accuracy is worse for SBLs, this would indicate less effective screening and low information value of personal credit history for risk assessment of entrepreneurs. We use "Refinance" as the main benchmark group for comparison because (1) this is the largest group, other loan purposes are too small and idiosyncratic to allow for robust modeling and comparison; (2) these customers are typical to marketplace platforms,

e.g. LC stated as a main goal on its website: "Our LC™ Marketplace Platform helps borrowers take control of their debt and empowers everyone to reach their financial goals".[c] In general, Chava & Paradkar (2018) claim that for marketplace borrowers in the U.S., more than 70% state that their primary reason for requesting funds is to pay off more expensive debt by replacing it with cheaper monthly re-payments.

4. Probability of Default (PD) Modeling

As a first step, we model PD using the LC risk sub-grades (their internal credit risk ratings that are available to investors), and dummy variables indicating if the account is SBL or Refinance in comparison to other loan purposes which are combined into the reference category. At this stage we use the whole dataset for 2012 with all loan purposes, the dependent variable is observed loan status (Default = 1), as illustrated in Fig. 1. If dummy variables are statistically significant, it would mean that there are omitted or unobserved predictors of default and the dummy variables proxy for them. This will give the evidence as to whether LC grades contain sufficient information to describe the PD risk for SBL and Refinance. We also control for loan duration.

In this section, we follow the principles of credit risk modeling for consumers (credit scoring) as described in Anderson (2007), Siddiqi (2006), Thomas (2009). Logistic regression is the most widely used algorithm in consumer credit risk, therefore, it is used in this section. The parameter estimates of this first model are given in Table 2.

The dummy for SBL is highly significant and positive, meaning that SBL have odds of default almost twice higher compared to other loans after controlling for LC risk sub-grades and loan duration. Since the dummy is significant, there is some unobserved information associated with default for small businesses. One can say that the risk of SBLs is not adequately captured by the information contained in risk grades. The dummy for Refinance has a negative sign, but is significant only at 10% level, implying the risk of refinancers is more adequately represented. The Odds ratio is close to 1, therefore, Refinance is not significantly different from other loan purposes.

We then focus on more detailed information which is available to investors when they decide which project to invest, in order to understand the level of accuracy that can be achieved when modeling default for Refinance and SBL; and statistically significant risk predictors for each segment. The characteristics that are available for default prediction are mainly credit bureau variables, and some limited information about the loan (term, amount, purpose) and borrower (home ownership, employment length, annual income). The full list of variables used in this paper together with the summary statistics is given in Appendix A. There are a lot of missing values in bureau variables in 2012 (a very common problem in practice), and this prevented

[c] https://www.lendingclub.com/.

Table 2. Logistic regression to model default from LC risk sub-grades.[a]

Effect	DF	Parameter estimate	Wald Chi-Square	*P*-value	Odds ratio
LC sub- grade	34		2023.5723	< 0.0001	
Term 60 months	1	0.1749	108.0577	< 0.0001	1.419
SBL	1	0.5853	70.5919	< 0.0001	1.796
Refinance	1	−0.0566	3.3233	0.0683	0.945

Note: Model fit: AIC = 44905.38, Pseudo R-sq = 0.044, $n = 53367$.
[a]The dependent variable Default = 1 (Charged Off or 16 + Days Past Due). The reference (omitted) dummy is "Other Loan Purposes". DF stands for Degrees of Freedom. Category parameter estimates for LC sub-grade are available on request.

previous studies from using the full set of potential independent variables. Thus, Mach *et al.* (2014) use FICO score, income, loan amount, term, home ownership, employment length. We overcome this limitation by categorizing or binning variables with missing values—an approach common in credit scoring and machine-learning. Missing values become one of the categories, and this approach also has an advantage of dealing with outliers and any nonlinear patterns.

The logistic regression model is trained on all loans in 2012 (training sample) with stepwise selection of variables (statistically significant at 5% level). Then tailored models are developed for Refinance and SBL, in order to check if default prediction improves when these segments have separate risk assessment models.

The principles of credit scoring require the models to be tested on the out-of-time sample, in order to evaluate how the model will perform beyond the time period it is trained on, because the main objective of predictive models is to predict into the future. In order to evaluate the predictive accuracy the loan performance needs to be known, therefore our out-of-time test sample consists of loan applications made in 2013, with loan performance observed between 2013 and 2018.

The predictive accuracy is measured by two standard measures used in both business failure prediction and consumer credit scoring: Area under the Curve (AUC) and Accuracy Ratio (AR) (Bishop 2006, Engelmann *et al.* 2003, Thomas 2009). AUC is obtained by plotting percentage of correctly predicted defaults (1-Type 1 error) against percentage of incorrectly predicted non-defaults (Type 2 error) for all values of predicted PD. AUC summarizes the quality of prediction over all possible cut-offs with higher values indicating better prediction: from 0.5 (no separation between classes, a random prediction) to 1 (maximum possible separation). Accuracy Ratio (AR) is related to AUC (Engelmann *et al.* 2003) and takes values from 0 (no separation) to 1 (perfect separation): AR = 2 × AUC-1.

Table 3 presents Accuracy Ratio and Area under the curve (AUC). The performance is assessed separately on Refinance and SBL segments. It makes sense to compare the models to internal LC assessment process, thus LC risk grades (A is the best, G - the worst) are used as a benchmark.

Table 3. Accuracy ratio and area under the curve for LC risk grades, Logistic regression developed and applied to the whole portfolio, "Refinance" and "SBL", training and testing samples.

	Dataset segment					
	2012 (train)			2013 (test)		
	All	Refinance	SBL	All	Refinance	SBL
Model	Accuracy ratio (AR)					
LC risk subgrade	0.304	0.306	0.308	0.33	0.336	0.236
Logistic all	0.4	0.396	0.382	0.362	0.364	0.242
Refinance logistic		0.4			0.364	
SBL logistic			0.424			0.23
SBL logistic with cross-validation			0.376			0.222
	Area under the curve (AUC)					
LC risk subgrade	0.652	0.653	0.654	0.665	0.668	0.618
Logistic all	0.7	0.698	0.691	0.681	0.682	0.621
Refinance logistic		0.7			0.682	
SBL logistic			0.712			0.615
SBL logistic with cross-validation			0.688			0.611

Predictive accuracy reported in Table 3 is relatively modest. There is some improvement for logistic regression as compared to LC risk grades, more pronounced for 2012, which is to be expected because it is the training set. Nevertheless, even in 2013, which is used as a test sample, there is still almost 10% improvement in Accuracy Ratio over LC risk subgrades. There is practically no difference when the logistic model is re-trained for Refinance segment separately, and this is to be expected, since Refinance constitutes almost 80% of all loans.

For SBL segment re-training the separate model notably improves the predictive accuracy in 2012, but the most likely reason is overfitting, since when applying the model to 2013, any gains disappear, and there is practically no difference between LC, Logistic all and SBL logistic. This can be attributed to the instability of predictors. Table 4 shows that there are only six variables selected into the model as compared to

Table 4. Significant (5% level) variables in SBL model, logistic regression.

#	Variable	DF	Chi-sq	P-value
1	Term	1	36.1665	< 0.0001
2	FICO score	1	25.039	< 0.0001
3	Number of accounts opened in past 24 months	4	22.4685	0.0002
4	Annual income	2	14.5684	0.0007
5	Loan amount	3	15.5548	0.0014
6	Number of credit bureau inquiries in past 6 months	4	16.4391	0.0025

Note: Model fit: AIC = 1452.69, Pseudo R-Square 0.104, $n = 1386$. Full details of parameter estimates are available on request, estimates for levels of categorical variables are not included for the sake of compact presentation.

Table 5. Significant (5% level) variables in refinance model, logistic regression.

#	Variable	DF	Chi-sq	*P*-value
1	Term	1	964.927	< 0.0001
2	FICO score	1	695.192	< 0.0001
3	Annual income	8	265.216	< 0.0001
4	Number of accounts opened in past 24 months	8	241.573	< 0.0001
5	Loan amount	8	135.966	< 0.0001
6	Months since most recent inquiry	9	86.425	< 0.0001
7	Number of instalment accounts	4	59.86	< 0.0001
8	Months since most recent bankcard opened	9	57.7826	< 0.0001
9	Purpose	1	40.6216	< 0.0001
10	Number of accounts opened in past 12 months	6	33.6335	< 0.0001
11	Ratio of total current balance to high credit/credit limit for all bankcard accounts	9	29.31	0.0006
12	Total high credit/credit limit for all bankcard accounts	8	26.0958	0.001
13	Months since most recent revolving account opened	9	25.4283	0.0025
14	The number of months since the borrower's last delinquency	9	24.7726	0.0032
15	Debt To Income (DTI)	8	24.3283	0.002
16	Percentage of all bankcard accounts > 75% of limit	8	24.0433	0.0023
17	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit	8	21.9431	0.005
18	The number of credit bureau inquiries in past 6 months	4	16.7973	0.0021
19	Total credit balance excluding mortgage	8	16.0575	0.0416
20	Home ownership	2	10.3922	0.0055
21	Number of open revolving accounts	2	9.3838	0.0092

Note: Model fit: AIC = 33804.61, Pseudo R-Square 0.0678, $n = 41176$. Full details of parameter estimates are available on request, estimates for levels of categorical variables are not included for the sake of compact presentation.

21 significant predictors for Refinance model (Table 5). Following the suggestion of an anonymous reviewer, in order to mitigate the overfitting, we perform 5-fold cross-validation for the SBL model. This reduces the AUC on the training set from 0.712 to a more realistic 0.688, however, the performance on the test set becomes also worse. If there is a weak association between the available predictors and default, it is impossible to develop the accurate model. It is better to assess SBL with the whole portfolio model, yet the result is still significantly lower as compared to Refinancers.

Small number of significant predictors means that even minor changes make the model less predictive on a test sample. Besides, a poor performance of SBL model raises the question about the information that can be used for modeling the risk of a small business. If the personal credit history of business owners does not contain the information with strong and stable associations with default, any model will not be able to capture true underlying drivers of default.

5. Differences Between Refinance and SBLs

In this section, we are going to analyze the information available to Lending Club at the point of application to show that SBLs differ on a number of characteristics

from Refinancers. First, we fit logistic regression to predict if the loan is likely to be an SBL as opposed to Refinance. We use the same predictors, i.e. personal credit history, as in the previous section and stepwise selection as before. The numeric variables that do not have missing values, such as *DTI* or *Loan Amount*, are entered as they are, variables with missing values are split into equally sized categories or bins (containing similar number of accounts) and entered as categorical with dummy variable parametrization. The bins are numbered in the ascending order, with higher values of the original variable corresponding to a higher category number, and "99" - to missing. The first (lowest) category is excluded from estimation as a reference one, i.e. the one that is used to interpret the parameter estimates of the other categories for this variable. For example, *Loan amount* has the reference category that contains the smallest amounts observed in the training sample, therefore, positive signs for statistically significant categories 5,6,7,8 mean that borrowers with larger loan amounts are more likely to be small businesses.

There are 14 variables statistically significant at 5% level (Table 6), and the model's predictive accuracy is better than when predicting default (AUC = 0.776,

Table 6. Variables significant in distinguishing SBL from refinance.

Variable	Category	Estimate	*P*-value	Odds ratio
Intercept		−0.4494	0.4563	
Revolving line utilization rate		−2.085	< 0.0001	0.124
DTI		−0.0535	< 0.0001	0.948
The number of credit bureau inquiries in past 6 months		0.213	< 0.0001	0.735
The number of open credit lines in the borrower's credit file		−0.0445	0.0002	0.956
The total number of credit lines currently in the borrower's credit file		−0.0162	< 0.0001	0.984
# of 30+ days past-due delinquency in the borrower's credit file for the past 2 years		0.1032	0.0194	1.109
Loan amount	1	−0.0564	0.6817	0.945
(Ref: smallest loan amounts)	2	0.0493	0.7169	1.051
	3	0.0203	0.8824	1.02
	4	−0.0477	0.737	0.953
	5	0.3001	0.0243	1.35
	6	0.456	0.0007	1.578
	7	0.5306	< 0.0001	1.7
	8	1.117	< 0.0001	3.056
Percentage of all bankcard accounts > 75% of limit	1	−0.4575	0.0009	0.633
(Ref: lowest number of cards that utilize more than 75% of credit limit)	2	−0.6064	< 0.0001	0.545
	3	−0.6397	0.0004	0.527
	4	−0.7323	< 0.0001	0.481
	5	−0.6967	< 0.0001	0.498
	6	−1.1366	< 0.0001	0.321
	7	−0.3439	0.3562	0.709
	9	−0.7977	< 0.0001	0.45
	99	−0.0569	0.8237	0.945

Table 6. (*Continued*)

Variable	Category	Estimate	*P*-value	Odds ratio
Total bankcard high credit/credit limit	1	−0.1401	0.3305	0.869
(Ref: lowest values)	2	−0.1097	0.4447	0.896
	3	−0.1077	0.4532	0.898
	4	−0.2541	0.0871	0.776
	5	−0.4934	0.0014	0.611
	6	−0.479	0.0018	0.619
	7	−0.4199	0.0044	0.657
	8	−0.4747	0.0013	0.622
	99	−0.2906	0.2633	0.748
Number of accounts opened in the last 12 months	1	−0.0557	0.7009	0.946
(Ref: lowest number)	2	0.1829	0.2132	1.201
	3	−0.3157	0.0941	0.729
	4	−0.171	0.4451	0.843
	6	0.0235	0.9241	1.024
	7	0.6925	0.1306	1.999
	99	0.3846	0.0031	1.469
Number of satisfactory accounts	1	−0.0975	0.5772	0.907
(Ref: lowest number)	2	−0.3061	0.0964	0.736
	3	−0.4417	0.0292	0.643
	4	−0.2632	0.4245	0.769
	99	0.0374	0.842	1.038
Employment length	n/a	−0.468	0.0395	0.626
(Ref: < 1 year)	1 year	0.131	0.3603	1.14
	2 years	0.1815	0.1705	1.199
	3 years	0.3249	0.0171	1.384
	4 years	0.2509	0.07	1.285
	5 years	0.0575	0.68	1.059
	6 years	0.1463	0.3111	1.158
	7 years	0.1419	0.3536	1.152
	8 years	0.1805	0.2757	1.198
	9 years	−0.0895	0.6365	0.914
	10+ years	−0.1886	0.1118	0.828
Total credit balance excluding mortgage	1	−0.1481	0.2613	0.862
(Ref: lowest balance)	2	0.1407	0.2866	1.151
	3	0.0873	0.5288	1.091
	4	0.2131	0.1362	1.238
	5	0.2106	0.1642	1.234
	6	0.3871	0.0096	1.473
	7	0.5757	0.0002	1.778
	8	0.6168	0.0001	1.853
Home ownership	Mortgage	0.232	0.0003	1.261
(Ref: Rent)	Own	0.1528	0.1761	1.165

Note: Predictors with missing values are categorized with higher values corresponding to a larger category number, dependent variable SBL = 1.

AR = 0.552). SBLs have lower Debt to Income (DTI) ratio, lower number of credit accounts, lower credit limit for bank cards and lower utilization rate, since the signs on the corresponding variables are negative. Yet they have higher loan amounts and total credit balances. The positive sign on *The number of credit bureau inquiries* can

be interpreted as a higher need for credit. This variable records the requests from lenders for credit history of credit applicants, and the positive sign indicates that SBL applicants shop around for credit more than Refinancers. It seems that despite having more spare credit capacity in comparison to Refinancers, entrepreneurs, nevertheless, are more credit-constrained. There is no information as to the terms and conditions of the credit lines outside LC, but it is possible to suggest that these conditions may drive business borrowers to seek more credit and/or better terms, or the existing spare capacity is not sufficient.

As for the credit performance, SBLs are likely to have lower number of satisfactory accounts as compared to Refinance. Besides, they are likely to have higher number of delinquencies on their credit file. They are also more likely to have mortgages which is consistent with Jagtiani & Lemieux (2016) who comment on the tendency of SMEs to use mortgages as a source of business funding. In terms of *Employment length* SBLs are likely to have 3–4 years of employment history.

In order to get a more comprehensive picture of the data structure, and as a robustness check we have also summarized all information available at the application point into several uncorrelated dimensions. This has been done by Factor Analysis for Mixed Data (FAMD) by Le *et al.* (2008) that combines PCA for numeric variables and Multiple Correspondence Analysis (MCA) for categorical ones. The results are consistent with those reported above and are available on request.

6. Robustness and Sensitivity Checks

In this section, we conduct the robustness checks for sample selection bias and to ensure that the discrepancy in information value between SBLs and Refinancers is not dependent on the variables chosen and/or classification algorithm used.

The LC loan applications go through preliminary selection as described in Sec. 3, therefore, the PD models developed on the accepted applications may be subject to selection bias. The information published on rejected applicants is much more restricted as compared to funded loans, for sample selection modeling we could only use the following variables: FICO score, DTI, Requested Loan Amount, Employment Length, Loan Purpose.

The selection bias correction consists in estimating the selection equation (usually by binary probit) and then using this equation to correct the parameters in the model of interest. We use the approach for binary response models proposed by Amemiya (1978) that involves estimating the two models simultaneously and this corrects for the endogeneity problem that occurs due to the selection bias. Greene (2006) argues that simultaneous or joint estimation is superior to the two-step approach that mimics the original model proposed for the continuous response by Heckman (1979). In line with the majority of literature on sample selection, we use the binary probit regression instead of logistic. These models are very similar (Greene 2006), as confirmed by our results in Table 7, where we first estimate probit PD model without sample correction. There are no or negligible differences with logistic PD.

Table 7. Predictive accuracy with selection bias correction.

	Dataset segment			
	2012 (train)		2013 (test)	
	Refinance	SBL	Refinance	SBL
Model	Accuracy ratio (AR)			
Refinance logistic	0.4		0.364	
Refinance probit	0.396		0.366	
Refinance probit with selection correction	0.388		0.364	
SBL logistic		0.424		0.23
SBL probit		0.424		0.232
SBL probit with selection correction		0.264		0.136
	Area under the curve (AUC)			
Refinance logistic	0.7		0.682	
SBL logistic		0.712		0.615
Refinance probit	0.698		0.683	
SBL probit		0.712		0.616
Refinance probit with selection correction	0.694		0.682	
SBL probit with selection correction		0.632		0.568

We are still interested in the predictive accuracy, and if there is a discrepancy between Refinance and SBL observed in Sec. 4. Although there is a significant correlation between selection and PD models, and there are changes in PD parameter estimates, the accuracy of PD estimation for Refinance is almost the same, as can be seen from Table 7. It may seem surprising, however, Banasik & Crook (2004) report similar results for a different credit portfolio. Nevertheless, there is a striking change for SBL, as predictive accuracy deteriorates dramatically. Note that this cannot be taken as the evidence that SBLs are disproportionately rejected, this is the consequence of adjusting the parameter estimates of already unstable model, which leads to even worse performance.

In short, we could not achieve the improvement in the predictive accuracy. Since the main objective of our study is to evaluate the predictive value of personal credit history, we did not pursue this analysis further. In subsequent checks, we used the models estimated on accepted applications only, since they gave the best predictive performance.

Further, we estimate the PD models for SBLs and Refinancers using the most powerful machine-learning (ML) techniques. The predictive accuracy is reported in Table 8 for the models with binned predictors (similar to earlier analysis) and feature selection. However, we have also experimented with a different method of dealing with missing values and used all variables that are provided by LC in their original form, i.e. we do not perform any feature selection or transformation. Namely, we have performed MCMC multiple imputation (Rubin 1987, Graham 2012), which have been shown to perform well in the context of credit risk modeling

Table 8. Predictive performance as measured by area under the ROC-curve of different machine-learning algorithms.

SBL	AUC			Refinance	AUC		
Algorithm	Train	Test	Train - Test	Algorithm	Train	Test	Train - Test
Linear SVM with balancing	0.712	0.616	0.096	Linear SVM with balancing	0.702	0.678	0.024
NN	0.716	0.616	0.1	XG Boost Linear	0.692	0.677	0.015
Random Forest	0.872	0.595	0.277	XG Boost Tree	0.726	0.658	0.068
XG Boost Linear	0.689	0.591	0.098	Decision Tree (CHAID) with balancing	0.681	0.646	0.035
XG Boost Tree	0.827	0.587	0.24	NN with balancing	0.735	0.644	0.091
Logistic regression	0.712	0.615	0.097	Logistic regression	0.7	0.682	0.018

(Florez-Lopez 2010). Yet in our analysis MCMC multiple imputation has demonstrated slightly inferior predictive accuracy on the out-of-time test sample, and therefore, is not included here. These results are available on request.

The choice of machine-learning techniques is based on studies that compared different algorithms in credit context (Brown & Mues 2012, Lessmann *et al.* 2015) and include Neural Networks (NN)—Multi-Layer Perceptron, Support Vector Machines (SVM)—Linear and with Radial Basis Function (RBF), Random Forests, Bayesian Networks, Gradient Boosting of Decision Trees (XG Tree) and of Linear Models (XG Boost). We have also experimented with balancing the training sample—a technique common in machine-learning in order to bring the target classes (Default/Non-Default) to the 50:50 ratio. It has improved the performance of several, but not all models. A detailed description of the algorithms is beyond the scope of this paper, but the interested reader can refer to Bishop (2006).

Table 8 reports the predictive accuracy of five algorithms that have shown best predictive performance, with a standard Logistic Regression for comparison. Similar to the analysis in previous sections and in line with machine-learning methodology we test the models on the independent out-of-time sample—the loans originated in 2013. The algorithms in Table 8 are sorted by descending Area under the ROC-curve (AUC). The results demonstrate similar pattern as in Sec. 4. All algorithms perform well on the training sample, some of them achieve a remarkable performance of above 0.8, yet when applied to independent out-of-time test sample, the performance drops. Again similar to Sec. 4, the drop in performance is much more pronounced for SBLs as compared to Refinancers. For the latter, the maximum drop is 0.091 (or 12.38% from the ranking accuracy achieved on the training sample) for Neural Networks. For the former, the smallest drop is 0.096 (or 13.48%) for Linear SVM and the maximum reaching 0.277 (or 31.77%). This confirms our initial observations that predictive value of personal credit history is higher for consumers rather than entrepreneurs.

7. Discussion

Our results show that the credit risk models developed on the personal credit history predict into the future far less accurately for entrepreneurs as compared to consumers. This is important because personal credit history is often used to evaluate the risk of entrepreneurs and small business owners in absence of other information. This is especially common when entrepreneurs apply for retail credit in order to finance their business needs, which is the case with Lending Club (LC) — the P2P lender used in the analysis.

We first estimate the Probability of Default (PD) with the LC internal credit risk ratings, loan duration and dummy variables for small business loans (SBLs) and Refinancers. We find that the dummy for SBL is highly significant and positive, with odds of default for SBLs twice higher as compared to other loans after controlling for LC risk ratings and loan duration. One can say that the risk of SBLs is not

adequately captured by the information contained in risk grades, and the dummy variable proxies for omitted or unobserved predictors of default. The dummy for Refinancers is significant only at 10% level, implying the risk of refinancers is more adequately represented by LC risk grades.

We then proceed with the analysis of detailed information about loans that LC makes available to investors (e.g. FICO scores, annual income, etc.), in order to understand to what extent this information is helpful in predicting default, and therefore, in making a good investment, and whether there are differences in its predictive value/accuracy for SBLs and Refinancers. We approach the information value from the Data Science perspective, and use standard measures for evaluating predictive accuracy. We estimate several logistic regression models, and measure predictive accuracy separately for SBLs and Refinance, and on the training sample (applications funded in 2012) and out-of-time test sample (applications funded in 2013).

The model developed on the whole sample provides a modest level of predictive accuracy which is slightly higher for Refinance than for SBL in 2012 (AUC = 0.698 vs 0.691). The difference becomes more pronounced on the test set, 0.682 vs 0.621. One would expect that the credit risk of entrepreneurs/small business owners is different to that of consumers and this may be the explanation for the observed discrepancy. It is known that small businesses differ from large corporates, and it is suggested that separate credit risk models should be developed for them (Altman & Sabato 2005). We follow the same logic and develop separate models for Refinance and SBL. A separate Refinance model shows a slight improvement if compared to the whole sample model in 2012, and the same performance in 2013. However, a separate SBL model whilst predicting well in-sample, demonstrates a drastic performance out-of-time and out-of-sample. The attempt to correct for overfitting with cross-validation does not improve the out-of-time/out-of-sample prediction. In fact, it is better to apply the whole sample model to SBL because it gives slightly better performance in 2013 – 0.621 as compared to 0.615 of the SBL model. Nevertheless, the difference is marginal, and both results are significantly below the predictive accuracy for Refinance.

We attribute this to low predictive value of personal credit history and low stability of association between credit history and probability of default (PD) of entrepreneurs. Variables significantly associated with PD in one year, become insignificant in the next one, making it difficult to screen the loans.

We show that for Refinance it is possible to achieve a stable screening of reasonable quality, yet it is a much bigger problem for entrepreneurs. The low screening quality of the information can be detrimental to closing the lending gap for small businesses, since inability to assess the risk constraints the credit growth. A higher default rate should not in itself restrict access to credit if the credit risk is accurately estimated and priced accordingly. There are risk-seeking investors with preference for higher returns. However, inability to predict risk is a serious problem.

As the next step in our exploratory analysis, we compare SBLs to Refinance, in order to understand if their risk profiles are indeed different. The latter group is the

largest category of loans in LC portfolio and one can argue these are "typical" customers not only of LC, but many marketplace lenders. These customers are also normally perceived as the highest risks. We find significant differences in the information/risk profiles of SBL and Refinance, with SBLs being less leveraged but having a greater demand for credit, with worse repayment history and containing a higher proportion of mortgage holders.

As robustness checks we correct for the selection bias. We also employ a wider range of most advanced machine-learning algorithms, to make sure that the results do not depend on a specific model. Whilst all algorithms perform well on the training sample, the performance on the independent out-of-time test sample is considerably worse, and the drop in performance is much more pronounced for SBLs as compared to Refinancers, confirming the results obtained with the logistic regression.

The results support the view that improvements in predictive accuracy are more likely to come from new types of information rather than from algorithmic side, which in turn reinforces the importance of our investigation of the value of the entrepreneurs' previous personal credit history for predicting their performance.

Despite the insights, the study is not without limitations. Our study uses one year of funding, future research can investigate earlier years or concentrate on loan term of 36 months to increase the time span and provide further insights into credit behavior of entrepreneurs applying to LC consumer platform. Another limitation is investigation of one marketplace lender. Whilst this lender is one of the largest in the world and is a good representative of the industry, in general, investigation of other platforms is necessary to confirm the generalizability of the results. Nevertheless, many new and traditional lenders use personal credit history as inputs in their credit risk models, and our results should inform them of the differential predictive value of this type of information for entrepreneurs as compared to consumers.

8. Conclusions

This paper presents an exploratory analysis of one year of loans granted by Lending Club (LC), with the aim to understand the information value of personal credit history for risk evaluation of small business loans (SBL). We compare SBL to consumers that seek to refinance/consolidate their existing debt (Refinance), for whom previous credit history is a logical type of information, traditionally used in credit risk assessment. We find there are significant differences in the profiles of SBL and Refinance borrowers, with the most notable ones being that SBLs are less leveraged but have a greater demand for credit. Their credit performance is likely to be worse, and they are more likely to have mortgages which is consistent with previous findings that SMEs use mortgages as a source of business funding (Jagtiani & Lemieux 2016).

When it comes to predicting Probability of Default (PD), we find that the personal credit history does have some predictive power (better than a random model), but this power is modest and does not allow for the effective discrimination between Defaults and Non-Defaults for SBL in comparison to Refinance. Whilst it is possible

to achieve higher predictive accuracy on the training sample (year of 2012), any gains disappear when the model is applied out-of-time and out-of-sample (year 2013). There are only six statistically significant variables in the PD model for SBLs as compared to 21 variables in Refinance PD model, and it is not surprising that the SBL model cannot achieve high predictive accuracy when applied to new samples.

This suggests that information value of personal credit history is lower for business borrowers in comparison to consumers, and this has important implications in terms of restricting their access to retail credit products. We advocate the need for additional information in order to improve the quality of screening for entrepreneurs.

These exploratory findings raise several questions for further investigation that should be of interest to researchers in the areas of credit risk and entrepreneurial finance. One line of inquiry concerns further investigation of different types of information that would improve the quality of credit screening of entrepreneurs. How can entrepreneurs signal their quality? And to what extent different signaling mechanisms can be connected to the subsequent performance of the loan? Answers to these questions will help to solve the problem of small business lending gap, in particular for start-ups.

Another direction of future research follows from the indications that entrepreneurs applying for marketplace lending are credit constrained. This prompts an investigation of where the marketplace lending comes in the pecking order of funding sources; and what types of entrepreneurs turn to these novel sources of external debt.

Appendix A

Table A.1. The list variables used with description.

Short name	Description
Acc_Open_Past_24Mths	Number of trades opened in past 24 months
Annual Income	The self-reported annual income provided by the borrower during registration
Avg_Cur_Bal	Average current balance of all accounts
Bc_Open_To_Buy	Total open to buy on revolving bankcards
Bc_Util	Ratio of total current balance to high credit/credit limit for all bankcard accounts
Chargeoff_Within_12_Mths	Number of charge-offs within 12 months
Collections_12_Mths_Ex_Med	Number of collections in 12 months excluding medical collections
Credit history	The number of months from the borrower's earliest reported credit line
Default	Repayment performance: 1 if Loan Status is Charged Off or Late; 0 otherwise
Delinq_2Yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
DTI	A ratio of total monthly debt payments on total debt obligations divided by self-reported monthly income

Table A.1. (*Continued*)

Short name	Description
Employment length	Employment length in years between 0 (less than one year) and 10 (ten or more years)
FICO	The upper boundary range the borrower's FICO at loan origination belongs to
Home ownership	The home ownership status: RENT, OWN, MORTGAGE, OTHER
Inq_6Mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
LC Grade	LC assigned loan grade from A (low risk) to F (high risk)
LC sub-grade	LC assigned loan sub-grade from A1 (low risk) to F5 (high risk)
Loan amount	The listed amount of the loan applied for by the borrower
Mo_Sin_Old_Il_Acct	Months since oldest bank installment account opened
Mo_Sin_Old_Rev_Tl_Op	Months since oldest revolving account opened
Mo_Sin_Rcnt_Rev_Tl_Op	Months since most recent revolving account opened
Mo_Sin_Rcnt_Tl	Months since most recent account opened
Mort_Acc	Number of mortgage accounts.
Mths_Since_Last_Delinq	The number of months since the borrower's last delinquency
Mths_Since_Last_Major_Derog	Months since most recent 90-day or worse rating
Mths_Since_Last_Record	The number of months since the last public record
Mths_Since_Recent_Bc	Months since most recent bankcard account opened
Mths_Since_Recent_Bc_Dlq	Months since most recent bankcard delinquency
Mths_Since_Recent_Inq	Months since most recent inquiry
Mths_Since_Recent_Revol_Delinq	Months since most recent revolving delinquency
Num_Accts_Ever_120_Pd	Number of accounts ever 120 or more days past due
Num_Actv_Bc_Tl	Number of currently active bankcard accounts
Num_Actv_Rev_Tl	Number of currently active revolving trades
Num_Bc_Sats	Number of satisfactory bankcard accounts
Num_Bc_Tl	Number of bankcard accounts
Num_Il_Tl	Number of installment accounts
Num_Op_Rev_Tl	Number of open revolving accounts
Num_Rev_Accts	Number of revolving accounts
Num_Rev_Tl_Bal_Gt_0	Number of revolving trades with balance > 0
Num_Sats	Number of satisfactory accounts
Num_Tl_120Dpd_2M	Number of accounts currently 120 days past due (updated in past 2 months)
Num_Tl_30Dpd	Number of accounts currently 30 days past due (updated in past 2 months)
Num_Tl_90G_Dpd_24M	Number of accounts 90 or more days past due in last 24 months
Num_Tl_Op_Past_12M	Number of accounts opened in past 12 months
Open_Acc	The number of open credit lines in the borrower's credit file
Pct_Tl_Nvr_Dlq	Percent of trades never delinquent
Percent_Bc_Gt_75	Percentage of all bankcard accounts $> 75\%$ of limit
Pub_Rec	Number of derogatory public records
Pub_Rec_Bankruptcies	Number of public record bankruptcies
Purpose	A category provided by the borrower for the loan request
Refinance	Refinancing consumers: 1 if Purpose ="credit_card" or "debt_consolidation", 0 otherwise
Revol_Bal	Total credit revolving balance
Revol_Util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit
SBL	Small Business Loan: 1 if Purpose ="small_business", 0 otherwise
Tax_Liens	Number of tax liens

Table A.1. (*Continued*)

Short name	Description
Term	The number of payments: 0 for 36 months; 1 for 60 months
Tot_Coll_Amt	Total collection amounts ever owed
Tot_Cur_Bal	Total current balance of all accounts
Tot_Hi_Cred_Lim	Total high credit/credit limit
Total_Acc	The total number of credit lines in the borrower's credit file
Total_Bal_Ex_Mort	Total credit balance excluding mortgage
Total_Bc_Limit	Total bankcard high credit/credit limit
Total_Il_High_Credit_Limit	Total installment high credit/credit limit
Total_Rev_Hi_Lim	Total revolving high credit/credit limit

Table A.2. Summary statistics for numeric variables.

Short name	Mean	Std dev	Minimum	Maximum	% Miss
Acc_Open_Past_24Mths	3.91	2.67	0	40	14.04%
Annual_Inc	69720.23	58654.51	4800	7141778	0.00%
Avg_Cur_Bal	13235.01	16821.35	0	800008	51.98%
Bc_Open_To_Buy	8251.61	13482.97	0	497445	15.03%
Bc_Util	66.28	27.36	0	187.9	15.08%
Chargeoff_Within_12_Mths	0.0015	0.0430	0	3	0.00%
Collections_12_Mths_Ex_Med	0.0004	0.0194	0	1	0.00%
Credit History	175.75	81.44	36	650	0.00%
Delinq_2Yrs	0.2017	0.6361	0	18	0.00%
DTI	16.66	7.59	0	34.99	0.00%
Fico_Range_High	705.39	32.34	664	850	0.00%
Inq_6Mths	0.83	1.01	0	8	0.00%
Loan amount	13461.71	8086.93	1000	35000	0.00%
Mo_Sin_Old_Il_Acct	122.57	50.93	1	649	54.19%
Mo_Sin_Old_Rev_Tl_Op	172.05	84.93	9	658	51.98%
Mo_Sin_Rcnt_Rev_Tl_Op	14.05	16.29	0	264	51.98%
Mo_Sin_Rcnt_Tl	8.73	9.63	0	174	51.98%
Mort_Acc	1.61	2.17	0	24	14.04%
Mths_Since_Last_Delinq	36.69	21.52	0	152	58.84%
Mths_Since_Last_Major_Derog	42.41	20.85	0	152	89.91%
Mths_Since_Last_Record	91.66	23.20	1	119	97.24%
Mths_Since_Recent_Bc	24.79	28.94	0	538	14.89%
Mths_Since_Recent_Bc_Dlq	41.39	21.16	0	152	87.43%
Mths_Since_Recent_Inq	6.69	5.99	0	24	24.90%
Mths_Since_Recent_Revol_Delinq	38.05	21.22	0	152	73.92%
Num_Accts_Ever_120_Pd	0.32	0.92	0	26	51.98%
Num_Actv_Bc_Tl	3.66	2.08	0	30	51.98%
Num_Actv_Rev_Tl	5.53	2.93	0	34	51.98%
Num_Bc_Sats	4.60	2.46	0	32	30.08%
Num_Bc_Tl	9.12	4.88	0	44	51.98%
Num_Il_Tl	7.54	6.34	0	57	51.98%
Num_Op_Rev_Tl	7.91	3.80	0	39	51.98%
Num_Rev_Accts	14.77	7.34	0	57	51.98%
Num_Rev_Tl_Bal_Gt_0	5.55	2.94	0	34	51.98%
Num_Sats	10.94	4.53	1	46	30.08%
Num_Tl_120Dpd_2M	0.0001	0.0108	0	1	51.98%

Table A.2. (*Continued*)

Short name	Mean	Std dev	Minimum	Maximum	% Miss
Num_Tl_30Dpd	0.0004	0.0216	0	2	51.98%
Num_Tl_90G_Dpd_24M	0.0627	0.3490	0	12	51.98%
Num_Tl_Op_Past_12M	1.84	1.59	0	25	51.98%
Open_Acc	10.62	4.47	1	49	0.00%
Pct_Tl_Nvr_Dlq	94.93	7.64	15	100	51.98%
Percent_Bc_Gt_75	53.34	34.83	0	100	15.03%
Pub_Rec	0.0289	0.1782	0	5	0.00%
Pub_Rec_Bankruptcies	0.0230	0.1530	0	5	0.00%
Revol_Bal	15100.76	14776.98	0	975800	0.00%
Revol_Util	0.5790	0.2423	0	1.044	0.09%
Tax_Liens	0.0008	0.0397	0	5	0.00%
Tot_Coll_Amt	47.82	653.36	0	55009	51.98%
Tot_Cur_Bal	129454.83	154569.27	0	8000078	51.98%
Tot_Hi_Cred_Lim	156387.77	169017.98	500	8592561	51.98%
Total_Acc	23.61	10.93	3	99	0.00%
Total_Bal_Ex_Mort	40148.59	36484.87	0	994496	14.04%
Total_Bc_Limit	19609.34	18471.92	0	522210	14.04%
Total_Il_High_Credit_Limit	33104.30	35893.60	0	902504	51.98%
Total_Rev_Hi_Lim	28392.23	25001.40	0	988000	51.98%

References

U. Ahmed, T. Beck, C. McDaniel & S. Schropp (2016) Filling the gap: How technology enables access to finance for small- and medium-sized enterprises, *MIT Press Journals: Innovations* **10** (3–4), 35–48.

L. Allen, D. G. DeLong & A. Saunders (2004) Issues in credit risk modeling of retail market, *Journal of Banking and Finance* **28** (4), 727–752.

E. Altman & G. Sabato (2005) Effects of the new Basel capital accord on bank capital requirements for SMEs, *Journal of Financial Services Research* **28** (1), 5–42.

E. Altman & G. Sabato (2007) Modeling credit risk for SMEs: Evidence from the US market, *ABACUS* **43** (3), 332–357.

E. I. Altman & A. Saunders (1997) Credit risk measurement: Developments over the last 20 years, *Journal of Banking and Finance* **21** (11), 1721–1742.

E. Altman, G. Sabato & N. Wilson (2010) The value of non-financial information in small and medium-sized enterprise risk management, *Journal of Credit Risk* **6**, 1–33.

E. Altman, M. Drozdowska, E. Laitinen & A. Suvas (2017) Financial and non-financial variables as long-horizon predictors of bankruptcy, *Journal of Credit Risk* **12** (4), 49–78.

E. Altman, M. Esentato & G. Sabato (2018) Assessing the creditworthiness of Italian SMEs and mini-bond issuers, *Global Finance Journal*, doi. org/10.1016/j.gfj.2018.09.003.

E. I. Altman (1968) Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, *Journal of Finance* **23** (4), 589–609.

T. Amemiya (1978) The estimation of a simultaneous equation generalized probit model, *Econometrica* **46**, 1193–1205.

R. Anderson (2007) *The Credit Scoring Toolkit.* Oxford: Oxford University Press.

J. Banasik & J. Crook (2004) Does reject inference really improve the performance of application scoring models? *Journal of Banking and Finance* **28**, 857–874.

A. N. Berger & W. S. Frame (2007) Small business credit scoring and credit availability, *Journal of Small Business Management* **45** (1), 5–22.

A. N. Berger, W. S. Frame & N. Miller (2005) Credit scoring and the availability, price and risk of small business credit, *Journal of Money, Credit and Banking* **37**, 191–222.

C. Bishop (2006) *Pattern Recognition and Machine Learning*. New York: Springer.

I. Brown & C. Mues (2012) An experimental comparison of classification algorithms for imbalanced credit scoring data sets, *Expert Systems with Applications* **39** (3), 3446–3453.

G. Bruton, S. Khavul, D. Siegel & M. Wright (2015) New financial alternatives in seeding entrepreneurship: Microfinance, crowdfunding, and peer-to-peer innovations, *Entrepreneurship Theory and Practice* **39** (1), 9–26.

S. Chava & N. Paradkar (2018) Winners and losers of marketplace lending: Evidence from borrower credit dynamics, *Presentation at the 18th Annual Bank Research Conference*, FDIC https://www.fdic.gov/analysis/cfr/bank-research-conference/annual-18th/24-paradkar.pdf.

F. Ciampi & N. Gordini (2013) Small enterprise default prediction modeling through artificial neural networks: An empirical analysis of Italian small enterprises, *Journal of Small Business Management* **51** (1), 23–45.

R. Cole (2018) How did bank lending to small business in the United States fare after the financial crisis? *U.S. Small Business Administration Research Study* 439.

M. Dietsch & J. Petey (2004) Should SME exposures be treated as retail or as corporate Exposures? A comparative analysis of default probabilities and asset correlation in French and German SMEs, *Journal of Banking and Finance* **28** (5).

J. Duarte, S. Siegel & L. Young (2012) Trust and credit: The role of appearance in peer-to-peer lending, *Review of Financial Studies* **25** (8), 2455–2484.

Dun & Bradstreet (2021) What's the difference between personal and business credit?, https://www.dnb.co.uk/resources/personal-vs-business-credit.html.

R. Emekter, Y. Tu, B. Jirasakuldech & M. Lu (2015) Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending, *Applied Economics* **47** (1), 54–70.

B. Engelmann, E. Hayden & D. Tasche (2003) Measuring the discriminative power of rating systems, Discussion Paper, Series 2: Banking and Financial Supervision, Deutsche Bundesbank, https://www.econstor.eu/escollectionhome/10419/24.

R. Florez-Lopez (2010) Effects of missing data in credit risk scoring: A comparative analysis of methods to achieve robustness in the absence of sufficient data, *Journal of the Operational Research Society* **61**, 486–501.

S. Freedman & G. Jin (2014) The information value of online social networks: Lessons from peer-to-peer lending, NBER Working Paper 19820, http://www.nber.org/papers/w19820.

L. Gonzalez & Y. Loureiro (2014) When can a photo increase credit? The impact of lender and borrower profiles on online peer-to-peer loans, *Journal of Behavioral and Experimental Finance* **2**, 44–58.

J. W. Graham (2012) *Missing Data: Analysis and Design*. New York: Springer.

W. Greene (2006) A general approach to incorporating selectivity in a model, Working Papers 06-10, New York University, Leonard N. Stern School of Business, Department of Economics.

J. Heckman (1979) Sample selection bias as a specification error, *Econometrica* **47** (1), 153–161, doi: 10.2307/1912352.

A. Hertzberg, A. Liberman & D. Paravisini (2018) Screening on loan terms: Evidence from maturity choice in consumer credit, *Review of Financial Studies* **31** (9), 3532–3567.

R. Iyer, A. Khwaja, E. Luttmer & K. Shue (2016) Screening peers softly: Inferring the quality of small borrowers, *Management Science* **62** (6), 1554–1577.

D. M. Jaffee & T. Russell (1976) Imperfect information, uncertainty, and credit rationing, *Quarterly Journal of Economics* **90**, 651–666.

J. Jagtiani & C. Lemieux (2016) Small business lending after the financial crisis: A new competitive landscape for community banks, *Federal Reserve Bank of Chicago Economic Perspectives* **40** (3), https://www.chicagofed.org/publications/economic-perspectives/2016/3-jagtianilemieux.

L. Larrimore, L. Jiang & J. Larrimore (2011) Peer to peer lending: The relationship between language features, trustworthiness, and persuasion success, *Journal of Applied Communication Research* **39**, 19–37.

S. Le, J. Josse & F. Husson (2008) FactoMineR: An R package for multivariate analysis, *Journal of Statistical Software* **25** (1), 1–18, http://www.jstatsoft.org/v25/i01/.

Lending Club (2010) Lending Club Prospectus, https://www.lendingclub.com/fileDownload.action?file=Clean_As_Filed_20101015.pdf&type=docs.

S. Lessmann, B. Baesens, H. V. Seow & L. C. Thomas (2015) Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research, *European Journal of Operational Research* **247** (1), 124–136.

M. Lin, N. Prabhala & S. Viswanathan (2013) Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to peer lending, *Management Science* **59**, 17–35.

D. Liu, D. Brass, Y. Lu & D. Chen (2015) Friendships in online peer-to-peer lending-pipes, prisms, and relational herding, *MIS Quarterly: Management Information Systems* **39** (3), 729–742.

T. L. Mach, C. M. Carter & C. R. Slattery (2014) Peer-to-peer lending to small businesses, Finance and Economics Discussion Series, Divisions of Research & Statistics and Monetary Affairs, Federal Reserve Board, Washington, D.C.

K. Mills & B. McCarthy (2014) The state of small business lending: Credit access during the recovery and how technology may change the game, Harvard Business School Working Paper 15-004.

K. Mills & B. McCarthy (2016) The state of small business lending: Innovation and technology and the implications for regulation, Harvard Business School Working Paper 17-042.

D. Rubin (1987) *Multiple Imputation for Non-Response in Surveys*. New York: Wiley.

P. Rudegeair (2020) Lack of bank credit hits small business, *Wall Street Journal*, https://www-proquest-com.ezproxy.is.ed.ac.uk/newspapers/lack-bank-credit-hits-small-business/docview/2471353688/se-2?accountid=10673.

C. Serrano-Cinca, B. Gutiérrez-Nieto & L. López-Palacios (2015) Determinants of default in P2P lending, *PloS One* **10** (10).

N. Siddiqi (2006) *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. New Jersey: Wiley.

J. E. Stiglitz & A. Weiss (1981) Credit rationing in markets with imperfect information, *American Economic Review* **71**, 393–410.

L. C. Thomas (2009) *Consumer Credit Models*. Oxford: Oxford University Press.

C. F. Vallini, F. Ciampi, N. Gordini & M. Benvenuti (2009) Are credit scoring models able to predict small enterprise default? Statistical evidence from Italian small enterprises, *International Journal of Business & Economics* **8** (1), 3–18.