NOVA IMS Information Management School

MEGI

Mestrado em Estatística e Gestão de Informação

Master Program in Statistics and Information Management

EVALUATING CREDIT DEFAULT RISK IN P2P LENDING

A Market Maturity Perspective

Rita Serras Celorico Da Silva Fialho

Dissertation presented as partial requirement for obtaining the Master's degree in Statistics and Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

EVALUATING CREDIT DEFAULT RISK IN P2P LENDING

by

Rita Serras Celorico Da Silva Fialho

Dissertation presented as partial requirement for obtaining the Master's degree in Statistics and Information Management

Advisor: Prof. Dr. Jorge Miguel Ventura Bravo, PhD

November 2021

ABSTRACT

The Peer-to-Peer (P2P) lending industry has grown significantly in recent years, owing its traction to a growing interest in the market from both individual borrowers and lenders alike. Such platforms have greatly benefited from the advent of digital transformation and expanding internet footprint across people of all ages and backgrounds. These platforms leverage novel ways of interacting to facilitate an alternative means of financing which provides more opportunities to borrowers who may, otherwise not have access to debt through conventional mechanisms, and more investment opportunities to lenders who may be seeking to diversify their portfolios.

This study leverages a dataset made available by the Lending Club P2P lending platform, containing more than 2 million loan entries, amassed over more than 10 years. Being a pioneer in the industry, Lending Club's track record and data provide a good window into the inner workings of such platforms and their ability to assess the creditworthiness of borrowers and loan applications.

We set out to build upon prior work done in this space by understanding and analyzing this dataset to identify the determinants of default of loans issued through P2P lending platforms. Our analysis is also employed to create a predictive model which is then tested against our dataset. This approach builds upon previous studies by outlining an end-to-end process to analyze and assess a platform's ability to adequately predict credit default risk.

We have found that, in alignment with prior work, such platforms are indeed able to adequately assess credit default risk, in the way that grades are assigned to individual loans. The logistic regression model which we have built has also yielded good results in predicting defaulted loans, while exhibiting mediocre performance in classifying fully repaid loans as likely cases of default.

KEYWORDS

Credit Risk; Credit Default; Default Risk; P2P Lending; Lending Club; Logistic Regression; Data Analysis; Determinants of Default

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ANNs	Artificial Neural Networks
AUC	Area Under the ROC Curve
BME	Bayesian Model Ensembles
DA	Discriminate Analysis
DT	Decision Trees
DTI	Debt to Income
FPR	False Positive Rate
GA	Genetic and Evolutionary Algorithms
LR	Logistic Regression
MLM	Maximum Likelihood Method
NN	Neural Networks
P2P	Peer-to-peer
OR	Odds Ratio
RF	Random Forest
ROC	Receiver Operating Characteristic curve
SEC	Securities and Exchange Commission
SVM	Support Vector Machine
TPR	True Positive Rate

1 INTRODUCTION

In modern history, lending is considered to be one of the key activities carried out by financial institutions. It is through the instrument of credit that various economic activities see their spread and diversification, from production to consumption. It is through credit that businesses and individuals fuel their needs to grow and improve.

In the activity of "lending" money, two parties enter a transaction in which a lender "lends" money to a borrower, with a promise of gradual repayment. Regardless of the terms of repayment, there is always a risk associated with the borrower not being able to pay back the loan. This risk is classified as "credit risk" and is, hence, a measure of the likelihood that a borrower may default on an issued loan, thus failing to meet the terms defined in the contract entered by both parties.

Given the nature and uncertainty of the credit market and considering the risks inherent to the money-lending activity, financial institutions have long sought to both develop and improve the mechanisms through which they vet and check the quality of borrowers to whom they lend money, as a means to reducing credit risk and, ultimately, making back their initial investment.

The financial crisis of 2007-2010 had a widespread and global impact on the entire financial industry and vastly affected other sectors of the economy, leading to a reduction in the possibility of recourse to traditional credit by consumers and small businesses and distrust and dissatisfaction with commercial banks and other established financial institutions.

The advent of the internet and the increase in popularity of web services and new business models based on leveraging the benefits of interconnectivity and fast, ubiquitous access to the internet, coupled with the aforementioned distrust and discontent with traditional financial institutions and intermediaries, further accelerated the rise of alternative finance in the context of a globalized world economy.

Alternative financial services compose an array of financial services offered by providers that may operate outside of the umbrella of regulated institutions. Despite many of the products and services provided by them not being "alternative", but rather the same as, or similar to those of, traditional providers, institutions in this segment are often characterized by having more inherent risk to their operations and by executing traditional functions/services via alternative means of interaction (e.g. leveraging the Internet to connect financial peers in a transaction) (Bradley et al., 2009).

Peer-to-peer (P2P) lending is the practice in which individuals lend money to other individuals through, generally, an online platform that directly connects lenders with borrowers without the involvement of a traditional financial intermediary. Where in conventional lending a financial institution, such as a bank, would lend money to an individual, frequently taking some type of collateral to secure it (such as a house or an automobile), this new form of lending complements the market by opening lending as an investment vehicle directly available to individuals, and expanding the opportunities provided to a wider range of borrowers with, potentially less restrictions or constraints (such as age, creditworthiness and employability).

P2P lending platforms offer simple registration processes for both borrowers and lenders alike, allowing them to submit details of their credit needs and funding interest, respectively. After

conducting the necessary due diligence, they connect borrowers and lenders to fill loan requests and provide funding across a range of risk levels (based on the credit rating provided for each loan).

The first surfacing of practical, generally accepted and functional platforms traces back to Zopa, a British P2P lending company, founded in 2005 in the UK. Soon after it, other companies followed the same model, with many non-viable candidates and alternatives coming to be. Other successful platforms were also started in the US, with Prosper (2006) and Lending Club (2007) being the most notable mentions. They are currently the two largest P2P companies in the US, respectively holding figures of US$38 billion in issued loans and over 2.5 million customers for Lending Club and US$13 billion in issued loans and over 820 thousand customers for Prosper (at the time of writing).

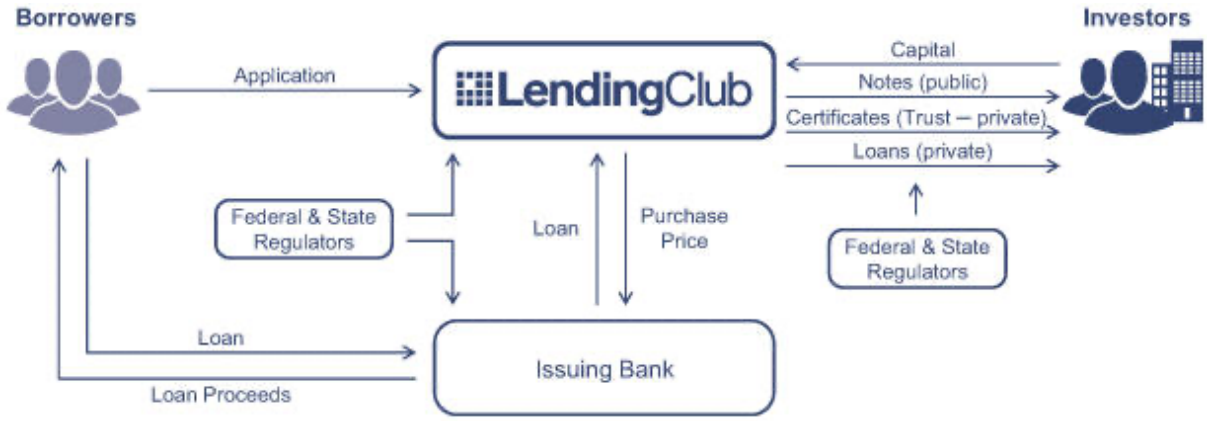
Despite the respectable figures presented above, and the growth trend evidenced in the sector (which is expected to grow to a US$897.85 billion market by 2024), the P2P lending segment still holds a considerably small percentage of all consumer credit issued, when compared to the global figures of US$3.9 trillion. (Transparency Market Research, 2016)

P2P lending platforms do, nonetheless, merit the attention they receive. They continue the trend of disintermediation that online business models have established in modern-day Internet and, thus, provide direct access to unsecured consumer credit as an alternative investment vehicle for individual lenders and do so at a much reduced "intermediation" cost due to deviating from the standard practice of traditional credit organizations. They also democratize access to consumer credit, by employing the aforementioned model, and bring it to the reach of the otherwise non-creditworthy, by applying the necessary safeguards and adequately pricing risk. (Herzenstein, Andrews, Dholakia, & Lyandres, 2008)

1.1 LENDING CLUB

To allow for better comprehension of the data which will be used in later chapters, it is important that we first develop an understanding of the subject under study (Lending Club), by reviewing how it operates and what externalities condition its operation.

Figure 1 – Lending Club's Business Model



Source: Lending Club Corp Form 10-K for Fiscal Year Ended December 31, 2014. Securities and Exchange Commission.

Figure 1 illustrates Lending Club's business model which can be observed in the following steps. (Securities and Exchange Commission, 2015)

First, the borrower creates an account on the platform and is required to provide personal information such as individual details (address, contact, salary information, marital status, etc.), credit history and also the details of the loan they expect to receive (purpose, amount, maturity and other details).

Secondly, Lending Club performs a background check by verifying the information provided by the loan applicant. Taking these details and coupling them with the individual's FICO credit score[1], Lending Club then proceeds to use a score-based model to determine the loan-borrower grade (which can range from A to G, representing lowest and highest risk, respectively) from which it derives a fixed interest-rate (Lending Club, 2019c). By fixing the interest rate in a centralized manner, Lending Club can factor into it the credit risk associated with the borrower, thus transparently protecting the lender and avoiding lender competition for borrowers, which could result in interest rates being driven down (inherently increasing risk).

In the third step of the process, if the borrower meets certain criteria, Lending Club will present them with various loan options (different maturity, interest rate, loan amount, and other options). The borrower will then choose if he wants to proceed with any of the suggested options.

The fourth step is initiated once the offered conditions are accepted by the borrower. Lending Club will list the loan on the platform, for investors to invest in, and once the loan is fully invested (i.e. the loan amount has been filled with capital committed by investors), the Lending Club partner bank will issue the loan.

Lenders can then choose to fund a given loan with the information provided on the platform. The online resources provided by Lending Club facilitate a list to the end-user detailing the borrower's grade, loan amount, amount funded, and time until which the loan is accepting funding. This process mimics the approach of "crowdfunding" as it enables multiple lenders to fund the same loan, while giving them full control of the selection and allowing them to diversify their lending "portfolio" (the types of loans to which they choose to lend).

Finally, in the fifth step of the process, Lending Club purchases the loan from the bank holding the obligation of the loan contract (Nowak, Ross, & Yencha, 2018).

A powerful factor for the fast growth of the platform (and the industry in general) is how its alternative processes and low execution costs make it an attractive alternative for borrowers and lenders alike. Lenders can benefit from a novel investment vehicle that gives them some transparency over the risk of the investment to which they are allocating capital while benefiting from the low operating costs, fees, and levies that may (and would) derive from such activities in the regular context of lending money. Borrowers, on the other hand, benefit from more attractive interest rates provided by the already mentioned streamlining of processes, vetting, and conveyance of information facilitated by the platform, while making access more widespread to a larger community of people which could,

[1] "A FICO score is a credit score created by the Fair Isaac Corporation (FICO). Lenders use borrowers' FICO scores along with other details on borrowers' credit reports to assess credit risk and determine whether to extend credit" - https://www.investopedia.com/terms/f/ficoscore.asp

otherwise, be barred from obtaining credit via conventional bank-borrowing (although interest rates provided may be substantially higher).

Despite the many benefits of P2P lending, there are fewer positive notes which we must identify and detail. One of the most commonly noted is the information asymmetry that exists between borrowers and lenders, the latter are at a disadvantage when compared to the former, given that they are likely to lack the data and tools to vet the borrower (which in traditional lending, a financial institution would generally have). Although this vetting of borrowers is done by Lending Club itself, the risk inherent to this process must be noted when lenders and borrowers are put side-by-side. The asymmetry exists if the platform's processes can't be audited or regulated, and it is up to the lender to decide on how this potential of asymmetrical information may influence his choice of loan/borrower to fund.

Given that the success of a money lender (whether institutional or otherwise) is directly tied with its ability to measure variables such as borrower creditworthiness, the likelihood of repayment of a loan and appropriate interest rates that offset borrower risk, we set out to understand the ability of P2P lending platforms to measure these factors, and correctly predict (or estimate) the risk of credit default. This is done through an analysis of a loan dataset from the P2P lending platform, Lending Club, by using the R Studio software to support the computations and data manipulation required of the work.

Our analysis complements previous work by assessing a more voluminous dataset which has been amassed over the course of more than 10 years. This provides a window into understanding the ability of such platforms to mature their credit risk estimation models over time, improving the quality of decision-making in approving/rejecting loan applications. We further expand the scope of existing work to consolidate the end-to-end process of identifying relevant predictors of credit default, applying them to build a prediction model and testing this model against our extensive dataset.

To achieve this, we have conducted an initial cleaning of the dataset under study to prune it for missing and incorrect data, outliers and irrelevant variables. We built upon this by carrying out an exploratory analysis to measure correlation between variables and validate our approach. After establishing the set of variables which are likely to be good determinants of default, we proceeded to build a predictive model, testing it against a sample of our dataset. Altogether this approach allowed us to understand (the dataset), identify (relevant variables), construct a model and validate it (by testing against available, real data).

Our findings are in line with those of previous work (Carmichael, 2014; Emekter et al., 2015; Polena & Regner, 2018; Serrano-Cinca et al., 2015). Our analysis arrives at similar sets of variables which constitute good determinants of default, additionally suggesting that of all information provided to lenders by the P2P lending platform, the platform-assigned "grades" are the best predictor of this. This result seems to indicate that platforms continually strive to improve their ability to recognize good funding opportunities, translating this into a grade which can provide direction to lenders.

Our prediction model yielded moderate results, given that although it was able to identify defaulted loans with a high level of fidelity, it also incorrectly identified a large number of repaid loans as having a high likelihood of default.

In the work that follows, we will be analyzing a dataset of the online P2P lending platform Lending Club. The platform makes data publicly available to registered users and has been in operation since

2007. Additionally, it was the first platform of its kind to become officially regulated by the Securities and Exchange Commission (SEC) in the United States, conferring it legitimacy without rivals in the industry. This makes Lending Club a good case study for understanding the variables at play that affect a user's likelihood of defaulting on a loan and hence allows us to study the determinants of credit risk in the context of the organization's operating environment.

The rest of this dissertation is organized as follows. **Section 2** (Literature Review) analyses previous work on the subjects that are of relevance to our study, such as the key topics of Credit Risk, P2P Lending and Determinants of Default. **Section 3** (Methodology) details the approach taken to analyze and utilize the provided Lending Club dataset. **Section 4** (Dataset Analysis) reviews existing data points and addresses potential limitations and biases to be later raised in the collected empirical results. **Section 5** (Results) presents the applied processes and methodologies and the outputs that they yielded after being applied to the dataset in question. **Section 6** (Conclusion) takes the outputs of previous sections to arrive at conclusions over the question at hand – "What are the determinants of credit risk in P2P lending?". Additionally, this section lays the foundation for continuing work to be carried out on the topic, detailing potential points of improvement and limitations that were observed during the development of this dissertation.

2 LITERATURE REVIEW

2.1 ONLINE P2P LENDING HISTORY AND CHARACTERIZATION

The rise in popularity of P2P lending platforms has become a self-evident fact from the extensive focus that academic literature has placed both on it and on understanding its multiple facets (Zhao et al., 2017). Hulme & Wright (2006) position said platforms as being the result of evolutionary trends in Social Lending across most of the last 3 centuries and establish their horizontal structure (as opposed to hierarchical structures between borrowers and lenders in traditional credit markets) as a key aspect responsible for their growth. The surfacing of a more developed Internet, one that more directly connects its users employing social networking platforms, and the advent of "Web 2.0", is also established as a growth-driving factor by other work (Emekter, Tu, Jirasakuldech, & Lu, 2015; Tapscott & Williams, 2007). Havrylchyk et al. (2016) further assess contributing factors and acknowledge that aside from the expansion of Internet-based services, an overall deterioration in customer experience for the traditional consumer finance market (either resulting from poor ease of access or inept supply) is paramount in understanding the growth trend.

Understanding the online P2P lending market requires comparing modern online lending platforms with their alternative counterparts in traditional consumer finance, by individually assessing both functions they are expected to execute as financial intermediaries: that of brokering the relationship between borrowers and lenders and that of transforming maturity and risk (Boot & Thakor, 1997; Greenbaum, Thakor, & Boot, 2019). Havrylchyk & Verdier (2018) survey existing platforms and conclude that, although they do indeed carry out the role of a broker in the market (and add to it a degree of flexibility in exploiting novel techniques and leveraging on data other than "hard information"), they do not transform maturity and risks. The lack thereof is mitigated, in some cases, by introducing securitization (although a word of caution must be given here since excessive demand for it may lead to disastrous consequences for the market Keys et al. (2010)) and secondary markets, provision funds and diversification and bundling strategies leveraging on computational automation (Havrylchyk & Verdier, 2018).

Many authors praise the potential carried by such online platforms, naming advantages such as wider access to unsecured consumer credit for both borrowers and lenders (Emekter et al., 2015), lower intermediation fees that result in more attractive rates for borrowers (Corporate Finance Institute, n.d.), the facilitation of direct investment into community development assets (Galloway & others, 2009), the elimination or reduction of gender and racial biases when funding loans (Gonzalez & Loureiro, 2014) and even the introduction of soft information into the borrower vetting and credit scoring cycle, either in an automated fashion, or by making such information available to prospective lenders (Herzenstein et al., 2008; Iyer et al., 2009).

2.2 RISKS OF ONLINE P2P LENDING MARKETS – THE GOOD AND THE BAD

Notwithstanding the aforementioned potential, there is no shortage of literature classifying online P2P lending platforms as generally riskier than their traditional counterparts and identifying the plethora of caveats and tradeoffs that prevail in the market (Käfer, 2018). Two main domains of concern are widely identified when authors study the dynamics that lending platforms help sustain, between lenders and borrowers: information asymmetry (the difference in access to borrower and loan-related

information that exists between both parties in the market) and the inability of platforms or lenders to adequately assess loans to be issued, due to lack of training or information resources (since there are low barriers to entry into the market, lenders may lack the proficiency to select good loan funding opportunities, or overestimate borrower quality from an abundance of "soft information"; platforms, on the other hand, may be ill-equipped for properly pricing the risk of default of a given loan).

Klafft (2008) study the above topics and conclude that the notable lack of skills in lenders' ability to evaluate investment risks leads to lower quality in financing loan applications. Lee & Lee (2012) identify the phenomena of "herding" behaviors as a result of that lack of skills in platforms supporting auctioning dynamics and find strong evidence of it leading to diminishing marginal returns on funded loans. Emekter et al. (2015) conduct a comprehensive study of a dataset of issued loans from online P2P lending platform Lending Club and observe that the higher interest rates charged to higher-risk borrowers do not make up for the added likelihood of default, i.e. platforms are unable to adequately price the risk of default for riskier borrowers/loans (Mild, Waitz, & Wöckl, 2015). Käfer (2018) surveys operational aspects of online P2P lending platforms and, despite acknowledging that the value of analyzing such aspects is largely time-bound due to the volatile nature of platform operating practices and models, posits that such platforms are riskier than their traditional industry counterparts.

Other works present more positive results on how these platforms can mitigate, or even in some cases eliminate, the risks arising from the previously noted domains of concern. Information asymmetry is indeed viewed as a concern but Iyear et al. (2009) establish that despite the inherent lack of proficiency of players in the market, lenders can satisfactorily infer borrower creditworthiness by resorting not only to "hard" banking data but also "soft" information in their judgment. Although the use of such data can be seen as harmful, other studies further stress how its presence can have a positive effect, rather than deterring lenders from observing hard evidence that a borrower may present an unserviceable level of risk (Berger & Gleisner, 2009; Herzenstein et al., 2008).

Lin et al. (2013) and Weiss et al. (2010) find that P2P lending platforms have an inherent motivation to provide good borrower screening services to mitigate adverse selection and sustain that funding success is heavily reliant on trustworthy information. "Herding", another phenomenon seen as potentially hazardous by multiple authors, is observed by Herzenstein et al. (2011) to take on strategic contours that improve loan funding and create a better environment for borrowers and lenders alike, furthering the understanding that although lenders may resort to the "wisdom of the crowd", they remain sensitive and focused on inescapable data and the observable force of market economics (Yum, Lee, & Chae, 2012).

It is noteworthy that establishing the concept of online P2P lending platforms as a homogenous market segment is an unwise step in attempting to comprehend its dynamics. As is observable in the market, platforms behave and are structured in a variety of ways, some platforms making use of social networks for enabling direct contact between borrowers and lenders, to emphasize the importance of soft information in the loan funding decision process (Prosper.com, 2016), others developing stringent taxonomies and undertaking the responsibility of selecting borrowers and loans, grading/scoring them and retrieving any information that may be considered by lenders when deciding which loans to fund (Lending Club, 2019b) and, in other cases, providing minimal information to lenders and leaving research, vetting and decision-making to them (Chorzempa, 2018).

This diversity can be seen as a risk and, although the prevalence and increasing traction of such platforms has been brought to the eyes of regulators (Barth, 2012), articulating a generalized set of regulatory constraints to be enforced for the overall market will not be an easy task and is likely not expected to come to bear within the near future. Lenders are, hence, advised to take strong care when weighing and considering which data points should be included for consideration in their own decision-making.

2.3 DETERMINANTS OF DEFAULT IN ONLINE P2P LENDING MARKETS

Numerous works have been published on the subject of identifying and categorizing the information gathered and made available by online P2P lending platforms on borrowers and loans to be issued. Although limited, literature seems to indicate that, aside from the additional "soft information" made available to prospective lenders, the more traditional "hard" data provides invaluable insight into the risk of default of loan applications.

In the case of the online lending platform Lending Club, (Serrano-Cinca, Gutierrez-Nieto, & López-Palacios, 2015) establish that key factors used by the platform to calculate a loan "grade" are indeed relevant and produce satisfactory results, noting ones such as loan purpose, annual income, housing situation, credit history and indebtedness as key determinants of default. Similar conclusions are drawn by Emekter et al. (2015), Lust (2017) and Möllenkamp (2017), thus establishing current platform practices as ones yielding good guidelines (loan grade) upon which lenders can make loan funding decisions.

Polena & Regner (2018) develop prior studies of determinants further and utilize loan risk levels as a categorization mechanism under which loans are individually studied and benchmarked – findings are consistent with prior literature at a general, non-categorized, level and additionally indicate that, for low loan-risk classes, indebtedness and past delinquencies are relevant factors, whereas indicators such as the borrower's credit history are only relevant for high loan-risk classes.

Finally, Carmichael (2014) breaks down loans by their purpose and establishes both hard and soft information datapoints, such as macroeconomic indicators and application-specific data, as good predictors of loan default, to an extent where the developed scoring model can outperform the "sub-grading system" facilitated by the Lending Club.

Previous literature indicates, from the above, that a core set of determinants have an increased and prevalent relevance when estimating the probability of default for any given loan application in online P2P lending platforms. Despite their discrete application in specific contexts or for specific risk ranges/categories, variables such as borrower annual income, past inquiries, FICO score (which itself is an amalgamation of borrower credit history datapoints (FICO, 2018; Langager, 2019)) and revolving line utilization seem to indeed be good determinants to take into consideration when estimating risk scores, pricing said risk and determining a default horizon for potentially delinquent loans.

2.4 RISK SCORING IN ONLINE P2P LENDING MARKETS

Much of the work produced in surveying, analyzing and developing techniques for carrying out credit risk scoring, risk pricing and loan default prediction follows in the footsteps of literature from the larger credit risk domain, since online P2P lending markets deal with and utilize much of the same information employed by traditional credit organizations in such practices.

To assess credit risk, in developed markets lenders typically consider historical loan application and loan performance data collected regularly from a small number of sources based on long-standing banking and credit relationships to develop credit-scoring models. Traditional credit-scoring models applying single-period classification techniques (e.g., logit, probit) to classify credit customers into different risk groups and to estimate the probability of default are still the most popular data mining techniques used in the industry (Chamboko & Bravo, 2016, 2019a,b, 2020). Individual classifiers employing single statistical or operational research methods include linear and multiple discriminate analysis (DA), logistic regression (LR), probit analysis, linear and quadratic programming and data envelopment analysis. Classifiers using machine learning methods such as neural networks (NN), support vector machine (SVM), decision trees (DT), genetic and evolutionary algorithms (GA), and Bayesian Model Ensembles (BME) have also been investigated (Zhao et al., 2017; Ashofteh & Bravo, 2019, 2021a,b; Bravo et al., 2021; Bravo & Ayuso, 2020, 2021).

Olson, Delen & Meng (2012) carry out an extensive comparison between Decision Trees and other conventional data mining techniques such as Neural Networks and Support Vector Machines in an attempt to develop more transparent and transportable decision support tools for predicting bankruptcy and similar comparative work is done by Huang et al. (2004).

Dirick et al. (2017) employ survival analysis techniques to achieve higher levels of fidelity in calculating a time horizon for when a given loan will default, thus facilitating a return-focused perspective (as opposed to binary classifiers which are ubiquitous in the field). A similar approach is taken by Byanjankar (2017) for the online P2P lending market where promising results are achieved for predicting the survival probabilities of borrowers at different times.

Đurović (2017) analyses the probability of default in a dataset of issued loans from online P2P lending platform Lending Club, by employing a survival analysis model to analyze the probability of default for loan term length and purpose. Findings yield that there are riskier loan purpose classes but also assert that a longer term is associated with a higher likelihood of default. Tan et al. (2018) study a similar dataset and propose an approach relying on differentiating loans based on negative outcomes (either “charge-off” or “pre-payment”) and, leveraging on deep neural networks, outperforming the risk analysis capabilities of traditional risk assessment capabilities of traditional survival analysis techniques.

Mezei, Byanjankar & Heikkilä (2018) use linguistic data transformation as a discretization step in traditional supervised learning algorithms to successfully improve their classification performance. Byanjankar, Heikkilä & Mezei (2015) develop a credit scoring model using Artificial Neural Networks (ANNs) to evaluate credit risk in P2P lending, achieving satisfactory results.

More comprehensive works tackle a variety of options for determining risk. Jin & Zhu (2015) develop a more comprehensive framework for approaching such datasets and combine an introductory step of identifying the determinants of risk through Random Forest (RF) and importance analysis techniques to then serve a multitude of data mining algorithms for comparison. They conclude that SVMs outperform other methods, although by a marginal amount. Carmichael (2014) uses a discrete-time hazard model (dynamic logistic regression) to estimate the probability of default of P2P loans. By using a mix of both hard and soft information, the model can achieve much better results when compared to the surveyed platform’s credit grading system.

3 METHODOLOGY

3.1 THE MODEL - LOGISTIC REGRESSION

In this work, we employ a logistic regression model to assess the factors that determine the loan performance of P2P loans. Logistic regressions model the probability of an event occurring depending on the values of a set of independent variables, predicting their effect on a binary variable response and classifying observations by estimating the probability that an outcome has in a particular category.

This statistical technique has been the most widely used in credit scoring applications (Abdou & Pointon, 2011) and is considered the industry standard for developing credit scoring models (Ala'raj & Abbod, 2015; Lessmann, Baesens, Seow, & Thomas, 2015).

For this study, *loan status* is our dependent variable with the following binary outcome: 0 (zero) if the loan status is "*Fully Paid"*, meaning that borrower was able to repay the loan and 1 (one) if the loan status is *"Default"*, meaning that the borrower is unable to fully pay the loan and default[2]. In this section, the work of Hosmer Jr et al. (2013) was closely followed.

Considering the above, the probability of a loan default (conditional probability) given the independent variable can be written as $Pr(y_i = 1|x_i) = \pi(x_i)$, where the borrower i will default given the information x_i. If this probability is greater than 50%, the model predicts that the borrower belongs to the "*Default*" class, otherwise the model predicts that the borrower instead belongs to the "*Fully Paid*" class.

The first step to determine this probability is to estimate a linear regression function:

$$g(x_i) = ln\frac{\pi(x_i)}{1-\pi(x_i)} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_j x_{ij} \qquad i = 1, 2, \dots, n \qquad \textbf{(1)}$$

where β_0 is a constant, β_j is a vector of regression coefficients and $\vec{x}_i$ is a vector of independent variables. Thus, x_{ij} is the value of the variable X_{ij} for the i^{th} borrower, with $j = 1, \dots, \mathcal{K}$ ($\mathcal{K}$ is the number of independent variables).

However, in a linear regression function, the probabilities can be less than 0 and greater than 1, hence the model needs to restrict the probability between 0 and 1 by including a non-linear function (logit) onto equation (1) transforming into logistic regression function (equation 2):

$$\pi(x_i) = \frac{e^{g(x_i)}}{1+e^{g(x_i)}} \qquad i = 1, 2, \dots, n \qquad \textbf{(2)}$$

Having the model specified, the next step is to estimate the coefficients β_0 and β_j. The method used to fit the non-linear model is the Maximum Likelihood Method (MLM). This method generates values for the unknown coefficients that maximize the probability of obtaining the observed set of data. The likelihood function is shown in equation (3).

[2] Please refer to the next chapter, Data, for a further enlightenment on the variable loan status

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i}\left[1 - \pi(x_i)\right]^{1-y_i} \quad \textbf{(3)}$$

If the variable y_i takes the value 1 (the borrower default), the likelihood is π_i; if not, the likelihood is equal to $1 - \pi_i$. To simplify, mathematically, the optimization of the likelihood function, the logarithm of the equation is normally maximized. The log-likelihood is defined as:

$$L(\boldsymbol{\beta}) = \ln[l(\boldsymbol{\beta})] \sum_{i=1}^{n} y_i \, ln\left[\pi(x_i)\right] + \left(1 - y_i\right) ln\left[1 - \pi(x_i)\right] \quad \textbf{(4)}$$

Therefore, to find the value of $\boldsymbol{\beta}$ that maximizes the computation $L(\boldsymbol{\beta})$ we can isolate the cases β_0 and β_j, setting the equations equal to zero and solving the problem to obtain the desired result. Due to the nonlinearity of logistic regression equations, software-based computation methods must be leveraged to compute them, addressing the iterative calculations required.

3.2 PERFORMANCE MEASUREMENTS OF THE MODEL

To assess the predictive capacity of our model, we record and compare its predictions to the actual, known, observations. A True Positive (TP) occurs whenever a **positive prediction** (outcome equal to "1", or "default") made by the model coincides with the observation, whereas a False Positive (FP) is recorded when the model's **positive prediction** is wrongly classified and contradicts the observation. True Negatives (TN) and False Negatives (FN) denote a similar logic, but for **negative predictions** (outcome equal to "0", or "fully paid").

Table 1 - Confusion Matrix

		Observed	
		Positive	**Negative**
Predicted	**Positive**	TP	FP
	Negative	FN	TN

Source: Author's production.

To verify the predictive capacity of our logistic regression model, it is necessary to understand its discrimination function. For this, and considering the above confusion matrix, we can calculate the following performance measures:

Accuracy measures the overall ratio of correctly predicted observations and can give an overall view of the model's performance. However, looking at this measure alone can build a false perception of the model due to its narrow focus.

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \quad \textbf{(5)}$$

Sensitivity corresponds to the probability of the correct classification of the event of interest to occur, i.e., how well the model correctly classifies true positives.

$$Sensitivity = \frac{TP}{TP + FN} \quad \textbf{(6)}$$

Specificity is defined as the proportion of observations correctly predicted to belong to the negative class.

$$Specificity = \frac{TN}{TN + FP} \quad \textbf{(7)}$$

Alternative methods to evaluate the performance of a classification model with binary classes include the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) (Wendler & Gröttrup, 2016). The ROC shows the values for which there is greater optimization of Sensitivity as a function of Specificity, i.e., plots the true positive rate (TPR or Sensitivity) of a discrete classifier against the false positive rate (FPR or 1-Specificity). The AUC, in turn, being a function of both the TPR and the FPR, provides a measure of the model's ability to differentiate between classes, or its discrimination power – how effectively it can predict negative and positive outcomes correctly.

As a baseline we can understand how the optimal value for AUC would be 1 – the curve would pass by point (0,1), which would then maximize AUC and give the optimal result of 100% TPR and 0% FPR. Contrarily to this, the baseline function drawn for $\pi(x_i)$ in the ROC curve plot provides the lower bound for a model without classification power and a calculated value of 0.5 for AUC.

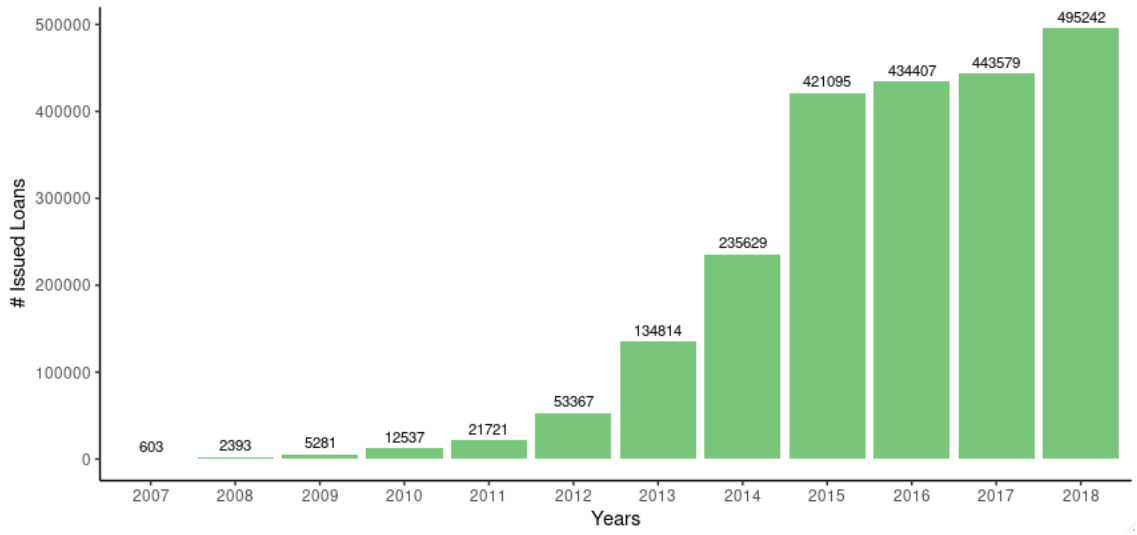
4 DATA

This chapter is divided into 4 sections. In the first section, “Data Description”, we provide a brief, high-level description of the data collected. In the second part, “Data Refinement”, we proceed to describe the steps taken to clean up the dataset and elaborate on how the data refinement was accomplished. After that treatment of the starting dataset, we conduct an exploratory analysis of the new, refined, dataset in the third section “Exploratory Data Analysis”. There we summarize the key observations on variables of the dataset, identifying outliers and handling missing values. At the end of this section, we provide the descriptive statistics and correlation between the selected variables.

4.1 DATA DESCRIPTION

The data under analysis were retrieved from the Lending Club website[3], the largest P2P lending platform in the United States. The data collected contains information on more than 2,200,000 loans issued through the platform from 2007 to 2019. Although the latest observations in the dataset are from March 2019, it contains only loans issued until December 2018. Figure 2 shows the number of loans issued throughout the years and in it, we can observe a prominent growth trend in the total number of issued loans until 2015, where the number of loans approximately doubles year-to-year. From 2015 onwards there is still an increase, but it is subdued when compared to the previous time frame.

Figure 2 - Number of loans issued by year



Source: Author’s production.

The original dataset is composed of 151 variables covering details across 3 main topics, which are focused on: borrower details, loan characteristics and final loan status (where available). Borrower information includes data points such as occupation, employment details, credit rating and other additional information which provides Lending Club users with an overview of who the borrower is

[3] Downloaded from https://www.lendingclub.com/info/download-data.action

and what their current financial/labor status is. Loan characteristics include details such as the amount of the loan, interest rate, monthly installment, purpose of the loan, and other loan-specific elements, that clarify how the loan is to be repaid. Finally, the information provided on the final loan status informs of the latest status of the loan, at the time of observation (i.e. the loan can either be paid off, defaulted, overdue or current) and provides additional context, depending on that same status – e.g. is there a payment plan for the defaulted loan. The full description of all 151 variables can be found in Appendix A. Out of 151 variables, we have selected 22 independent variables, as not all the variables are relevant for this analysis. The selection of the variables in the study is explained further in the following sub-sections.

4.2 DATA REFINEMENT

In this section, we describe all the steps involved in the treatment of the dataset and how we have proceeded to refine the data therein.

The first step in our treatment of the dataset consists of narrowing down the time frame under study. We proceed to filter out the loans that have not reached maturity yet, to guarantee consistency between the data points available. With this, we have removed all loans with a 36-month repayment term that were issued after February 2016 and all loans with a 60-month repayment term that were issued after February 2014. The loans issued in 2007 were also removed as the borrower information initially requested was different from the following years.

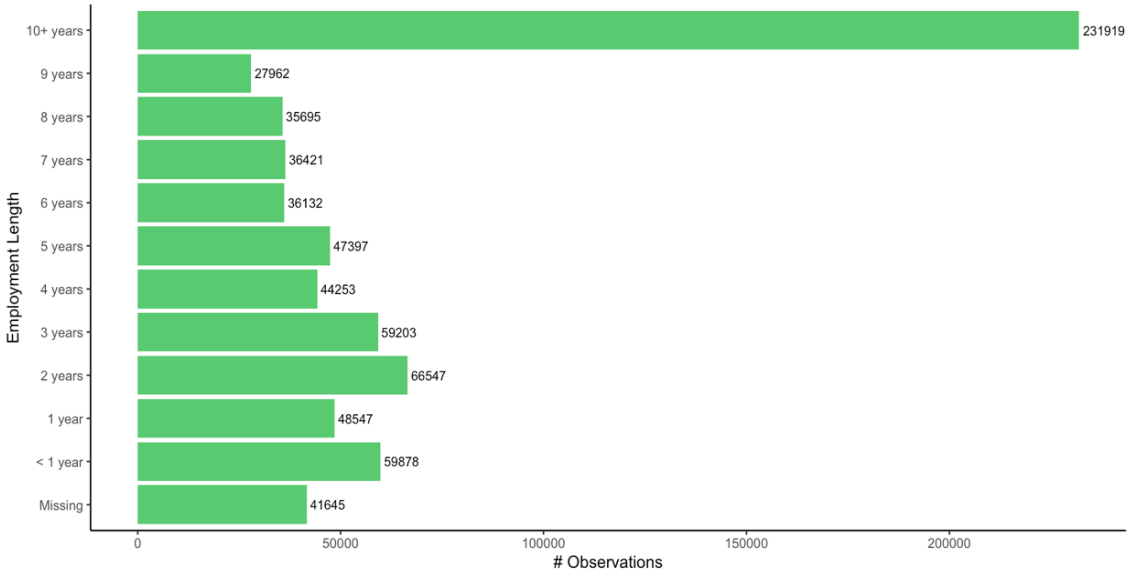
As previously mentioned, not all the variables are relevant to our study. Our goal within this section is to identify predictors of default so that at the time of the loan request, the platform can use those variables as a reference to decide on whether to approve or reject the loan. Due to this, any variables that are not available at the time of the loan application cannot be used as predictors for credit approval. For this reason, we removed variables such as *next_pymnt_d*, *hardship_flag*, *pymnt_plan* among others related to these.

We leave out some variables because they represent the same data as others, such as *funded_amount* and *funded_amount_inv* that represent the same as *loan_amnt*. Other variables are also deemed to not be relevant as they don't have any impact on the borrower's likelihood of default, such as *id*, *member_id* and *policy_code*, and given this, they are also removed. We also proceeded to remove dataset columns such as *emp_title* and *url*, because they contain too many unique values, making them difficult to categorize and analyze, so we opt to leave them out of the study.

Another reason to delete columns is the high frequency of missing data in some variables. We have decided to leave out all the variables with more than five percent (5%) of missing values. Therefore, we have left out variables such as *mths_since_last_record* and *mths_since_last_delinq*.

A further refinement is conducted on the *emp_length* variable, due to it having a substantial volume of missing values. To prevent discarding these data points (which amount to approximately 6% of the total dataset), we emploied a 'binning" technique and group all missing values into a separate group labeled "missing". Figure 3 illustrates the distribution of the number of observations by the length of employment.

Figure 3 - Number of observations by *employment length*



Source: Author's production.

It can be observed in Figure 3 that the majority of the categories have a substantially reduced number of observations when compared with the category "10+ years". We chose to aggregate some categories together, resulting in two new categories ("0 - 4 years" and "5 - 9 years"). This step was executed to simplify the categories available in variable *emp_length*. Table 2 below shows the new distribution of the variable *emp_length*.

Table 2 - Summary of the *employment length* variable

Missing	0 - 4 years	5 - 9 years	10+ years
41,645	278,428	183,607	231,919

Source: Author's production.

The only other variables with missing values were *revol_util* and *pub_rec_bankruptcies*, with 0.06% and 0.11% of missing values, respectively. As these percentages represent a total of 1,208 observations we decided to remove these missing values.

As previously stated, loan status is the dependent variable of this study and it is divided into six categories described in Table 3. To adapt the variable to our model, we have to transform it into a variable containing a binary outcome/value, 0 or 1. When the *loan_status* assumes the value of one (1) it represents the failure of payment, meaning that the borrower is unable to fully pay the loan and eventually defaults. On the other hand, when *loan_status* assumes the value zero (0), this means that the borrower had paid the loan, fulfilling the established contract.

Table 3 - Number of loans by *loan status*

Initial Data Set Distribution		Final Data Set Distribution	
Loan Status	# of Loans	Loan Status	# of Loans
Charged Off	109,258	1 - Default	109,797
Current	433	0 - Fully Paid	623,770
Does not meet the credit policy. Status: Charged Off	539		
Does not meet the credit policy. Status: Fully Paid	1,521		
Fully Paid	622,249		
In Grace Period	40		
Late (16-30 days)	34		
Late (31-120 days)	317		
Total	734,391	Total	733,567

Source: Author's production.

To proceed with the transformation of the variable, we have removed all the loans with "*Current"* status as the borrowers are still making monthly payments and we don't yet know the final status. Loans with the "*In Grace Period"* status represent the loans that are at the most 15 days with a late payment and loan statuses with names *"Late (16-30 days)"* and "*Late (31-120 days)"* represent the loans have a delayed installment between 16-30 and 31-120 days, respectively. While these loans are in some way in default, we filtered them out in this study as the probability of them being paid back is significant (26% to 72%) compared with the "*Default"* status (11%) (Lending Club, 2019a).

We then proceeded to aggregate the status "Charged Off" with "Does not meet the credit policy. Status: Charged Off" into a newly created status labeled as "Default", since both correspond to the same outcome. The status "Does not meet the credit policy. Status: Fully Paid" is also combined into the "Fully Paid" category. Table 3 shows the initial distribution of the number of loans by status and the final distribution after the adjustments.

The loan description provided by the borrower on why the loan is needed is contained in the variable *desc* and it is optional information that most of the borrowers leave blank. The variable by itself is difficult to analyze since this is a free-style field. To retrieve value out of this variable, we created a dummy variable, *writing_skills*, that is derived from the original user input. If the borrower has left a description on this field, he or she will have automatically 10 points. Then, we check if there are any spelling mistakes on the description - if detected, 2 points will be deducted from the initial 10. We also check if the text has no end marks, i.e., if it is missing ending punctuation - if detected, another 2 points will be deducted. The possible values of *writing_skills* are 0, if the borrower hasn't written any description, 6 if there is any word misspelled and any missing punctuation, 8 if there is any word misspelled or any missing punctuation, and 10 if there is a description with no spelling mistakes and with correct punctuation.

Although the *desc* variable is only filled in by approximately 17% of borrowers, earlier works have indicated that this variable may be a good predictor of default (Carmichael, 2014). We employ this

composition of textual analysis methods to derive numeric data which is easier to analyze and to use within the context of our prediction model.

Lastly, the date of the earliest credit line opened by the borrower (month and year) is shown in *earliest_cr_line*. We have created a dummy variable, *credit_history*, where we have subtracted the loan's issue date by the date of the earliest credit line. We now have the number of years of credit history considering the date when the borrower's earliest reported credit line was opened.

The aforementioned refinement steps result in the reduction of variables from a starting base of 151 to a total of 25. In the sub-section that follows we proceed to analyze the resulting dataset and conduct any additional refinement that may be deemed necessary, before building our model.

4.3 EXPLORATORY DATA ANALYSIS

To further refine our dataset and to prepare it for usage in building our prediction model, we have conducted an exploratory analysis of the provided data.

We begin this exploratory data analysis by investigating the outliers that may exist on the dataset. Our first finding focuses on the variable *annual_inc*. Table 4 provides a summary of this variable and at first glance, we can detect some values that seem unrealistic for this data. The maximum annual income of $9,000,000 compared with the median annual income of $62,000 and the 3rd quartile of $89,000 suggests that there are some outliers. Figure 4 shows the distribution of borrowers' annual income. Observing Table 4 along with Figure 4, we can confirm the evidence of outliers.

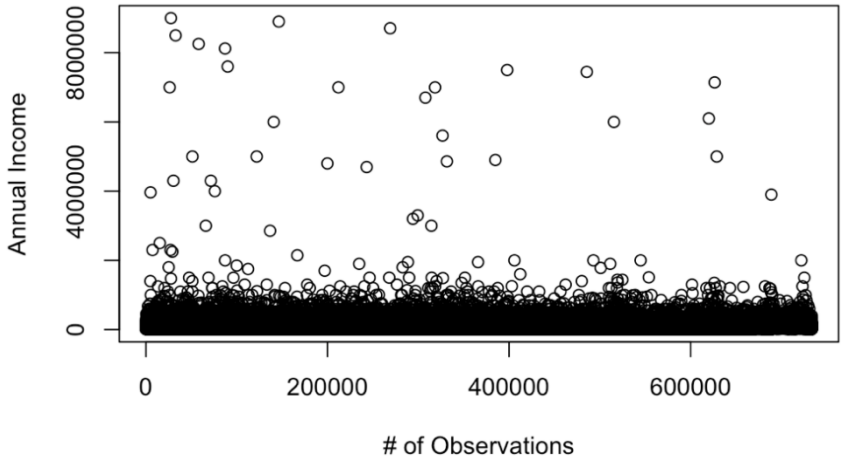
Table 4 - Summary of the *annual income* variable

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	45,000	62,000	73,689	89,000	9,000,000

Source: Author's production.

A deeper analysis of the remaining information provided for these observations shows that the majority of the borrowers are requesting small loan amounts for debt consolidation. Taking all these details into account, we consider these records to be the result of human error and, therefore, we have limited the highest value of annual income to $1,000,000. In total, there were 139 observations removed.

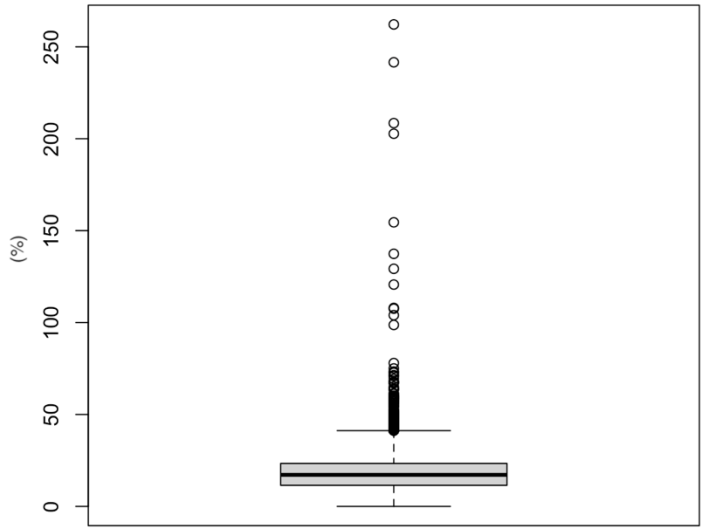
Figure 4 - Distribution of *annual income*



Source: Author's production.

Another finding concerns the *dti* variable (debt-to-income ratio), which represents the ratio of the borrower's total monthly debt payments over their monthly income. Figure 5 displays the boxplot of the variable *dti*.

Figure 5 - *Dti* outliers

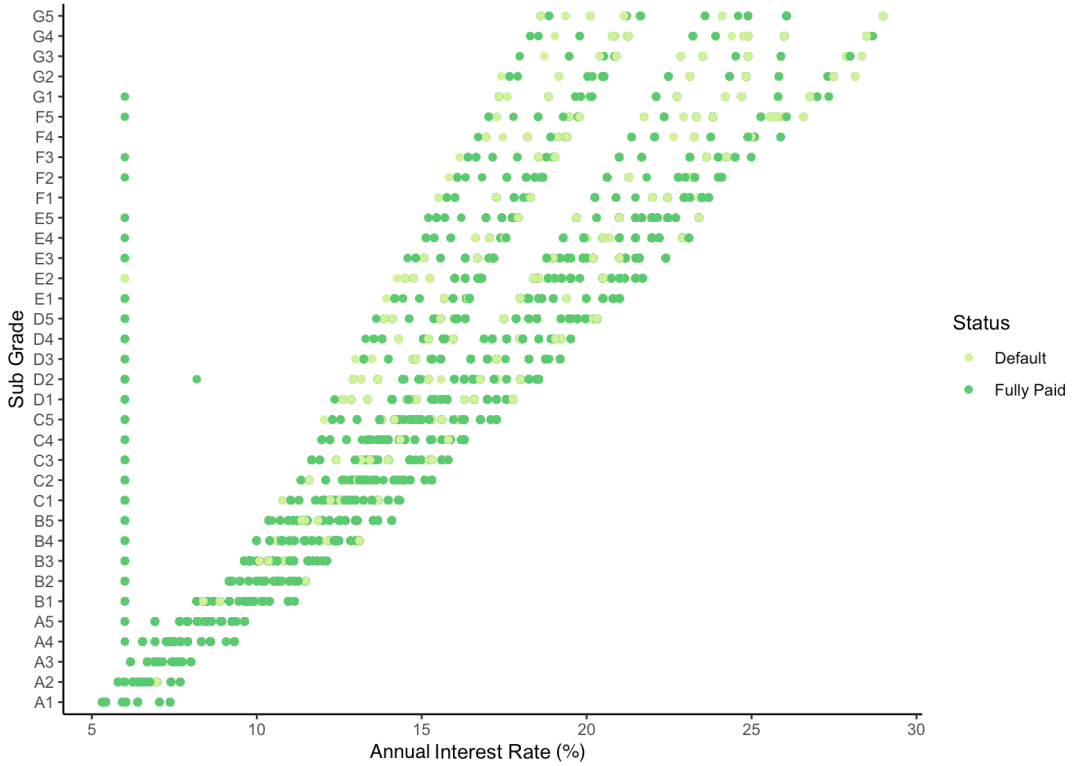


Source: Author's production.

In observing Figure 5 it becomes noticeable that the majority of loans are concentrated in the range of values between 10 and 25 and there is evidence of several outliers. We believe that these outliers are caused by an error as the data provided would indicate that some borrowers have significantly more outstanding debt than their current income – we also observed the value of *dti* to be inconsistent with some of the other data points provided, for some borrowers – thus, we have removed all the outliers. Compared with other studies, for instance, Carmichael (2014), Emekter et al. (2015) and Möllenkamp (2017), we can see that after the removal of the outliers, the maximum value for this variable is as expected. The exclusion of *dti*'s outliers decreases our dataset by 484 observations.

From a lender's perspective, the *interest rate* is an important variable in deciding whether or not to fund a loan. We examine how this variable interacts with the variable *subgrade* by computing a scatter plot. Naturally, the better the subgrade, the lower the interest rate and vice-versa. This tendency is shown in Figure 6.

Figure 6 - Relationship between *interest rate* and *subgrade*



Source: Author's production.

From Figure 6 we can observe that most of the loans are graded between A and C and that the amount of defaulted loans increases with worse grades. One intriguing observation is that almost every subgrade has observations with the same interest rate (6%). These observations seem to be outliers created by a potential error on the update of the interest rate, as by analyzing their loan characteristics they are consistent with other loans within the same grade. Therefore, we have left out 180 observations from the dataset.

After analyzing the outliers, we examine the variable *home ownership*. Table 5 shows the distribution of the loans by each category of *home ownership*. Looking at the factors of this variable: other, any and none, they represent a very small part of this variable (0.03% of the full dataset) and since they are undefined by Lending Club (which complicates its analysis and use within our study) we have chosen to remove them from our dataset. With this change, there were an additional 224 observations removed from the dataset.

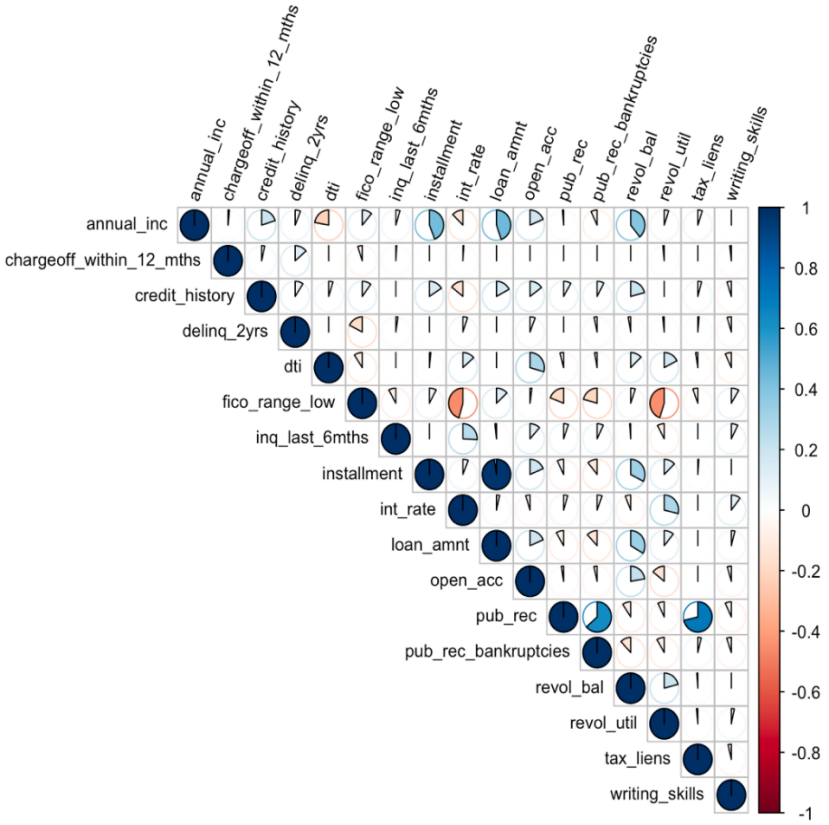
Table 5 - Distribution of loans by *home ownership*

ANY	MORTGAGE	NONE	OTHER	OWN	RENT
2	350 430	42	180	73 809	308 300
0.0003%	47.82%	0.01%	0.02%	10.07%	42.07%

Source: Author's production.

Lastly and for a better understanding of the way that variables are connected, we computed a correlation matrix for the continuous variables. The matrix below provides a visualization of the correlation that variables have between themselves. The detailed values of the information displayed in this graph can be found in Table 10 in Appendix B.

Figure 7 - Correlation Matrix of numerical variables



Source: Author's production.

Results show that the highest correlation of 0.9777 is obtained between *loan amount* and *installment*. These two variables are highly correlated and to ensure the precision of our model, we decide to remove the latter from the study. *Tax liens* and *public record bankruptcies* bear the second and third highest correlation against *public records* with 0.7131 and 0.6351, respectively. It is expectable that a certain level of correlation among the independent variables will exist and, even though some of the variables display a high degree of correlation (excluding *installment*), that same degree is still moderate. We have decided not to remove any of these variables as we could lose valuable information for the study.

Additionally, we have removed *subgrade* as we believe that this variable, created by Lending Club, is already an outcome of the platform's internal processes/models used to identify the risk of default of the loans and it provides the same information as the variable *grade*.

4.4 FINAL VARIABLES

After the refinement was conducted, we arrived at a new, sanitized dataset where we were able to reduce the dataset's contents from 151 variables to 22 variables across a total of 732,540 loan observations. In the final dataset 109,619 loans (15%) defaulted and 622,921 loans (85%) were fully paid.

Tables 11 and 12, presented in Appendix C, report the descriptive statistics of the numerical and categorical variables, respectively.

Upon reviewing the last column of Table 11, titled "Average by loan status", we can observe some differences between the average values for loans with the status of *Fully Paid* and *Default,* across other variables. Loans that are *Fully Paid* have significantly higher *annual income*, longer *credit history,* lower *debt-to-income* ratio, lower *interest rate*, higher *revolving balance,* and a lower *revolving utilization ratio* than the *Default*ed loans. On the other hand, borrowers that were not able to pay have higher average *loan amount*, more delinquencies in the past 2 years, lower FICO score, more inquiries in the last 6 months, more public records with more bankruptcies, and more accounts open than the borrowers that paid in full.

Table 12 displays the distribution of the categorical variables in our dataset and the default rate for each category. As we can see on the variable *employment length*, the category "missing" stands out among the others, with the least number of observations (5.7%) and with the highest default rate (21%). This seems to indicate that borrowers who do not have an employment history, choose not to provide it or work in a field which lacks employment stability (like freelancing) are at higher risk of default. On the other hand, the remaining categories have the same trend regarding the number of observations, varying from 25% to 37.8%, and the default rate, from 13.8% to 15%.

Regarding the variable *grade*, we can see that most of the issued loans are graded "B" (almost 33% of observations) and about 80% of the issued loans have their grade between "A" and "C". As credit grades get worse, the chance of the loan defaulting increases. Grade "A" has a default rate of 5.5% while "F" and "G" have the highest default rate, 34.80% and 37.90% respectively. Even though the lower grades are riskier, they only represent 1.8% of the total observations. These findings are consistent with the findings displayed in Figure 6 and suggest that Lending Club is actively and effortfully endeavoring to adequately price the risk of issued loans, following their assigned grades.

Observing the variable *home_ownership*, "mortgage" (47.8%) and "rent" (42.1%) are the home situation that most borrowers live in. Somehow, borrowers that own their home have a default rate higher than a borrower that is paying a mortgage, 15.5% and 12.8% respectively. However, the borrowers that are paying rent are the riskiest, with a 17,3% of default rate.

Analyzing the variable *purpose*, we can conclude that the main purpose for a credit application continues to be "debt consolidation", with 57.2% of loans, followed by the repayment of "credit card"

debt, with 23.8%. Although these are the most popular reasons for a loan, they are not the ones with the highest default rate. “Small business” loans and loans for “educational” purposes are the 2 leading categories and the ones offering the highest risk of default, with default rates of 24.4% and 20.2%, respectively. Contrarily to this, “car” loans and “wedding” loans offer the lowest risk of default, with default rates of 11.9% and 12.3%, respectively.

Looking at the variable *term*, the majority of the loans have 36 months duration and the 60-month term only represents 8.7% of loans in the dataset. Despite the small fraction on the dataset, the latter duration has a very high default rate (25.3%) when compared with loans with a 36-month duration (14%).

The last variable under analysis, *verification_status*, is divided into three different categories. As per Lending Club[4], the definition of each category is as follows:

- when a borrower states that they earn X amount income and the platform can verify an income level that is within an acceptable range of X, the loan is labeled as “verified”;
- when a borrower claims that they work at a certain company and Lending Club can verify that the borrower does indeed work there, the loan is categorized as “source verified”;
- a loan is categorized as “not verified” if there is no kind of verification.

With regards to this dataset, the majority of the loans have some kind of verification, either where the borrower works (34.9%) or how much he/she earns (33%). Surprisingly, the category with the lowest default rate is “not verified” (11.7%), while “verified” and “source verified” have default rates of 18% and 15.1%, respectively.

[4] Please refer to: https://www.lendingclub.com/investing/investor-education/income-verification

5 RESULTS AND DISCUSSION

5.1 MODEL AND RESULTS

To build our logistic regression model, we first proceeded to split the final (post-refinement) dataset into two sub-datasets, namely a "**training**" and a "**test**" datasets. The **training** dataset, which is composed of 2/3 (66,67%) of the final dataset, is used to construct (train) the model and arrive at one that best explains (or predicts) the dependent variable as a function of the explanatory variables. The test dataset, which is composed of the remaining 1/3 (33,33%) of the final dataset, is used to assess the predictive capacity of the generated model.

A selection of the independent variables in the study was done before the construction of the model[5]. Once the data treatment was completed, we proceed to fit the model using stepwise regression to ensure that we have a well-fitted model. Such fitting methods work by gradually selecting and/or deleting variables from their working set, based on algorithmic checks that are conducted to assess the "importance" of variables, and proceeding to include/exclude them (Hosmer Jr et al., 2013).

In the case of this study, we employed the backward stepwise approach, where the model deletes the variables that have the most statistical insignificance for the model.

For the model developed in our study, the following significant variables were identified for their correlation with estimating the likelihood of borrower default: loan amount, term, interest rate, grade, employment length, home ownership, annual income, verification status, purpose, debt-to-income, FICO score, inquiries last 6 months, open accounts, charged-off within 12 months, public records, revolving balance, revolving line utilization rate, writing skills and credit history.

Table 6 presents the regression results of the model. The first column shows the estimates of the regression parameters. These parameters give the sign of the variables' effect on the probability of default, however, they are not easy to interpret by themselves.

To assist in our analysis of the results, we compute the "odds ratios" of each variable as a way to assess the level of association between the given independent variables and the dependent variable. This value is obtained via the calculation of the exponential function of the regression coefficient, yielding the expected variation in the dependent variable for a one-unit increase in the independent variable.

These additional computations can be reviewed in the fifth column of Table 6. Predictors with a positive sign on the first column and an odds ratio higher than 1 indicate a higher probability of loan default. The negative sign on the first column indicates the opposite.

The results in Table 6 show that 26 out of 41 variables are highly significant considering the estimation done with MLM.

[5] Refer to section 4

Table 6 - Logistic Regression results

Predictors	Estimate	Std. Error	z value	Pr(>|z|)		Odds Ratio
(Intercept)	0.30591	0.16235	1.884	0.060	.	1.358
loan_amnt	0.00001	6.667E-07	18.607	< 0.000	***	1.000
term 60 months	0.32868	0.01511	21.747	< 0.000	***	1.389
int_rate	0.04228	0.00323	13.073	< 0.000	***	1.043
gradeB	0.40148	0.01988	20.191	< 0.000	***	1.494
gradeC	0.67665	0.02690	25.155	< 0.000	***	1.967
gradeD	0.79376	0.03613	21.967	< 0.000	***	2.212
gradeE	0.84355	0.04545	18.559	< 0.000	***	2.325
gradeF	0.85330	0.05768	14.793	< 0.000	***	2.347
gradeG	0.79135	0.07857	10.072	< 0.000	***	2.206
emp_length10+ years	-0.01574	0.01061	-1.483	0.138		0.984
emp_length5-9 years	0.02372	0.01069	2.219	0.026	*	1.024
emp_lengthMissing	0.41624	0.01802	23.096	< 0.000	***	1.516
home_ownershipOWN	0.12392	0.01456	8.509	< 0.000	***	1.132
home_ownershipRENT	0.24525	0.00947	25.891	< 0.000	***	1.278
annual_inc	-2.67E-06	0.00000	-19.384	< 0.000	***	1.000
verification_statusSource Verified	0.11782	0.01098	10.732	< 0.000	***	1.125
verification_statusVerified	0.05793	0.01161	4.989	0.000	***	1.060
purposecredit_card	0.02303	0.04381	0.526	0.599		1.023
purposedebt_consolidation	0.09179	0.04311	2.129	0.033	*	1.096
purposeeducational	0.66054	0.17547	3.764	0.000	***	1.936
purposehome_improvement	0.11179	0.04650	2.404	0.016	*	1.118
purposehouse	0.16889	0.07127	2.37	0.018	*	1.184
purposemajor_purchase	1.11E-01	0.05197	2.132	0.033	*	1.117
purposemedical	0.19839	0.05638	3.519	0.000	***	1.219
purposemoving	0.17971	0.06090	2.951	3.17E-03	**	1.197
purposeother	0.08760	0.04587	1.91	0.056	.	1.092
purposerenewable_energy	0.20129	0.14305	1.407	0.159		1.223
purposesmall_business	0.50026	0.05270	9.493	< 0.000	***	1.649
purposevacation	0.01789	0.06553	0.273	0.785		1.018
purposewedding	-0.19333	0.08965	-2.157	0.031	*	0.824
dti	0.01599	0.00058	27.689	< 0.000	***	1.016
fico_range_low	-0.00538	0.00021	-26.107	< 0.000	***	0.995
inq_last_6mths	0.08103	0.00408	19.841	< 0.000	***	1.084
open_acc	0.01003	0.00091	11.024	< 0.000	***	1.010
pub_rec	0.01100	0.00694	1.586	0.113		1.011
revol_bal	-2.24E-06	0.00000	-7.187	0	***	1.000
revol_util	-0.00111	0.00022	-5.028	0.000	***	0.999
chargeoff_within_12_mths	-0.07241	0.03866	-1.873	0.061	.	0.930
writing_skills	-0.00715	0.00130	-5.502	0	***	0.993
credit_history	-0.00543	0.00062	-8.7	< 0.000	***	0.995

Source: Author's production. Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Analyzing the results, our model finds *grade* to be highly significant on all categories and that the probability of default increases as the borrower's grade worsens. The grades D, E, F and G are the variables with a higher odds ratio having the most significant impact on our model and thus on the probability of default. Grade F registers the highest positive correlation among all categories and also an odds ratio of 2.347. This variable is a good predictor of default and good for investors to get a substantial insight into the borrower's creditworthiness. These results are aligned with the majority of the literature presented (Emekter et al., 2015; Möllenkamp, 2017; Serrano-Cinca et al., 2015).

The same happens with the *interest rate*. This variable is also highly significant on the prediction of the default that with the increase of the rate, the probability of default also increases. This finding is consistent with the findings in Figure 6 represented in the previous chapter.

Loan amount and *term* are other variables to be highly significant in predicting the failure of the loan payment. The positive sign on these variables suggests that borrowers with a larger loan amount and a longer period of loan payment are more likely to default. Everything else equal, this proves that borrowers who request larger loan amounts on a longer-term and therefore will have larger installments for a longer period of payment are more likely to fail the payments.

The variable *home ownership* shows that a borrower paying rent or owning a house is more likely to default than a borrower that pays a mortgage. These results are in line with the data previously revealed on the exploratory analysis where these two categories have a higher default rate than the latter one.

Further to this, the two categories that have some kind of verification on the variable *verification status* are also highly significant predictors. The results are in line with the data examined in the previous chapter. This may suggest that Lending Club verifies the borrowers that they believe to have a higher risk of default.

Annual income, FICO and dti are also highly significant variables. The negative sign on the two first variables implies that a borrower with a higher income and a greater FICO score is less likely to default. Contrarily, the last of these variables with a positive sign indicates that borrowers with higher amounts of debt struggle more to repay their loans.

The borrower's *purpose* for requesting a loan has 14 categories. Of those 14, 10 categories are significant at different levels. Borrowers needing a loan to help them finance their "education", "small business" or "medical expenses" are riskier borrowers than those who are financing their "wedding", "vacation" or repaying their "credit card" debt. Actually, the category "wedding" registers a negative correlation with the likelihood of default revealing to be the only purpose that affects negatively the probability of default with an OR of 0.824.

On the variable *employment length*, although the category 5-9 years is significant (at 5% level), only the category "missing" is highly significant at the 0.1% level. Also, the latter has a coefficient with a positive sign and OR of 1.516, having a big impact on the likelihood of default. The category 10+ years, despite not being significant, is the only one from this variable with a negative sign, having the opposite effect on the probability of default. These results are aligned with the data in Table 12 of Appendix C. A possible justification for the increased likelihood of a borrower defaulting, when registered under the "missing" category, may be due to him being unemployed and, hence, being more likely to default.

Regarding the variables related to the credit history of the borrower, *inquiries last 6 months, open accounts, revolving balance, revolving line utilization rate* and *credit history* are also highly significant. This is in line with our expectations, *inquiries last 6 months* and *open accounts* show the borrowers lack repayment of a loan, increasing the probability of default. On the other hand, *revolving balance, revolving line utilization rate,* and *credit history* show a negative relationship with the likelihood of default of a borrower.

Surprisingly, the variable *charged off within 12 months* has a negative impact on the probability of default of the borrower, i.e., the borrower with more charge-offs (on other loans) in the last 12 months is more likely to fully pay back the loan.

The variable *writing skills* is highly significant and affects the probability of default negatively. When the value of writing skills increases, the probability of the borrower paying also increases.

Regarding *public records*, the variable seems not to have any significance on the model investigated, however has an odds ratio of 1.011, which is higher than other variables with a higher level of significance.

5.2 MODEL PERFORMANCE

The predictive capacity of the model was tested by leveraging the remaining 33% of the sample dataset which were not used for building our logistic regression model (i.e. used in the "training" set). We then proceeded to use this "testing" set to retrieve a series of performance metrics to analyze the model, as described in section 3.2.

Table 7 provides the summary of the outcomes in a contingency table.

Table 7 - Confusion Matrix

		Observed	
		Positive	**Negative**
Predicted	**Positive**	207745	79
	Negative	36292	64

Source: Author's production.

Based on the values of the above table, we calculated the following performance measures.

Table 8 - Model Performance Measures

Accuracy	0.851
Sensitivity	0.851
Specificity	0.448

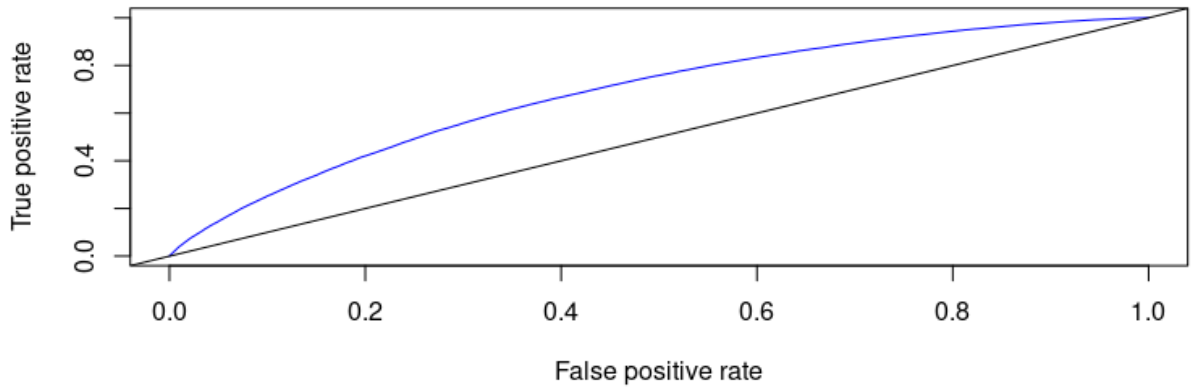
Source: Author's production.

The accuracy value of the test dataset in the study has a very good outcome. Nevertheless, we have in mind that this result is dependent on the manual split of the data that we did earlier and that this measure alone is not a reliable indicator of the model's performance.

Concerning the other two measures, the model performs better in distinguishing borrowers that defaulted than borrowers that fully paid, as it yields better results for sensitivity than specificity.

Figure 8 shows the ROC curve of our model. As we can observe, the curve is very close to the baseline (black line that represents the balance TPR=FPR).

Figure 8 - ROC Curve



Source: Author's production.

By reviewing the AUC indicator, its result reveals a value of 0.6835, showing that, although our model can accurately predict loans that will default, it is far less capable at predicting loans that will not do so. This represents a lack of ability to differentiate between the 2 classes of our dependent variable – "default" VS "fully paid".

We can conclude from the results and related performance measures that the model's classification power is at the best "average". We believe that this lack of quality may be due to the imbalanced classification in our dependent variable as it is a common limitation within loan classifications and credit risk modeling. As per Singh, Tsai, & Ramiah (2014), this imbalance negatively impacts algorithms such as Logistic Regression that optimizes across the entire training set.

6 CONCLUSIONS

In an increasingly financialized world, where new financial technologies and service models keep on disrupting the established and traditional practices and industries, peer-to-peer (P2P) lending seeks to democratize access to credit for a wider range of borrowers (including those deemed by conventional finance to be non-creditworthy) and also to unsecured consumer debt, for a wider range of investors. In doing so, it is bridging a demand and a supply gap which currently exists in this largely untapped market.

Lending Club, being one of the pioneers and biggest P2P lending platforms, provides a good object of study and valuable insight into the inner workings of these platforms and their ability to adequately predict the risk of credit default. This last function turns Lending Club (and other analogous platforms) into more than just a mediator connecting supply and demand, assigning it the responsibility of curating and attesting borrower information and risk.

In the beginning of this work, we set out to study a dataset of loan observations from the Lending Club platform, containing information for loans issued between 2007 and 2018. With it we endeavored to further the current understanding and established knowledge of what are the determinants of loan default and what performance we can expect from a prediction model built atop this data.

In order to achieve this, we first conducted a cleaning and refinement of the data set, eliminating incorrect and missing data, where required. This process was followed by an exploratory analysis of the outputted data, to gain an overall understanding of the existing variables, their correlation and driving the selection of the final variables to be used for our study of determinants of default.

To further our analysis, we built a logistic regression model to analyze the dataset using the stepwise backward technique in order to arrive at a final, model-derived, set of independent variables that were the most significant for predicting loan default.

Our analysis aims at establishing a lifecycle of reviewing and understanding borrower default based on hard data readily available to the investor (or to P2P lending platforms). Unlike prior work, which focused on developing predictive models or identifying the determinants of default, our work lays out and end-to-end foundation for understanding the characteristics of borrowers loans within a platform (or set of loan observations), identifying determinants of default, building a predictive model and analyzing its efficacy in predicting loan default.

Our results seem to be consistent with existing literature (Carmichael, 2014; Emekter et al., 2015; Polena & Regner, 2018; Serrano-Cinca et al., 2015) and suggest that the variables which best explain (or predict) the borrowers likelihood of default are *grade*, *loan amount*, *interest rate*, *term*, *annual income*, *home ownership*, *debt-to-income ratio*, *FICO score*, *inquiries in last 6 months*, number of *open accounts* and *credit history*.

The most relevant independent variable is the *grade* of a borrower (as assigned by Lending Club) which, having the highest odds ratio, presents the best predictor of loan default. The higher the grade of a given loan (“A” being the highest grade and “G” the lowest, in alphabetical order), the higher the likelihood that it will be fully repaid.

Finally, although our model has shown a high accuracy rate of 0.8511, the AUC analysis produced a result of 0.6835, meaning that the overall performance of our model, as represented by its ability to distinguish between classes (i.e. borrowers that will default VS borrowers that will fully pay back a loan) is far from excellent and, at best, mediocre.

In summary, although our model has been able to effectively predict defaulted loans, its results have yielded a significant amount of "false negatives", i.e. fully paid loans that were predicted to default. We extrapolate from this that, although opportunities may be missed to fund good quality loans, it is best to err on the side of caution when dealing with unsecured consumer debt.

6.1 LIMITATIONS AND FUTURE WORK

Throughout this work, we identified several limitations and constraints that may have hindered our investigation and its subsequent results. We present those in this section and also attempt to lay out the foundation for future work to build upon the results provided herein.

During the initial analysis and clean-up/refinement of our dataset, we were forced to remove a substantial number of variables and loan observations due to missing data or incorrectly filled out forms. This approach, although consistent with established data analysis practices, may have caused the loss of potentially valuable data points for the study.

We hypothesize that the cause of this problem, with a special focus on incorrectly provided data, is likely tied to the fact that these platforms allow for the manual introduction of data by users. Such platforms also lack the manpower or capacity to handle and properly vet 100% of the data which they receive.

With regards to the results obtained with the developed logistic regression model, we estimate that the mediocre predictive performance, which we achieved, is a product of the dataset being imbalanced. This is a common issue in credit datasets as these will largely contain a disproportionate amount of non-defaulted loans (85% of non-defaulted loans in our dataset) and one that negatively impacts the predictive capacity of logistic regression models (Singh et al., 2014).

7 BIBLIOGRAPHY

Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent Systems in Accounting, Finance and Management*, *18*(2–3), 59–88.

Ala'raj, M., & Abbod, M. (2015). A systematic credit scoring model based on heterogeneous classifier ensembles. *2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA)*, 1–7.

Ashofteh, A. & Bravo, J. M. (2019). A non-parametric based computationally efficient approach for credit scoring. Atas da Conferencia da Associacao Portuguesa de Sistemas de Informacao 2019 [CAPSI 2019 - 19th Conference of the Portuguese Association for Information Systems, Proceedings. 4]. https://aisel.aisnet.org/capsi2019/4.

Ashofteh, A., & Bravo, J. M. (2021a). A Conservative Approach for Online Credit Scoring. Expert Systems With Applications, Volume 176, p. 1-16, 114835. https://doi.org/10.1016/j.eswa.2021.114835

Ashofteh, A. & Bravo, J. M. (2021b). Spark Code: A Novel Conservative Approach for Online Credit Scoring [Source Code]. https://doi.org/10.24433/CO.1963899.v1. Associated Publication: "A Conservative Approach for Online Credit Scoring", Expert Systems with Applications, https://doi.org/10.1016/j.eswa.2021.114835

Barth, C. (2012). Looking For 10% Yields? Go Online For Peer To Peer Lending. *Forbes*. Retrieved from https://www.forbes.com/sites/chrisbarth/2012/06/06/looking-for-10-yields-go-online-for-peer-to-peer-lending/#6396eafa3a8f

Berger, S. C., & Gleisner, F. (2009). Emergence of financial intermediaries in electronic markets: The case of online P2P lending. *BuR Business Research Journal*, *2*(1).

Boot, A. W. A., & Thakor, A. V. (1997). Financial system architecture. *The Review of Financial Studies*, *10*(3), 693–733.

Bradley, Christine; Burhouse, Susan; Gratton, Heather; Miller, R.-A. (2009). Alternative Financial Services: A Primer. *FDIC Quarterly*, *3*(1), 39–47. Retrieved from https://www.fdic.gov/bank/analytical/quarterly/2009-vol3-1/fdic140-quarterlyvol3no1-afs-final.pdf

Bravo, J. M. (2020). Longevity-Linked Life Annuities: A Bayesian Model Ensemble Pricing Approach. CAPSI 2020 Proceedings, 29. https://aisel.aisnet.org/capsi2020/29.

Bravo, J. M. (2021). Pricing participating longevity-linked life annuities: A Bayesian Model Ensemble approach. European Actuarial Journal. https://doi.org/10.1007/s13385-021-00279-w

Bravo, J. M., Ayuso, M. (2020). Mortality and life expectancy forecasts using bayesian model combinations: An application to the portuguese population. RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação, E40, 128–144. https://doi.org/10.17013/risti.40.128–145.

Bravo, J. M., Ayuso, M. (2021). Forecasting the retirement age: A Bayesian Model Ensemble Approach. Advances in Intelligent Systems and Computing, Volume 1365 AIST, 123–135 [2021 World Conference on Information Systems and Technologies, WorldCIST 2021] Springer, Cham. https://doi.org/10.1007/978-3-030-72657-7_12.

Bravo, J. M., Ayuso, M., Holzmann, R., Palmer, E. (2021). Addressing the Life Expectancy Gap in

Pension Policy. Insurance: Mathematics and Economics, 99, 200-221. https://doi.org/10.1016/j.insmatheco.2021.03.025.

Byanjankar, A. (2017). Predicting credit risk in Peer-to-Peer lending with survival analysis. *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1–8.

Byanjankar, A., Heikkilä, M., & Mezei, J. (2015). Predicting credit risk in peer-to-peer lending: A neural network approach. *2015 IEEE Symposium Series on Computational Intelligence*, 719–725.

Carmichael, D. (2014). Modeling default for peer-to-peer loans. *Available at SSRN 2529240*.

Chamboko, R., & Bravo, J. M. (2016). On the modelling of prognosis from delinquency to normal performance on retail consumer loans. Risk Management, 18(4), 264–287. https://doi.org/10.1057/s41283-016-0006-4

Chamboko, R., & Bravo, J. M. (2019a). Frailty correlated default on retail consumer loans in Zimbabwe. International Journal of Applied Decision Sciences, 12(3), 257–270. https://doi.org/10.1504/IJADS.2019.100436

Chamboko, R., & Bravo, J. M. V. (2019b). Modelling and forecasting recurrent recovery events on consumer loans. International Journal of Applied Decision Sciences, 12(3), 271–287. https://doi.org/10.1504/IJADS.2019.100440

Chamboko, R., & Bravo, J. M. (2020). A Multi-State Approach to Modelling Intermediate Events and Multiple Mortgage Loan Outcomes. Risks, 8(2), 64. https://doi.org/10.3390/risks8020064

Chorzempa, M. (2018). Massive P2P Failures in China: Underground Banks Going Under. *Peterson Institute for International Economics*, *21*. Retrieved from https://www.piie.com/blogs/china-economic-watch/massive-p2p-failures-china-underground-banks-going-under

Corporate Finance Institute. (n.d.). Peer-to-Peer Lending - Overview, How It Works, Pros & Cons. Retrieved from https://corporatefinanceinstitute.com/resources/knowledge/finance/peer-to-peer-lending/

Dirick, L., Claeskens, G., & Baesens, B. (2017). Time to default in credit scoring using survival analysis: a benchmark study. *Journal of the Operational Research Society*, *68*(6), 652–665.

Đurović, A. (2017). Estimating Probability of Default on Peer to Peer Market – Survival Analysis Approach. *Journal of Central Banking Theory and Practice*, *6*(2).

Emekter, R., Tu, Y., Jirasakuldech, B., & Lu, M. (2015). Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. *Applied Economics*, *47*(1), 54–70. https://doi.org/10.1080/00036846.2014.962222

FICO. (2018). *An evolution in ML innovations that helps both lenders and consumers*. Retrieved from https://www.fico.com/en/resource-download-file/6559

Galloway, I., & others. (2009). Peer-to-peer lending and community development finance. *Community Investments*, *21*(3), 19–23.

Gonzalez, L., & Loureiro, Y. K. (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.

Greenbaum, S. I., Thakor, A. V, & Boot, A. W. A. (2019). *Contemporary financial intermediation*. Academic Press.

Havrylchyk, O., Mariotto, C., Rahim, T.-U.-, & Verdier, M. (2016). What Drives the Expansion of the Peer-to-Peer Lending? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2841316

Havrylchyk, O., & Verdier, M. (2018). The financial intermediation role of the P2P lending platforms. *Comparative Economic Studies*, *60*(1), 115–130.

Herzenstein, M., Andrews, R. L., Dholakia, U. M., & Lyandres, E. (2008). The democratization of personal consumer loans? Determinants of success in online peer-to-peer lending communities. *Boston University School of Management Research Paper*, *14*(6), 1–36.

Herzenstein, M., Dholakia, U. M., & Andrews, R. L. (2011). Strategic herding behavior in peer-to-peer loan auctions. *Journal of Interactive Marketing*, *25*(1), 27–36.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (Vol. 398). John Wiley & Sons.

Huang, Z., Chen, H., Hsu, C.-J., Chen, W.-H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems*, *37*(4), 543–558.

Hulme, M. K., & Wright, C. (2006). Internet based social lending: Past, present and future. *Social Futures Observatory*, *11*, 1–115.

Iyer, R., Khwaja, A. I., Luttmer, E. F. P., & Shue, K. (2009). Screening in new credit markets: Can individual lenders infer borrower creditworthiness in peer-to-peer lending? *AFA 2011 Denver Meetings Paper*.

Jin, Y., & Zhu, Y. (2015). A data-driven approach to predict default risk of loan for online Peer-to-Peer (P2P) lending. *2015 Fifth International Conference on Communication Systems and Network Technologies*, 609–613.

Käfer, B. (2018). Peer-to-Peer Lending - A (Financial Stability) Risk Perspective. *Review of Economics*, *69*(1), 1–25.

Keys, B. J., Mukherjee, T., Seru, A., & Vig, V. (2010). Did securitization lead to lax screening? Evidence from subprime loans. *The Quarterly Journal of Economics*, *125*(1), 307–362.

Klafft, M. (2008). Online peer-to-peer lending: a lenders' perspective. *Proceedings of the International Conference on E-Learning, E-Business, Enterprise Information Systems, and E-Government, EEE*, 371–375.

Langager, C. (2019). How Is My Credit Score Calculated? Retrieved from Investopedia website: https://www.investopedia.com/ask/answers/05/creditscorecalculation.asp

Lee, E., & Lee, B. (2012). Herding behavior in online P2P lending: An empirical investigation. *Electronic Commerce Research and Applications*, *11*(5), 495–503.

Lending Club. (2019a). Demand and credit profile. Retrieved from Lending Club website: https://www.lendingclub.com/info/demand-and-credit-profile.action

Lending Club. (2019b). How are loans listed and approved for investing? – LendingClub. Retrieved from Lending Club website: https://help.lendingclub.com/hc/en-us/articles/215466748-How-are-loans-listed-and-approved-for-investing-

Lending Club. (2019c). Rates & Fees. Retrieved from Lending Club website: https://www.lendingclub.com/public/rates-and-fees.action

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

Lin, M., Prabhala, N. R., & Viswanathan, S. (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.

Lust, D. (2017). *Analysis of scoring in peer-to-peer lending*.

Mezei, J., Byanjankar, A., & Heikkilä, M. (2018). *Credit risk evaluation in peer-to-peer lending with linguistic data transformation and supervised learning*.

Mild, A., Waitz, M., & Wöckl, J. (2015). How low can you go? Overcoming the inability of lenders to set proper interest rates on unsecured peer-to-peer lending markets. *Journal of Business Research*, *68*(6), 1291–1305.

Möllenkamp, N. (2017). *Determinants of Loan Performance in P2P Lending*. University of Twente.

Nowak, A., Ross, A., & Yencha, C. (2018). Small Business Borrowing and Peer-to-Peer Lending: Evidence From Lending Club. *Contemporary Economic Policy*, *36*(2), 318–336. https://doi.org/10.1111/COEP.12252

Olson, D. L., Delen, D., & Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems*, *52*(2), 464–473.

Polena, M., & Regner, T. (2018). Determinants of borrowers' default in P2P lending under consideration of the loan risk class. *Games*, *9*(4), 82.

Prosper.com. (2016). What is the loan review process? How long does it take? – Help is on the way. Retrieved from Prosper.com website: https://prosper.zendesk.com/hc/en-us/articles/210013753-What-is-the-loan-review-process-How-long-does-it-take-

Securities and Exchange Commission. (2015). *LendingClub Corp Form 10-K for Fiscal Year Ended December 31, 2014*. Retrieved from https://www.sec.gov/Archives/edgar/data/1409970/000119312515070385/d851207d10k.htm

Serrano-Cinca, C., Gutierrez-Nieto, B., & López-Palacios, L. (2015). Determinants of default in P2P lending. *PloS One*, *10*(10), e0139427.

Singh, S., Tsai, K., & Ramiah, S. (2014). *Peer Lending Risk Predictor*. https://doi.org/10.13140/2.1.4810.6567

Tan, F., Hou, X., Zhang, J., Wei, Z., Yan, Z., & Weng, S.-C. (2018). A novel risk assessment scheme and practice for peer-to-peer lending. *Proc. ACM SIGKDD Workshop Data Sci. Fintech*.

Tapscott, D., & Williams, A. (2007). The new science of sharing. *The BusinessWeek Wikinomics Series*.

Transparency Market Research. (2016). *Peer-to-Peer Lending Market - Global Industry Analysis, Size, Share, Growth, Trends and Forecast 2016 - 2024*. Retrieved from https://www.transparencymarketresearch.com/peer-to-peer-lending-market.html

Weiss, G. N. F., Pelger, K., & Horsch, A. (2010). Mitigating adverse selection in P2P lending - Empirical evidence from prosper. com. *Available at SSRN 1650774*.

Wendler, T., & Gröttrup, S. (2016). *Data mining with SPSS modeler: theory, exercises and solutions*.

https://doi.org/10.1007/978-3-319-28709-6

Yum, H., Lee, B., & Chae, M. (2012). From the wisdom of crowds to my own judgment in microfinance through online peer-to-peer lending platforms. *Electronic Commerce Research and Applications*, *11*(5), 469–483.

Zhao, H., Ge, Y., Liu, Q., Wang, G., Chen, E., & Zhang, H. (2017). P2P lending survey: platforms, recent advances and prospects. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *8*(6), 72.

Zhao, D., Huang, C., Wei, Y., Yu, F., Wang, M., & Chen, H. (2017). An effective computational model for bankruptcy prediction using kernel extreme learning machine approach. Computational Economics, 49(2), 325–341. https://doi.org/10.1007/s10614-016-9562-7

8 APPENDIX

8.1 APPENDIX A

Table 9 - All Lending Club variables with description

Variables	Description
acc_now_delinq	The number of accounts on which the borrower is now delinquent.
acc_open_past_24 mths	Number of trades opened in past 24 months.
addr_state	The state provided by the borrower in the loan application.
all_util	Balance to credit limit on all trades.
annual_inc	The self-reported annual income provided by the borrower during registration.
annual_inc	The combined self-reported annual income provided by the co-borrowers during registration
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers.
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
Collection _ recovery _fee	Post charge off collection fee
collections_12 _mths_ex_med	Number of collections in 12 months excluding medical collections
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
delinq_amnt	The past-due amount owed for the accounts on which the borrower is now delinquent.
desc	Loan description provided by the borrower
dti	A ratio calculated using the borrower's total monthly debt payments on the total de bt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
dti_joint	A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested LC loan, divided by the co-borrowers' combined self-reported monthly income.
earliest_cr_line	The month the borrower's earliest reported credit line was opened
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means l ess than one year and 10 means ten or more years.
emp_title	The job title supplied by the Borrower when applying for the loan. (Employer Title replaces Employer Name for all loans listed after 9/23/2013)
fico_range_high	The upper boundary range the borrower's FICO at loan origination belongs to.
fico_range_low	The lower boundary range the borrower's FICO at loan origination belongs to.
funded_amnt	The total amount committed to that loan at that point in time.

Variables	Description
funded_amnt_inv	The total amount committed by investors for that loan at that point in time.
grade	LC assigned loan grade.
home_ownership	The homeownership status provided by the borrower during registration or obtaine d from the credit report. Our values are: ANY, RENT, OWN, MORTGAGE, OTHER.
id	A unique LC assigned ID for the loan listing.
il_util	Ratio of total current balance to high credit/credit limit on all install acct
initial_list_status	The initial listing status of the loan. Possible values are – W, F (w-whole; f-fractional)
inq_fi	Number of personal finance inquiries.
inq_last_12m	Number of credit inquiries in past 12 months.
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries).
installment	The monthly payment owed by the borrower if the loan originates.
int_rate	Interest Rate on the loan.
issue_d	The month which the loan was funded.
last_credit_pull_d	The most recent month LC pulled credit for this loan.
last_fico_range _high	The upper boundary range the borrower’s last FICO pulled belongs to.
last_fico_range _low	The lower boundary range the borrower’s last FICO pulled belongs to.
last_pymnt_amnt	Last total payment amount received.
last_pymnt_d	Last month payment was received.
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, t he credit department reduces the loan amount, then it will be reflected in this value.
loan_status	Current status of the loan.
max_bal_bc	Maximum current balance owed on all revolving accounts.
member_id	A unique LC assigned Id for the borrower member.
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_deli nq	The number of months since the borrower's last delinquency.
mths_since_last_maj or_derog	Months since most recent 90-day or worse rating.
mths_since_last_reco rd	The number of months since the last public record.
mths_since_rcnt _il	Months since most recent installment accounts opened
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.

Variables	Description
next_pymnt_d	Next scheduled payment date
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.
open_acc_6m	Number of open trades in last 6 months
open_il_12m	Number of installment accounts opened in past 12 months
open_il_24m	Number of installment accounts opened in past 24 months
open_act_il	Number of currently active installment trades
open_rv_12m	Number of revolving trades opened in past 12 months
open_rv_24m	Number of revolving trades opened in past 24 months
out_prncp	Remaining outstanding principal for total amount funded
out_prncp_inv	Remaining outstanding principal for portion of total amount funded by investors
pct_tl_nvr_dlq	Percent of trades never delinquent
percent_bc_gt_75	Percentage of all bankcard accounts > 75% of limit.
policy_code	Publicly available policy_code=1 new products not publicly available policy_code=2
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.
pymnt_plan	Indicates if a payment plan has been put in place for the loan
recoveries	post charge off gross recovery
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.
sub_grade	LC assigned loan subgrade
tax_liens	Number of tax liens

Variables	Description
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
title	The loan title provided by the borrower
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 currently in the borrower's credit file
total_bal_ex_mort	Total credit balance excluding mortgage
total_bal_il	Total current balance of all installment accounts
total_bc_limit	Total bankcard high credit/credit limit
total_cu_tl	Number of finance trades
total_il_high_credit _limit	Total installment high credit/credit limit
total_pymnt	Payments received to date for total amount funded
total_pymnt_inv	Payments received to date for portion of total amount funded by investors
total_rec_int	Interest received to date
total_rec_late_fee	Late fees received to date
total_rec_prncp	Principal received to date
total_rev_hi_lim	Total revolving high credit/credit limit
url	URL for the LC page with listing data.
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verifi ed
verified_status_ joint	Indicates if the co-borrowers' joint income was verified by LC, not verified, or if the income source was verified
zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application .
revol_bal_joint	Sum of revolving credit balance of the co-borrowers, net of duplicate balances.
sec_app_fico _range_low	FICO range (high) for the secondary applicant.
sec_app_fico _range_high	FICO range (low) for the secondary applicant.
sec_app_earliest _cr_line	Earliest credit line at time of application for the secondary applicant.
sec_app_inq_last _6mths	Credit inquiries in the last 6 months at time of application for the secondary applica nt.
sec_app_mort_acc	Number of mortgage accounts at time of application for the secondary applicant.
sec_app_open_acc	Number of open trades at time of application for the secondary applicant.
sec_app_revol_util	Ratio of total current balance to high credit/credit limit for all revolving accounts.
sec_app_open _act_il	Number of currently active installment trades at time of application for the seconda ry applicant.
sec_app_num _rev_accts	Number of revolving accounts at time of application for the secondary applicant.
sec_app_chargeoff _within_12 _mths	Number of charge-offs within last 12 months at the time of application for the secondary applicant.

Variables	Description
sec_app_collections _12_mths _ex_med	Number of collections within last 12 months excluding medical collections at time of application for the secondary applicant
sec_app_mths_since _last_major _derog	Months since most recent 90- day or worse rating at time of application for the secondary applicant
hardship_flag	Flags whether or not the borrower is on a hardship plan
hardship_type	Describes the hardship plan offering
hardship_reason	Describes the reason the hardship plan was offered
hardship_status	Describes if the hardship plan is active, pending, canceled, completed, or broken
deferral_term	Amount of months that the borrower is expected to pay less than the contractual m onthly payment amount due to a hardship plan
hardship_amount	The interest payment that the borrower has committed to make each month while t hey are on a hardship plan
hardship_start_date	The start date of the hardship plan period
hardship_end_date	The end date of the hardship plan period
payment_plan _start_date	The day the first hardship plan payment is due. For example, if a borrower has a har dship plan period of 3 months, the start date is the start of the three- month period in which the borrower is allowed to make interest-only payments.
hardship_length	The number of months the borrower will make smaller payments than normally obli gated due to a hardship plan
hardship_dpd	Account days past due as of the hardship plan start date
hardship_loan_status	Loan Status as of the hardship plan start date
orig_projected_additi onal_accrued _interest	The original projected additional interest amount that will accrue for the given hard ship payment plan as of the Hardship Start Date. This field will be null if the borrower has broken their hardship payment plan.
hardship_payoff_ balance_amount	The payoff balance amount as of the hardship plan start date
hardship_last_payme nt_amount	The last payment amount as of the hardship plan start date
disbursement_ method	The method by which the borrower receives their loan. Possible values are: CASH, DIRECT_PAY
debt_settlement_ flag	Flags whether or not the borrower, who has charged-off, is working with a debt-settlement company.
debt_settlement_ flag_date	The most recent date that the Debt_Settlement_Flag has been set.
settlement_status	The status of the borrower's settlement plan. Possible values are: COMPLETE, ACTIVE, BROKEN, CANCELLED, DENIED, DRAFT
settlement_date	The date that the borrower agrees to the settlement plan.
settlement_amount	The loan amount that the borrower has agreed to settle for.
settlement_percenta ge	The settlement amount as a percentage of the payoff balance amount on the loan
settlement_term	The number of months that the borrower will be on the settlement plan.

Source: Author's production.

8.2 APPENDIX B

Table 10 - Correlation matrix of LC variables

	loan_amnt	Int_rate	installment	annual_inc	dti	delinq_2yrs	fico_range_low	inq_last_6mths	open_acc	pub_rec	revol_bal	revol_util	chargeoff_within_12_mths	pub_rec_bankruptcies	Tax_liens	writing_skills	credit_history
loan_amnt	1	0.0356	**0.9777**	0.4469	0.0031	-0.0076	0.1316	-0.0177	0.1858	-0.0858	0.3339	0.109	-0.0004	-0.1211	0.0041	0.0401	0.1721
Int_rate	0.0356	1	0.0605	-0.1261	0.1352	0.0525	-0.4547	0.2584	-0.0494	0.0473	-0.062	0.2873	0.0111	0.0561	0.0072	0.1049	-0.1445
installment	**0.9777**	0.0605	1	0.4383	0.0165	0.0036	0.0793	-0.0039	0.1868	-0.0721	0.3262	0.1264	0.0027	-0.11	0.011	0.0061	0.1597
annual_inc	0.4469	-0.1261	0.4383	1	-0.2211	0.0614	0.1127	0.0477	0.1871	-0.0118	0.3904	0.05	0.0154	-0.0709	0.0517	-0.0097	0.201
dti	0.0031	0.1352	0.0165	-0.2211	1	-0.0016	-0.0934	-0.0075	0.2939	-0.0403	0.128	0.1749	-0.0029	-0.0298	-0.0242	-0.0704	0.0486
delinq_2yrs	-0.0076	0.0525	0.0036	0.0614	-0.0016	1	-0.1765	0.0238	0.0588	-0.0094	-0.0286	-0.0163	0.1388	-0.0355	0.014	-0.0478	0.0929
fico_range_low	0.1316	-0.4547	0.0793	0.1127	-0.0934	-0.1765	1	-0.0815	0.0212	-0.1915	0.0526	-0.4493	-0.0517	-0.2077	-0.0558	0.0865	0.1023
inq_last_6mths	-0.0177	0.2584	-0.0039	0.0477	-0.0075	0.0238	-0.0815	1	0.1161	0.0546	-0.0124	-0.0786	0.0105	0.0708	0.0056	0.0738	-0.0079
open_acc	0.1858	-0.0494	0.1868	0.1871	0.2939	0.0588	0.0212	0.1161	1	-0.0251	0.2233	-0.1438	0.0073	-0.0373	-0.002	-0.0462	0.1497
pub_rec	-0.0858	0.0473	-0.0721	-0.0118	-0.0403	-0.0094	-0.1915	0.0546	-0.0251	1	-0.0939	-0.0727	-0.0002	**0.6351**	**0.7131**	-0.0705	0.0855
revol_bal	0.3339	-0.062	0.3262	0.3904	0.128	-0.0286	0.0526	-0.0124	0.2233	-0.0939	1	0.2089	-0.01	-0.1203	-0.0119	0.0016	0.2096
revol_util	0.109	0.2873	0.1264	0.05	0.1749	-0.0163	-0.4493	-0.0786	-0.1438	-0.0727	0.2089	1	-0.0153	-0.086	-0.0149	0.0338	-0.004
chargeoff_within_12_mths	-0.0004	0.0111	0.0027	0.0154	-0.0029	0.1388	-0.0517	0.0105	0.0073	-0.0002	-0.01	-0.0153	1	-0.0071	-0.0005	-0.0173	0.036
pub_rec_bankruptcies	-0.1211	0.0561	-0.11	-0.0709	-0.0298	-0.0355	-0.2077	0.0708	-0.0373	**0.6351**	-0.1203	-0.086	-0.0071	1	0.0384	-0.0534	0.0778
tax_liens	0.0041	0.0072	0.011	0.0517	-0.0242	0.014	-0.0558	0.0056	-0.002	**0.7131**	-0.0119	-0.0149	-0.0005	0.0384	1	-0.0374	0.0427
writing_skills	0.0401	0.1049	0.0061	-0.0097	-0.0704	-0.0478	0.0865	0.0738	-0.0462	-0.0705	0.0016	0.0338	-0.0173	-0.0534	-0.0374	1	-0.0498
credit_history	0.1721	-0.1445	0.1597	0.201	0.0486	0.0929	0.1023	-0.0079	0.1497	0.0855	0.2096	-0.004	0.036	0.0778	0.0427	-0.0498	1

Source: Author's production.

8.3 APPENDIX C

Table 11 - Descriptive Statistics of Numerical Variables

Variables	Mean	Std.Dev	Min	Median	Max	Average by loan status	
						Fully Paid	Default
Annual Income	73,250	49,039	3,000	62,000	1,000,000	74,563	65,787
Charged Off within 12 months	0.01	0.11	0	0	10	0.01	0.01
Credit History	16.12	7.56	1	15	71	16.25	15.40
Delinquency past 2 years	0.31	0.86	0	0	30	0.31	0.34
Debt-to-Income	17.64	8.20	0	17.15	41.21	17.36	19.20
FICO	695	31	640	690	845	697	687
Inquiries last 6 months	0.70	1	0	0	17	0.67	0.88
Interest rate	12.42	4.17	5.32	12.29	28.99	12.06	14.47
Loan Amount	13,202	8,090	500	11,000	35,000	13,172	13,372
Open Accounts	11.31	5.22	1	10	84	11.28	11.48
Public Records	0.20	0.59	0	0	86	0.19	0.23
Public Record Bankruptcies	0.12	0.36	0	0	12	0.12	0.14
Revolving Balance	16,032	22,095	0	11,053	2,904,836	16,289	14,573
Revolving line utilization rate	54.07	23.95	0	54.80	892.30	53.56	56.98
Tax Liens	0.05	0.40	0	0	85.00	0.05	0.05
Writing Skills	1.51	3.39	0	0	10	1.50	1.53

Source: Author's production.

Table 12 - Descriptive Statistics of Categorical Variables

Variables	# issued loans (%)		Default rate (%)
Employment Length			
0 – 4 years	277,107	(37.8%)	15.0%
5 – 9 years	182,924	(25.0%	15.0%
10+ years	231,027	(31.5%)	13.8%
Missing	41,482	(5.7%)	21.0%
Grade			
A	156,599	(21.4%)	5.5%
B	238,806	(32.6%)	11.4%
C	195,344	(26.7%)	18.5%
D	94,012	(12.8%)	24.2%
E	34,607	(4.7%)	29.8%
F	10,974	(1.5%)	34.8%
G	2,198	(0.3%)	37.9%
Home Ownership			
Mortgage	350,431	(47.8%)	12.8%
Own	73,809	(10.1%)	15.5%
Rent	308,300	(42.1%)	17.3%
Purpose			
Car	8,371	(1.1%)	11.9%
Credit card	173,984	(23.8%)	12.4%
Debt consolidation	418,877	(57.2%)	15.8%
Educational	361	(0.05%)	20.2%
Home improvement	42,686	(5.8%)	13.2%
House	3,235	(0.4%)	18.9%
Major purchase	15,314	(2.1%)	13.3%
Medical	7,999	(1.1%)	17.5%
Moving	5,195	(0.7%)	19.5%
Other	39,582	(5.4%)	17.1%
Renewable energy	554	(0.1%)	19.7%
Small business	9,317	(1.3%)	24.4%
Vacation	4,748	(0.6%)	16.1%
Wedding	2,317	(0.3%)	12.3%
Term			
36 months	668,970	(91.3%)	14.0%
60 months	63,570	(8.7%)	25.3%
Verification Status			
Source Verified	255,845	(34.9%)	15.1%
Verified	242,057	(33.0%)	18.0%
Not Verified	234,638	(32.0%)	11.7%

Source: Author's production.