12-2019

Credit Risk Estimation in the Age of Peer to Peer Lending

Aboubacar Coulibaly
acoulibaly@stcloudstate.edu

Credit Risk Estimation in the Age of Peer to Peer Lending

by

Aboubacar Coulibaly

A Thesis

Submitted to the Graduate Faculty of

St. Cloud State University

in Partial Fulfillment of the Requirements

for the Degree of

Master of Science

in Applied Economics

December, 2019

Thesis Committee:
Manamperi Nimantha, Chairperson
Ratha Arthatrana
Lynn A. Collen

Abstract

Credit Risk in Peer to Peer Lending is an emerging field with practical implications for U.S banking system. Peer to Peer Lending is a type of online lending process which uses nontraditional bank channels. The inexorable rise of Fintechs has led to an extraordinary change in financial intermediation. This paper examines the factors that are critical in predicting default in Peer to Peer lending. The paper finds that FICO score, debt-to-income ratio , the loan amount, the credit grade assigned by the online lending platform are all critical factors of the credit risk evaluation process. Furthermore, models with hyperparameters such as neural networks and random forest do not reliably outperform classical logistic regression in the prediction of credit default. Finally, this paper makes vital policy recommendations to strengthen the efficiency of marketplace lending and provides a set of rules to prevent another crisis of the magnitude of the great recession.

Keywords: Peer-to-Peer Lending, Credit Risk, FICO, P2P Loan Default, Consumer Credit, Logistic Regression.

JEL Codes: C45, C51, E44, G21, G28, G50.

Acknowledgments

I would like to thank my parents for always encouraging me in my studies and never sparring any expense to help me achieve my potential. I would also like to thank my committee members for their guidance and inspiration at such a critical time of my academic development. Finally, my utmost gratitude to all my teachers from my home country Mali, New York and St. Cloud. And family members that always believed in me and encouraged me to give the best of myself and never give up.

Table of Contents

Page

List of Tables

Table Page

List of Figures

Figure Page

Chapter 1: Introduction

This paper estimates credit risk default in the peer to peer lending market space in the U.S. The Peer to Peer (P2P) lending is the process of lending online to individuals and businesses outside of the traditional banking networks. P2P Lending or marketplace lending seems inexorably poised to grab a sizable portion of the market share in credit lending, as much as $90 billion in originations by 2020 (Treasury, 2016). The largest players in the U.S market include platforms such as Lending Club, Prosper, Peerform, Upstarts and Funding Circle. These platforms intend to displace traditional banking through financial disintermediation and technology. Essentially matching investors directly to borrowers and doing so more efficiently than banks. Given their lower legacy costs and efficiency gains from technological innovations, these platforms can offer borrowers lower interest rates than traditional bank (Milne & Parboteeah, 2016). Additionally, investors receive a higher yield and diversify their portfolio by owning an asset class historically earmarked for banks.

However, P2P platforms possess less expertise in consumer lending, especially in assessing credit risk relative to banks. Credit risk in banking is defined as the likelihood that a borrower or counterparty fails to meet his/her contractual agreements to the lender. Moreover, asymmetric information issues are ubiquitous in P2P lending. Historically banks have minimized credit risk by acquiring a large amount of data on borrowers and putting a premium on relational rather than transactional lending (Calebe, Loriana, & Paolo, 2016). The P2P marketplace inherently lacks such historical data. Therefore, investors are more prone to credit risk due to information asymmetry.

This paper attempts to build predictive models which could be used by an investor to estimate credit default risk. To achieve this objective, data from a well-known platform Lending Club (LC) is utilized. The basic rationale for the study is that there is a need of investors to properly gauge the credit risk of a potential borrower hence a model could assist in the loan selection process.

Although there are many factors that could trigger default on a loan, this paper investigates factors provided by the platforms such as: the borrower's credit score, debt to income ratio (DTI), revolving debt utilization rate, annual income, length of credit history, public records, level of education.

This paper is organized in the following manner: Chapter II presents the literature review on credit risk in P2P marketplace. Chapter III describes the data and its origin.   Chapter IV presents the methodology in evaluating risk including logistic regression, neural networks and random forest. Chapter V presents the empirical results and finally chapter VI concludes with our findings and key policy recommendations.

Chapter 2: Literature Review

2.1 Genesis of P2P Lending

P2P Lending also known as marketplace lending stems from the development of the internet and the rise of e-commerce. The vast opportunities provided by the web and big data coupled with investors need for a higher yield have contributed to the birth of the industry. The first platform to see the day in the U.S was Prosper marketplace which was set up in 2006 to act as marketplace were investors and borrowers could transact as an “e-bay” for loans (Milne & Parboteeah, 2016). The marketplace lending quickly established itself as an alternative to bank loans and credit cards. Platforms like Lending Club and Funding Circle emerged to capture market share in the lucrative consumer finance space (Schumpeter, 2013). Several factors have contributed to the success of P2P Lending, (1) a lagging regulatory regime, (2) the opportunity for investors to earn a higher yield in a period of depressingly low interest rates, (3) the provision of credit in underserved segments, and (4) technological agility relative to the big banks (Milne & Parboteeah, 2016). However, as the industry matures small investors have been crowded out of the space by large institutional investors such as banks, hedge funds and asset managers (Demyanyk, Loutskina, & Kolliner, 2017). Furthermore, many policymakers and analysts have questioned the benefits of the industry to the average borrower. Many believe that P2P loans are predatory in nature and provide negligible funds to underserved areas and the underbanked population (Demyanyk et l., 2017). Moreover, some analysts fear the peer to peer industry exacerbates the level of financial distress by charging exorbitant rates. Additionally, sceptics claim an over supplying of credit to the financially distressed and minority segments of the consumer market with lower levels of education and higher levels of debt (Demyanyk et al., 2017). An increase in the supply of loans ceteris paribus will lead to a decrease in the prevailing market interest rate and increase access overall. However, given the lack of financial education in the general society marketplace lending may facilitate access to excessive and untenable amounts of debt. As attested by the Great Recession, borrowers and lenders alike can make suboptimal decisions leading to financial stress and higher levels of debt (Dore & Mach, 2019). To make matters worse, a study by Demyanyk et al. (2017) found that the debt level for P2P borrowers actually increased and credit scores decreased ex post receiving a loan. This is especially concerning given that marketplace lenders have always touted the product as a channel to pay down high interest credit card debt through consolidation loans (Balyuk, 2019) . According to LC, 66% of borrowers stated paying down credit card debt and debt consolidation in general as the intended use of funds, [see Figures 1 and 2], (Club, 2019) .

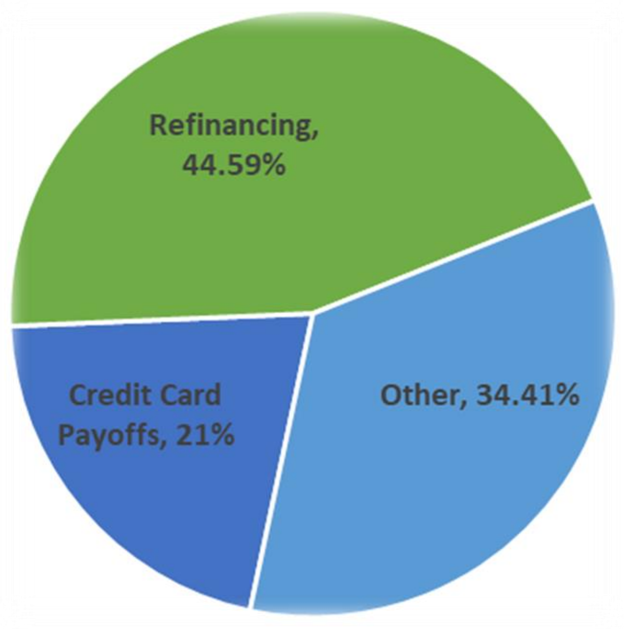


Figure 1: Reported use of funds (source: https://lendingclub.com).

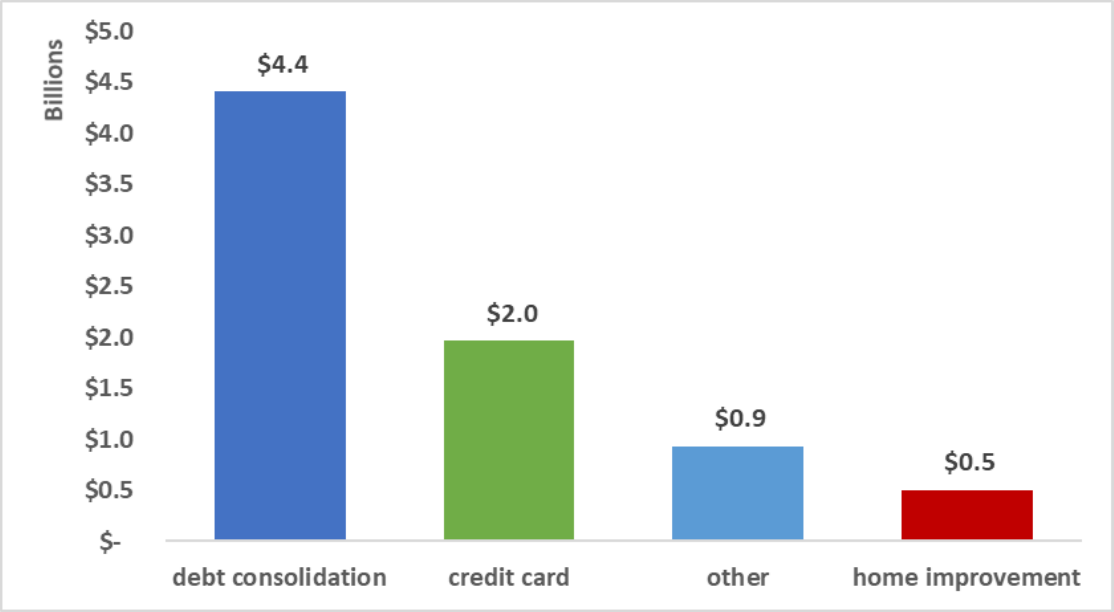


Figure 2: Balances by loan purpose (2018 originations).

Figure 2 shows that in 2018 almost 8 billion US dollars were borrowed on LC. The largest portion of the funds was borrowed for debt consolidation, followed by credit cards debt, “Other” includes things such as vacations, weddings and medical. Another half a billion went to home improvement purpose as stated by obligors. Debt consolidation and credit card accounted for $6.4 out of the $8 billion or 82% of the funds in 2018. This shows that debt consolidation is by far the most commonly stated usage of borrowed funds.

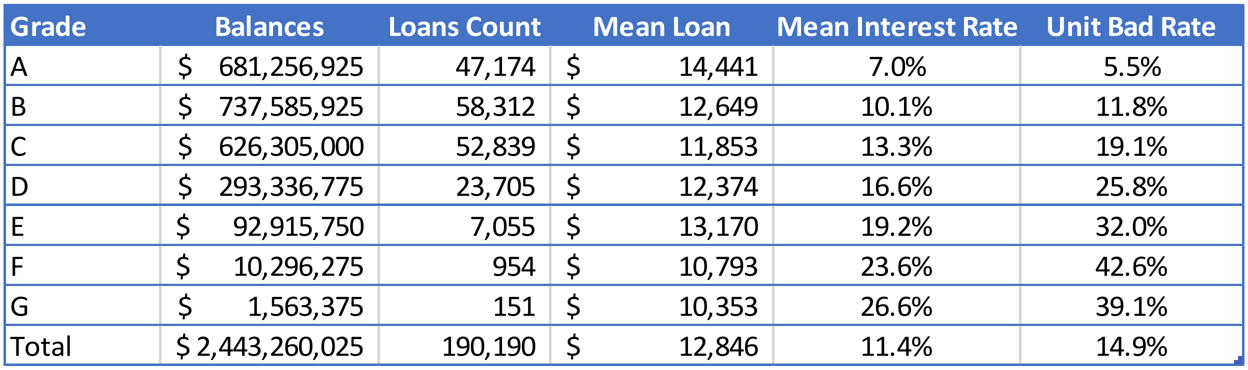
2.2 Credit Risk in P2P

The literature related to credit risk in P2P is nascent, however an enduring theme of said literature is information asymmetry. Information asymmetry as the name suggests is a situation in which one party to a transaction has significantly more information relative to the second party which can lead to market distortions. Investors in P2P industry face a large amount of information asymmetry (Serrano-Cinca, Gutiérrez-Nieto, & López-Palacios, 2015). Borrowers typically have a very good idea of their ability to repay based on private information such earning potentials as well as any outstanding obligations while lenders do not. Using data from Germany Calebe, Loriana, and Paolo in 2016 found that P2P platforms were catering to a riskier segment of borrowers and supplying them with microloans. Moreover, they found that the return on P2P lending after adjusting for the risk was lower relative to banks. Notably Hertzberg, Liberman, and Paravisini in 2019 found that riskier borrowers in the marketplace tend to self-select into longer term loans which are costlier in the long run and thus signal the borrower's likelihood of default. Zou, Huixin, and Zheng in 2017 found that lenders in the marketplace face a higher credit risk than traditional banks due to adverse selection. Using data from "PPDAI" a China based P2P lender, they found that loans with less stellar credit ratings and longer terms charge off at a much higher rate. They also found that the order of credit rating and borrowers with high credit lines tend to perform well. In another study the determinants of credit risk were found to be the existing amount of debt of the borrower, the credit grade assigned by the platform, loan purpose as well as housing situation (Serrano-Cinca et al., 2015). Furthermore, Cornaggia, Wolfe, and Yoo in 2018 found that P2P lenders can create adverse selection issues in the consumer credit space by cherry-picking high quality loans and leaving banks with lower quality loans to originate. Sometimes even creating a race to the bottom by causing banks to start loosening credit quality standards to sustain market share or stamp out competition. Emekter, Tu, Jirasakuldech, and Lu in their 2015 study of lending club data using logistic regression found that the key determinants of P2P credit risk include a borrower's credit grade, debt to income ratio (DTI) , their FICO score and the revolving line utilization rate. Neural network models have also been implemented to gauge credit risk in P2P settings. A study by Byanjankar, Heikkila, and Mezei in 2017 found that it was indeed possible to assess credit risk in P2P and they found that financial factors rather than demographic ones were more predictive of loan outcomes. Most studies have focused on identifying some determinants of credit risk and possible ways to improve the P2P market. This study contributes to the current literature in three ways: (1) it identifies key credit risk predictors in P2P lending market, (2) uses three quantitative methods for assessing default risk and (3) provides a predictive model that estimates ex ante credit risk at the loan level using available data provided by lending club.

Chapter 3: Data and Origins

The data for this analysis comes from the Lending Club (LC). It includes individual level data for 36 months maturity loans made by LC in 2015 for over 5.8 billion dollars. The data includes loan level credit attributes such as income, debt to income ratio, the credit score and the status of the loan at the end of Q2:2019. There are 190,190 observations and 145 variables in the dataset. The data set was split in training and testing sets with 60% of observations in training and 40% in the testing dataset .Default in the analysis is defined as any account that goes to charge-off. The variables such as FICO, the credit grade and interest rate have shown a high information value. The interest rate factor showed a suspiciously high information value pointing to endogeneity. Accounts with high interest rate will charge off at a higher rate because they are inherently riskier but also because the interest rate will compound the cost of repayment and increase the likelihood of a bad performance.

Table 1: A snapshot of the data set.



Grade	Balances	Loans Count	Mean Loan	Mean Interest Rate	Unit Bad Rate
A	$ 681,256,925	47,174	$ 14,441	7.0%	5.5%
B	$ 737,585,925	58,312	$ 12,649	10.1%	11.8%
C	$ 626,305,000	52,839	$ 11,853	13.3%	19.1%
D	$ 293,336,775	23,705	$ 12,374	16.6%	25.8%
E	$ 92,915,750	7,055	$ 13,170	19.2%	32.0%
F	$ 10,296,275	954	$ 10,793	23.6%	42.6%
G	$ 1,563,375	151	$ 10,353	26.6%	39.1%
Total	$ 2,443,260,025	190,190	$ 12,846	11.4%	14.9%

Table 1 shows a snapshot of 2015 originations with credit grade shown in descending order. The credit grade is an assessment of loan quality assigned by the LC platform and is highly predictive of loan performance. Credit grade A is the lowest rated risk and G is the highest rate risk loan. Most of the lending is taking place in credit grades A through C. The default rate increases sharply has one moves from credit A to G. Finally, the interest rate charged by LC increases in tandem with the credit grade and the probability of default reflecting risk reward dynamics.

Chapter 4: Methodology

This chapter examines the variable selection process and methods for estimating credit default using three well known classification methods: Logistic Regression, Random Forest and Neural Networks.

4.1 Variable Selection Process

An important endeavor in any scientific study is identifying factors that influence the phenomena one is trying to explain, it naturally follows that great attention must be given to identification and selection of pertinent factors. The process for selecting variables and subsequently modeling them was based on three pillars: (1) identifying the 10 best predictors with the highest information value, (2) identifying the top 10 variables that showed importance in random forest decision trees, (3) then the top 10 variables for methods (1) and (2) are used as inputs to model the probability of defaulting via logistic regression, random forest and neural networks.

Information value based variable selection techniques. These are well known in credit risk analysis and have been extensively practiced in building credit risk models, (Deloitte, 2016). The information value (IV) and a closely related concept, weight of evidence (WOE) rank variables based on their ability to separate good vs bad loans. The formula for information value is the following: $IV = \sum_{i=1}^{z}(Dist\ Good_i - Dist\ Bad_i) *\ WOE_i$ (1)

, where z is the number of bins for the factor and $WOE = \ln\left(\frac{Dist\ Good}{Dist\ Bad}\right)$. The next table illustrates the interpretation of the values of IV for a potential predictor.

Table 2: Interpretation of information value (Baesens, Daniel, & Scheule, 2016).

Information Value	Interpretation
Less than 0.02	The variable is not predictive
0.02 to 0.1	the variable is weakly predictive
0.1 to 0.3	the variable is moderately predictive
More than 0.3	the variable is strongly predictive

The next three figures and three tables show exploratory analyses that were performed on all variables to identify the most predictive factors of credit default. Factors with information values above 0.02 were further utilized in the model building process (Baesens et al., 2016). For instance, the credit grade assigned by LC is highly predictive given an information value of 0.38 (Figure 3). The default rate increases monotonically from grade A to grades D, E, F, and G. Grade A has 5.5% default rate overall and accounts for 24.8% of the data or 47,174 loans. At the other end of the spectrum, credit grades D, E, F, and G have a default rate of 27.7% and account for 16.8% of the data or 31,865 loans. Grade is clearly an important predictor in assessing credit risk for LC loans. The conditional distribution of default changes drastically from one grade to the next (Figure 4 and Table 4). Another variable that showed some predictive ability was the FICO score at the time of the credit decision (Figure 4 and Table 4). Borrowers with a FICO score below 685 had a default rate of 18.9% compared to only 6.6% for those with 730 or more FICO scores. The information value of 0.14 makes FICO an important factor in the analysis as well. The debt to income (DTI) was found to be somewhat predictive as well with an IV of 0.06. There is a clear correlation between the amount of debt the borrower has and his/her future performance on the loan (Figure 5 and Table 5).

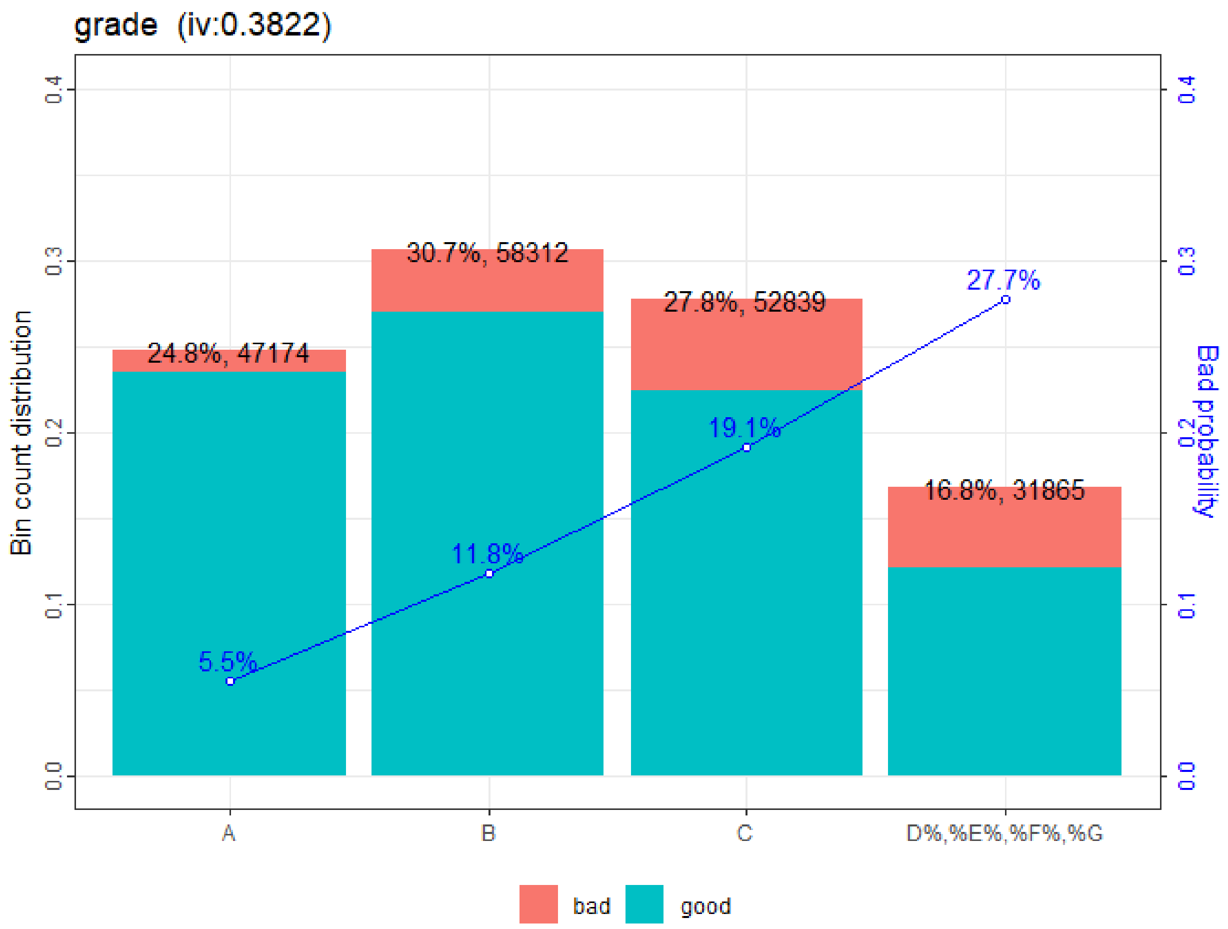
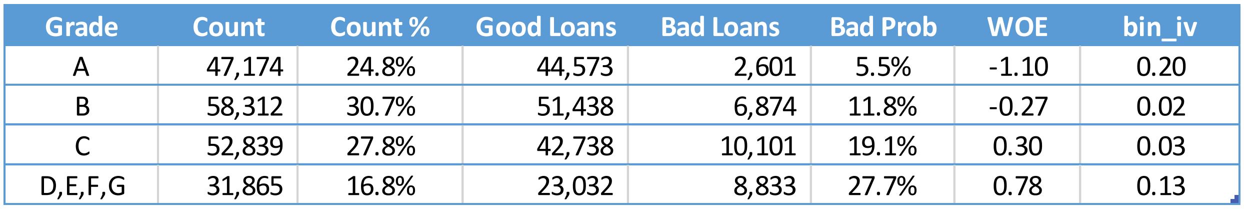


Figure 3: Default rate by credit grade.

Table 3: Information value for grade.



Grade	Count	Count %	Good Loans	Bad Loans	Bad Prob	WOE	bin_iv
A	47,174	24.8%	44,573	2,601	5.5%	-1.10	0.20
B	58,312	30.7%	51,438	6,874	11.8%	-0.27	0.02
C	52,839	27.8%	42,738	10,101	19.1%	0.30	0.03
D,E,F,G	31,865	16.8%	23,032	8,833	27.7%	0.78	0.13

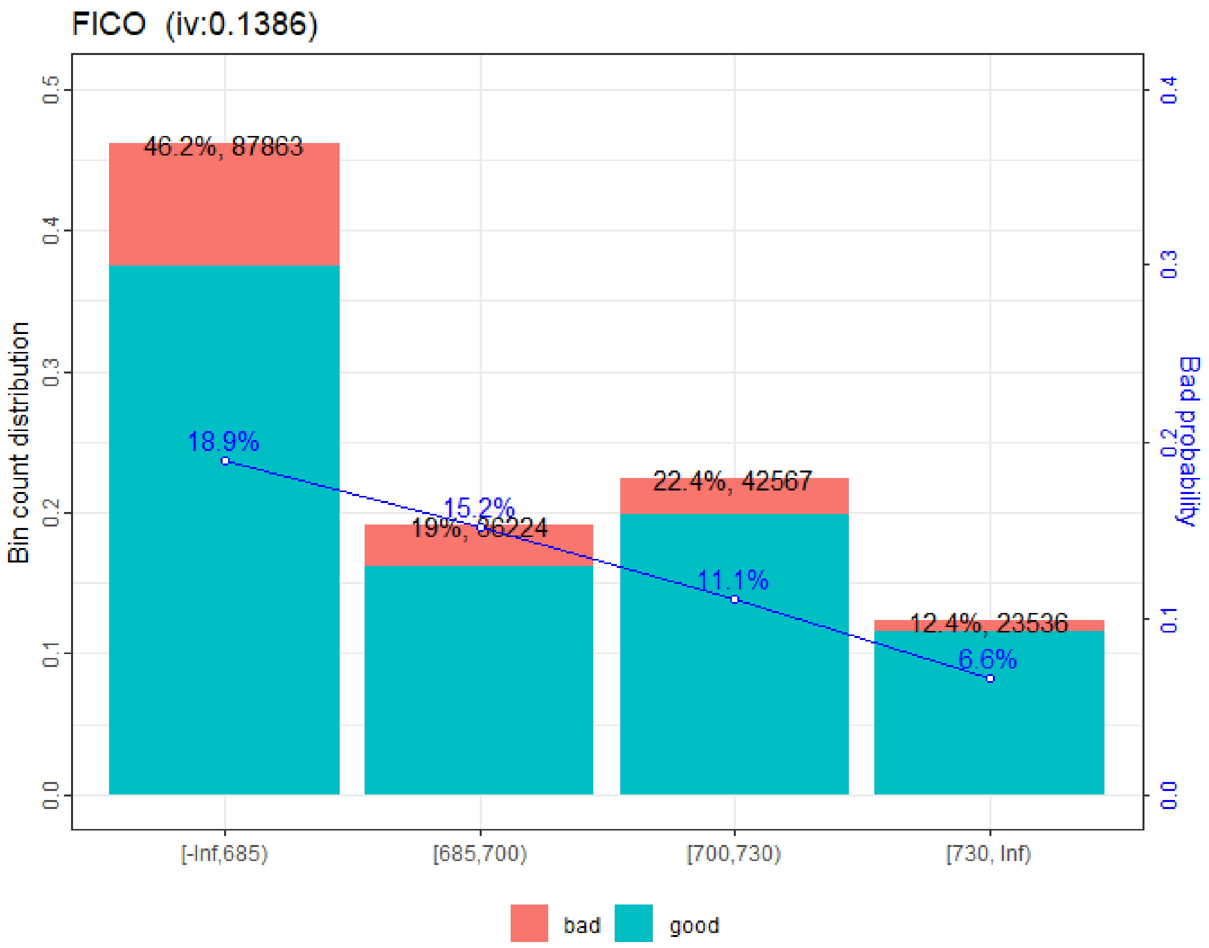
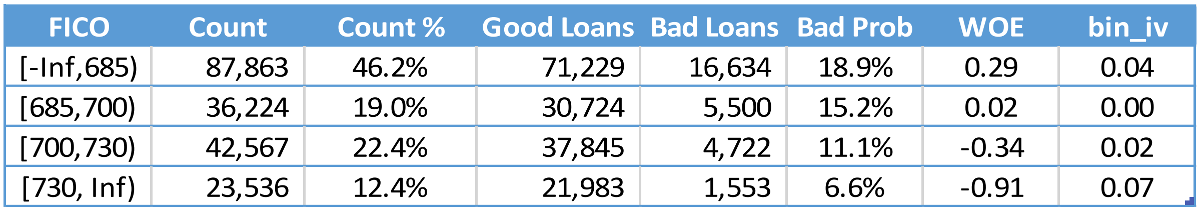


Figure 4: Default rate by FICO bands.

Table 4: Information value for FICO.



FICO	Count	Count %	Good Loans	Bad Loans	Bad Prob	WOE	bin_iv
[-Inf,685)	87,863	46.2%	71,229	16,634	18.9%	0.29	0.04
[685,700)	36,224	19.0%	30,724	5,500	15.2%	0.02	0.00
[700,730)	42,567	22.4%	37,845	4,722	11.1%	-0.34	0.02
[730, Inf)	23,536	12.4%	21,983	1,553	6.6%	-0.91	0.07

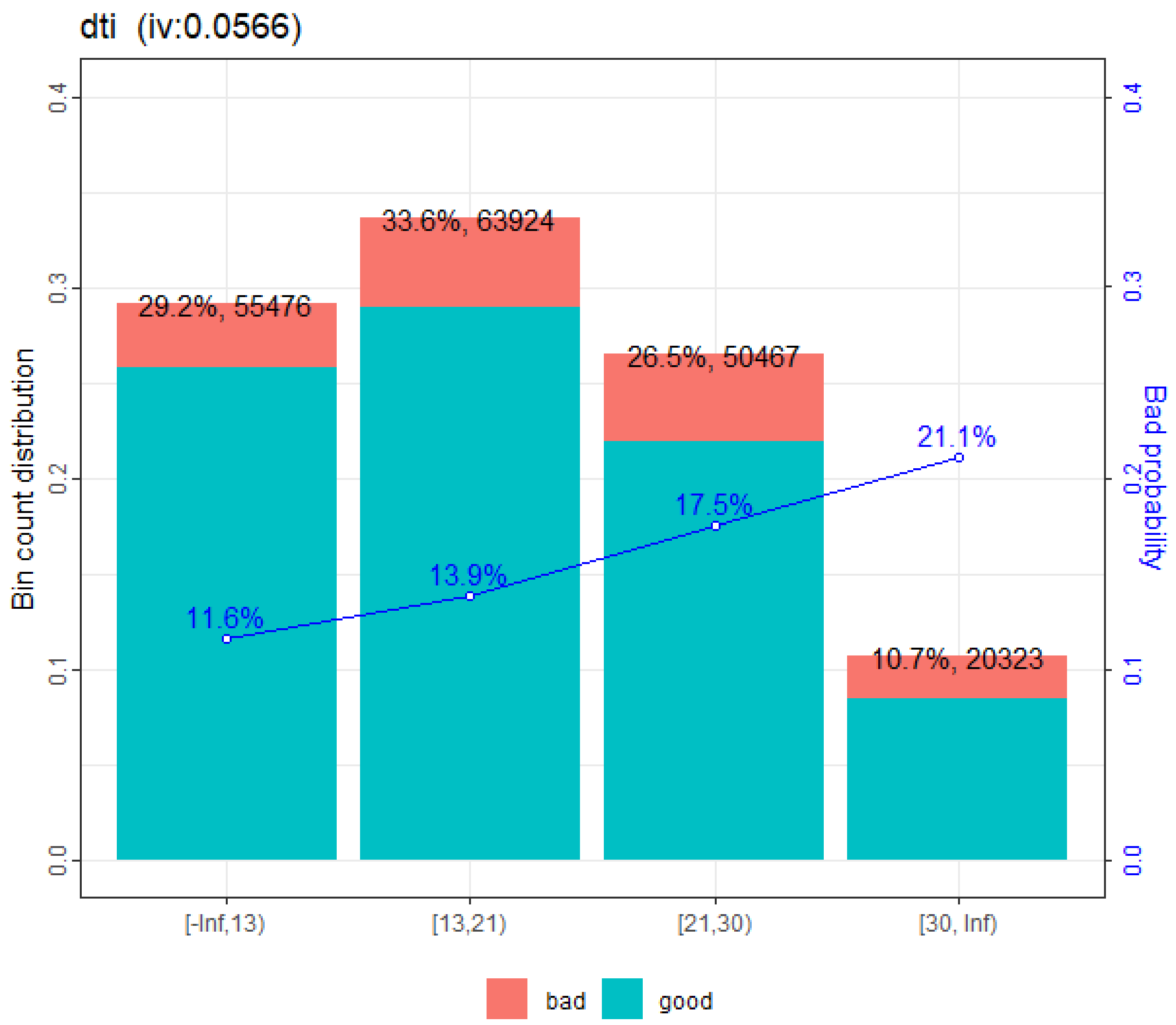


Figure 5: Default rate by DTI band.

Table 5: Information value for DTI.

DTI	Count	Count %	Good Loans	Bad Loans	Bad Prob	WOE	bin_iv
[-Inf,13)	55,476	29.2%	49,047	6,429	11.6%	-0.29	0.02
[13,21)	63,924	33.6%	55,054	8,870	13.9%	-0.09	0.00
[21,30)	50,467	26.5%	41,642	8,825	17.5%	0.19	0.01
[30, Inf)	20,323	10.7%	16,038	4,285	21.1%	0.42	0.02

Based on the information values installment amount, PMI and annual income of the borrower were the three most predictive factors. The interest rate, the subgrade, grade and DTI were also among the most predictive factors. The top 10 predictors identified through IV for further model building are shown below, (Figure 6).

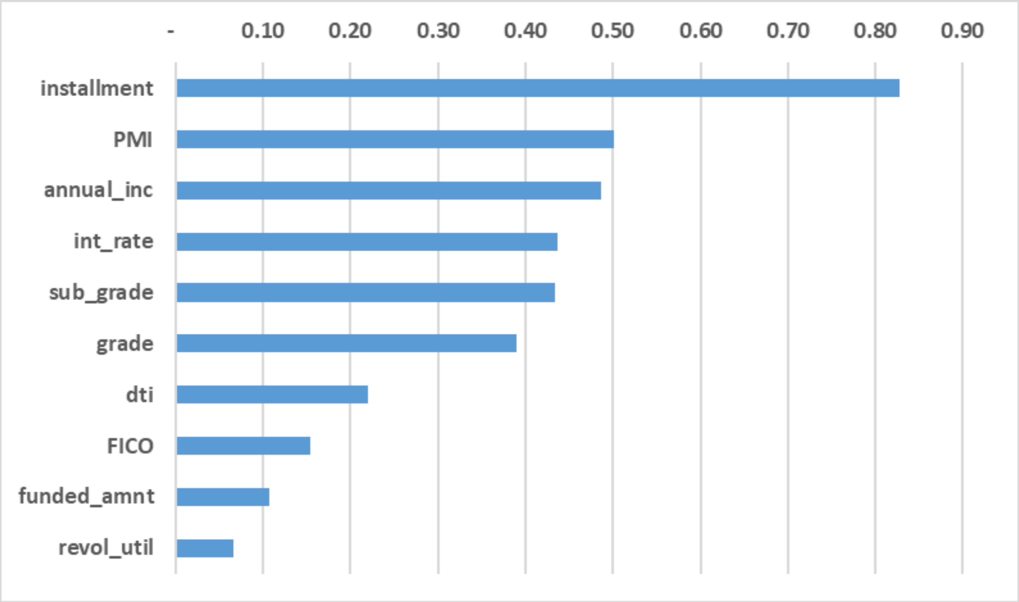


Figure 6: Top 10 predictive variables based on information value.

Random forest based variable selection techniques. These have been increasingly popular and are known to be efficient in building parsimonious models (Liaw & Wiener, 2002). This study focuses on variable selection based on the mean decrease in Gini coefficient. The Gini coefficient is a measure of node purity and ranges from 0, a case of complete purity in a node, to 1, a complete impurity in the node. The importance of a variable is then measured by assessing the degree of purity/impurity in the nodes when the variable is permutated or excluded from trees. Since the trees are being built randomly, variables that are important will keep performing well by having a higher impact in decreasing the Gini coefficient (Dinsdale &

Edwards, 2019). Essentially the permutation helps measure the importance of individual factors across all trees such that the importance of the variable is calculated as follows:

$$VI(x_i) = \frac{\sum_{t=1}^{ktree} VI^t(x_i)}{ktree} \qquad (2).$$

According to Figure 7, the random forest algorithm ranked the interest rate as the most important factor in predicting default. Variables such as FICO, the credit grade, DTI, revolving utilization rate were also found to be important. However, variables such as purpose, tax liens and inquiries were not found to be important. This result is consistent with earlier results found using information value. The rationale for exploring two distinct variable selection techniques is to arrive to a more robust collection of potential predictors. Given the technical nuances between information value based variable selection and random forest variable selection the emergence of similar variables from those distinct techniques reinforces the robustness of the factors in question. Information Value measures the predictive capability of a predictor in discerning good loans from bad loans. Random forest on the other hand builds several trees and permutates the variable in and out of the forest while assessing how pure, i.e. how homogeneous the nodes of the trees become in the presence/absence of the factor. Additionally, Random forest tends to favor continuous factors and variables with high cardinality.

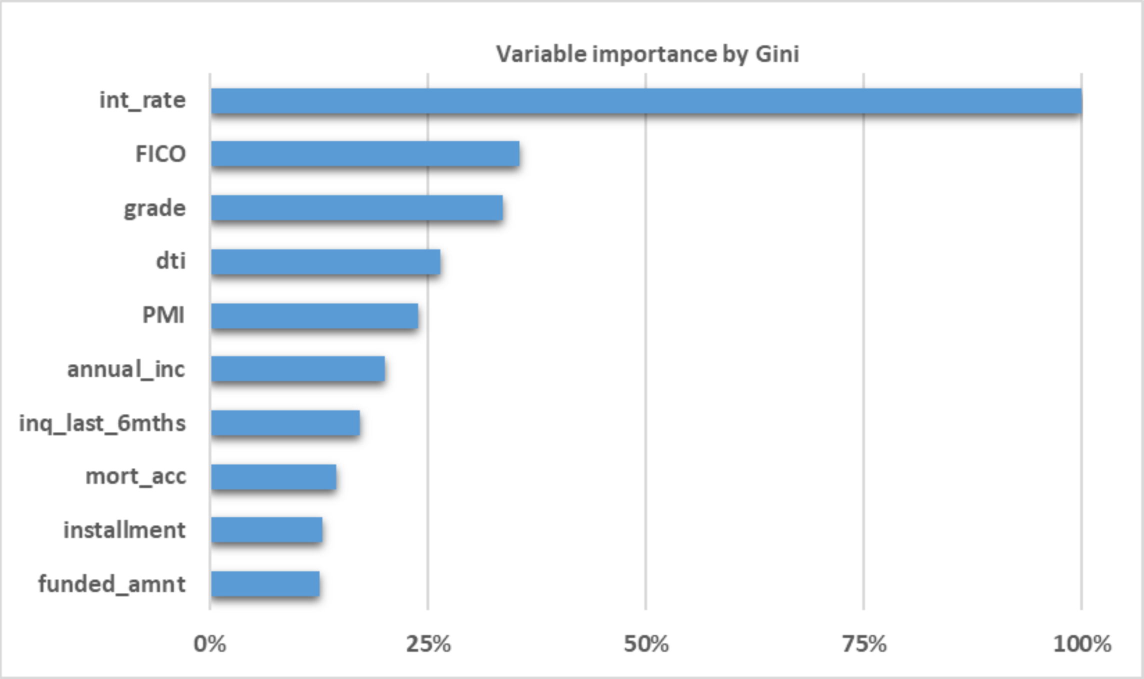


Figure 7: Variable importance: Random forest.

4.2 Modeling Techniques

We start with logistic regression which is widely used in the industry for default modeling and provides a solid theoretical framework for estimating default. Secondly, we explore the performance of Neural Networks and Random Forest algorithm which are widely popular machine learning algorithms and have proven to be effective in classification problems.

4.2.1 Logistic Regression

The logistic model is an extension of the linear regression model in which we attempt to predict a binary outcome (0 or 1). This method is well known in the credit risk arena and classification problems in general (Baesens et al., 2016). The method allows for better interpretation of model coefficients unlike black box models such as neural networks and random forest. Logistic regression estimates the conditional probabilities via a nonlinear function of factors. The logistic model outputs a probability by estimating the log linear relationship between the dependent and the explanatory variables:

$$p = \frac{\exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots\beta_k x_k)}{1+\exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots\beta_k x_k)} \qquad (3)$$

Where;

p is the probability of the event, X_k is the predictor k, β_k is the coefficient of predictor k and *k* is the number of predictors in the model.

From equation 1, the logistic regression can be written in its log odds form as follows:

$$\ln\left(\frac{p}{1-p}\right) = \beta_o + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k \qquad (4)$$

The maximum likelihood method is used for the optimization of this logit model. The paper uses multivariate logistic regression analysis to estimate several models based on the most predictive variables identifies in part 4.1.

4.2.2 Neural Networks

Neural networks are deep learning algorithms that rose from research in Artificial Intelligence (AI). They tend to perform well in classification types of problems and have been successfully implemented in credit risk analysis. Neural networks have gained prominence in recent years thanks to their ability to model complex phenomena. A neural network model typically includes several completely connected neurons. At the start of a network there is an input layer comprised of all predictors followed by one or more layers. At the end of the neural network we have the output layer which consists of only one node. All the nodes are fully connected which allows for backpropagation to take place. Backpropagation is the process which allows the network to learn from its mistakes and to adjust its predictions by sending signals back and forth between parameters while minimizing a loss function. Neural networks uses a sigmoid transformation to capture highly complex and non-linear relationships. The Sigmoid function is given by $\alpha(\pi) = \frac{1}{1+e-^{\pi}}$ (5).

Figure 7 shows a typical neural network with three layers. Firstly, there is one node for each predictor variable followed by a single hidden layer. The hidden node is also connected to a single output node. Furthermore, two control nodes are present. The first control nod is connected to the hidden layer and the second is connected to the output layer. The total number of edges is given by

$$pk + k + k + 1$$

In the case of Figure 7 we have , $8 * 1 + 1 + 1 + 1 = 11$ edges. The neural network algorithm iteratively searches for the most optimal sets of weights for each node, like coefficients estimates in linear regression analysis. Once the weights are estimates new predictions can be made by providing the set of predictors. At the end of the neural network we have the output layer which consists of only one node. The prediction of neural networks in one episode is the following:

$$y(t) = \sum_{J=1}^{N} f\left(\beta_k, x_k(t)\right) + \varepsilon(t) \qquad (6)$$

where β_k are the parameters resulting from backpropagation,

X_k are the inputs variables, $\varepsilon(t)$ is the error,

f is a nonlinear function and y(t) is the output value (Addo Martey, Guegan, & Hassani, 2018).

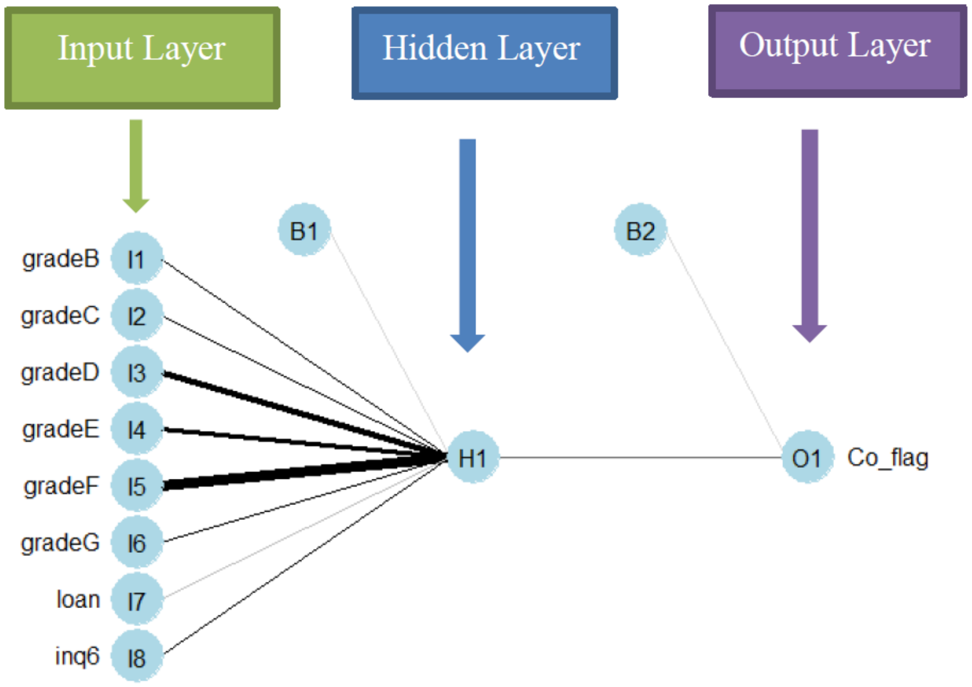


Figure 8: A neural network model.

4.2.3 Random Forests

Random forests algorithm is an ensemble model that fits several trees to arrive to a vote in the case of classification problems hence the term “Forest”. The algorithm was introduced by Leo Breiman (2001) and performs well in classification types of problems such as predicting loan default or predicting admission to a university. Random Forest is robust due to several bootstrapped resampling that is done at the tree level as well as nod level to prevent overfitting and to decorrelate the different trees in the forest. Additionally, only a random subset of predictors is considered each time a tree is built, or a split is done. Effectively the algorithm randomly select a subset of predictors k from all inputs predictors such that

$$k = \sqrt{z}$$

where z is the total number of predictors in the model. Another advantage of Random Forests is their ability to rank variable by importance and thus select those variables that provide purity based on the Gini Index or Entropy. The Gini Index and entropy are both measures of homogeneity of class in the nodes of a tree. While Gini measures how many times a randomly selected observation from the data set would be incorrectly classified. Entropy tries to gauge the disorder in classifying observations relative to the target variable, having observations from different classes in a node signifies disorder.

$$Entropy = \sum_{k=1}^{n} 1 - p(c_k)log_2(p(c_k)) \qquad (7)\,,$$

$$Gini = \sum_{k=1}^{n} p^2(c_k) \qquad (8)$$

Where $p(c_k)$ is the probability /percentage of class c_k in a node.

Random forest algorithm outputs predictions based on averaging the predictions of all the trees in the forest such that;

$$\frac{1}{k} = \sum_{k=1}^{K} k^{th}\ tree\ response \qquad (9)$$

Where the index k runs over the individual trees in the forest (Stasoft, 2019).

4.3 Model Performance Metric

Although there are several metrics for comparing binary classifiers, this analysis focuses on the area under the curve (henceforth AUC). AUC is one of the best-known model performance metrics in assessing the predictive ability of binary classifiers. The AUC curve is a measure of the area under the receiver operating curve (ROC). ROC is a graph displaying the performance of classification models at all possible probability thresholds. ROC measures the discriminative power of a classification model. AUC values range from 0 to 1 with higher AUC values suggesting higher predictive power of the model. In this paper a higher AUC value implies the model is correctly identifying good loans as being good and bad loans as being bad. For instance, an AUC of 0.9 implies that a randomly selected borrower from the defaulters group has a probability default value higher than that of a randomly selected borrower from the non-defaulter's group 90% of the time. AUC is leveraged in this paper to benchmark the performance of logistic regression models versus random forest and neural network models.

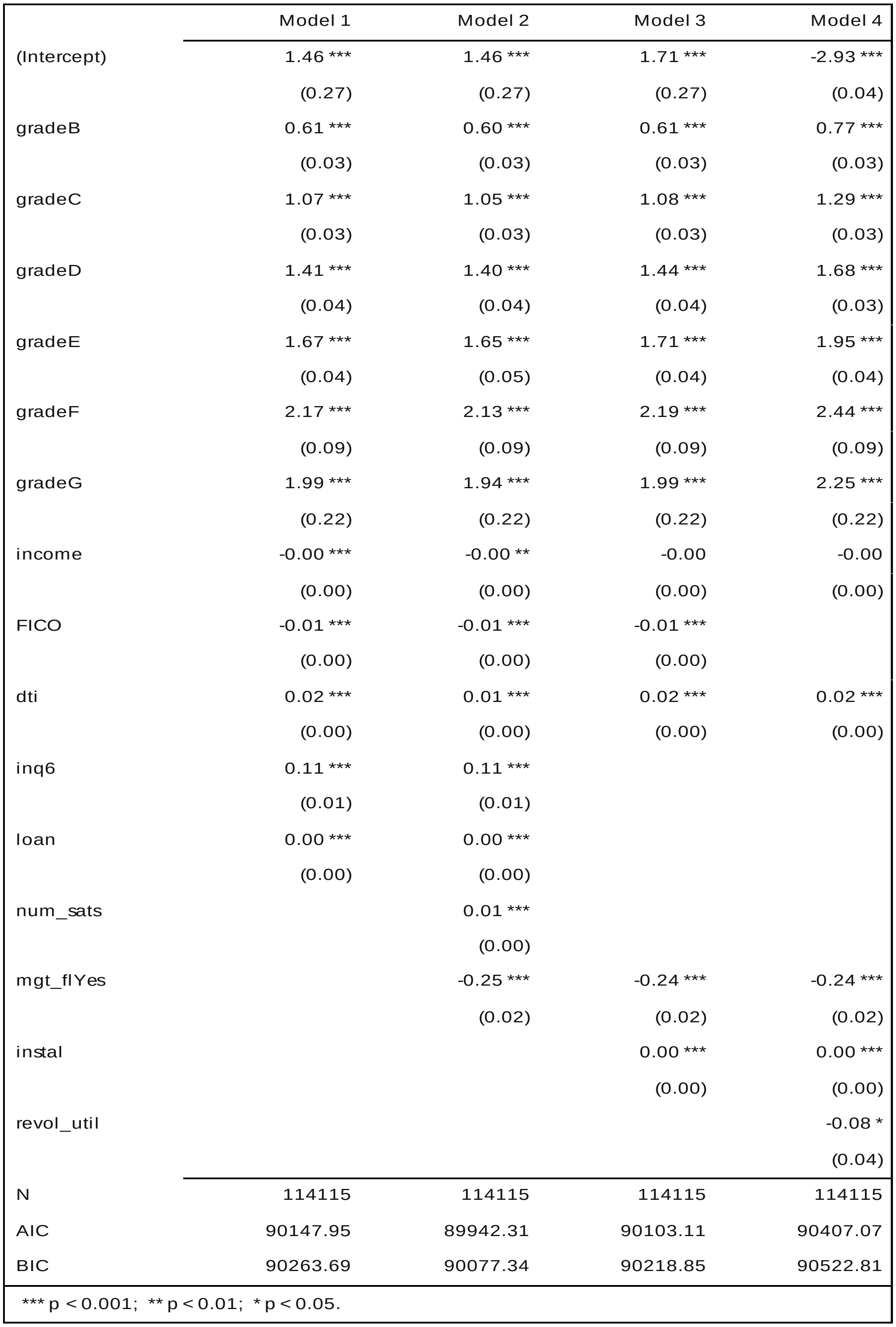
$$AUC = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} 1\, p_i > p_j \qquad (10)$$

Where i covers all the observations with true class 1 and j covers all the observations with true class 0 ; p_i and p_j denote the probability value assigned by classifier to observation i and j when the condition $p_i > p_j$ is met then the function outputs 1 (University, 2019) .

Chapter 5: Empirical Results

This chapter provides the results of predicting the likelihood of credit default based on identified credit risk attributes via three distinct classifiers. Moreover, the magnitude and direction of predictors based on four logistic regression models are displayed. Finally, the performance of logistic regression compared to random forest and neural networks based on AUC is presented. Based on the empirical results of four logistic regression models (Table 6), we find that credit grade is an important factor as well as FICO. Moreover DTI, the number of inquiries in past 6 months, number of satisfactory trades, having a mortgage account and the payment amount are also important predictors. Moreover, loan amount was found to be statistically significant although the magnitude of the coefficient suggests loan amount to have a limited impact on default in practice. A higher loan amount implies a higher monthly payment which can be difficult for borrowers especially those with existing debt. Additionally, revolving debt utilization rate, which is the percentage of total credit available currently utilized was found to be statistically significant. One surprising finding is the direction of the coefficient for revolving utilization rate. Intuitively one would expect borrowers with higher revolving line utilization to be more prone to defaulting since they already have a current debt load that requires servicing. The logistic regression results from Model 4 suggests that borrowers with higher revolving debt utilization are less likely to default once we account for other factors. One potential explanation of this is that people with higher revolving debt are more experienced with credit products and will on average default less even after accounting for the higher debt load they are currently carrying.

Table 6: Logistic regression results.



	Model 1	Model 2	Model 3	Model 4
(Intercept)	1.46 ***	1.46 ***	1.71 ***	-2.93 ***
	(0.27)	(0.27)	(0.27)	(0.04)
gradeB	0.61 ***	0.60 ***	0.61 ***	0.77 ***
	(0.03)	(0.03)	(0.03)	(0.03)
gradeC	1.07 ***	1.05 ***	1.08 ***	1.29 ***
	(0.03)	(0.03)	(0.03)	(0.03)
gradeD	1.41 ***	1.40 ***	1.44 ***	1.68 ***
	(0.04)	(0.04)	(0.04)	(0.03)
gradeE	1.67 ***	1.65 ***	1.71 ***	1.95 ***
	(0.04)	(0.05)	(0.04)	(0.04)
gradeF	2.17 ***	2.13 ***	2.19 ***	2.44 ***
	(0.09)	(0.09)	(0.09)	(0.09)
gradeG	1.99 ***	1.94 ***	1.99 ***	2.25 ***
	(0.22)	(0.22)	(0.22)	(0.22)
income	-0.00 ***	-0.00 **	-0.00	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)
FICO	-0.01 ***	-0.01 ***	-0.01 ***	
	(0.00)	(0.00)	(0.00)	
dti	0.02 ***	0.01 ***	0.02 ***	0.02 ***
	(0.00)	(0.00)	(0.00)	(0.00)
inq6	0.11 ***	0.11 ***		
	(0.01)	(0.01)		
loan	0.00 ***	0.00 ***		
	(0.00)	(0.00)		
num_sats		0.01 ***		
		(0.00)		
mgt_flYes		-0.25 ***	-0.24 ***	-0.24 ***
		(0.02)	(0.02)	(0.02)
instal			0.00 ***	0.00 ***
			(0.00)	(0.00)
revol_util				-0.08 *
				(0.04)
N	114115	114115	114115	114115
AIC	90147.95	89942.31	90103.11	90407.07
BIC	90263.69	90077.34	90218.85	90522.81

*** p < 0.001; ** p < 0.01; * p < 0.05.

Note: the names of factors have been abbreviated. The factor inq6 is inq_last_6mths , the factor instal is installment , loan is funded_amnt and mtg_fl_Yes is binary factor for mort_acc.

The logistic regression output for model 1 in Table 6 can be interpreted in the following manner:

- Having credit grade B, versus A increases the probability of default by 0.61 (grade A is the reference level).
- The probability of default increases monotonically as one goes from grade A to F with a slight decrease at grade G.
- Ceteris paribus, each 10-point increase in FICO score decreases the probability of default by 0.1.
- Ceteris paribus, each 10-point increase in DTI increases the probability of default by 0.2.
- Ceteris paribus, each additional credit inquiry in the past 6 months increases the probability of default by 0.11.

The magnitude of the coefficient for the number of recent credit inquiries is particularly large. This indicates that borrowers with large numbers of credit inquiries within 6 months will have a very high probability of credit default. Intuitively this makes sense, many credit inquiries by a borrower is likely due to a severe shortage in liquidity leading to what's known in the banking industry as "credit shopping". Credit shopping refers to the process of contacting several financial institutions in a short period of time to accumulate debt which is unlikely to be repaid. Models created with random forest and neural networks are highly complex and lack the ease of interpretability of coefficients as in the case of logistic regression. However, we can infer from those models using variable importance plots. The random forest and neural network models found similar results to the logistic regression. Both the neural network and random forest algorithms found the most predictive factors were the interest rate, FICO score, credit grade, PMI, DTI and the number of inquiries in the past six months. The random forest model built 500 trees with two variables used for splitting at each node. The neural network model with four hidden layers was found to be the most optimal model among several neural network models with a maximum ROC of 0.68 (Figure 9). Further comparison of the logistic regression model versus random forest and neural network shows that a neural network is marginally the best based on ROC (Table 7). One interesting finding is that although neural network models are much more complex, they do not always provide better performance. The neural network model performed only slightly better than the logistic model on the testing data set (Figure 9, Table 7). This finding is of practical importance given that logistic models provide better clarity and interpretability compared to black box models such as neural networks and random forest.

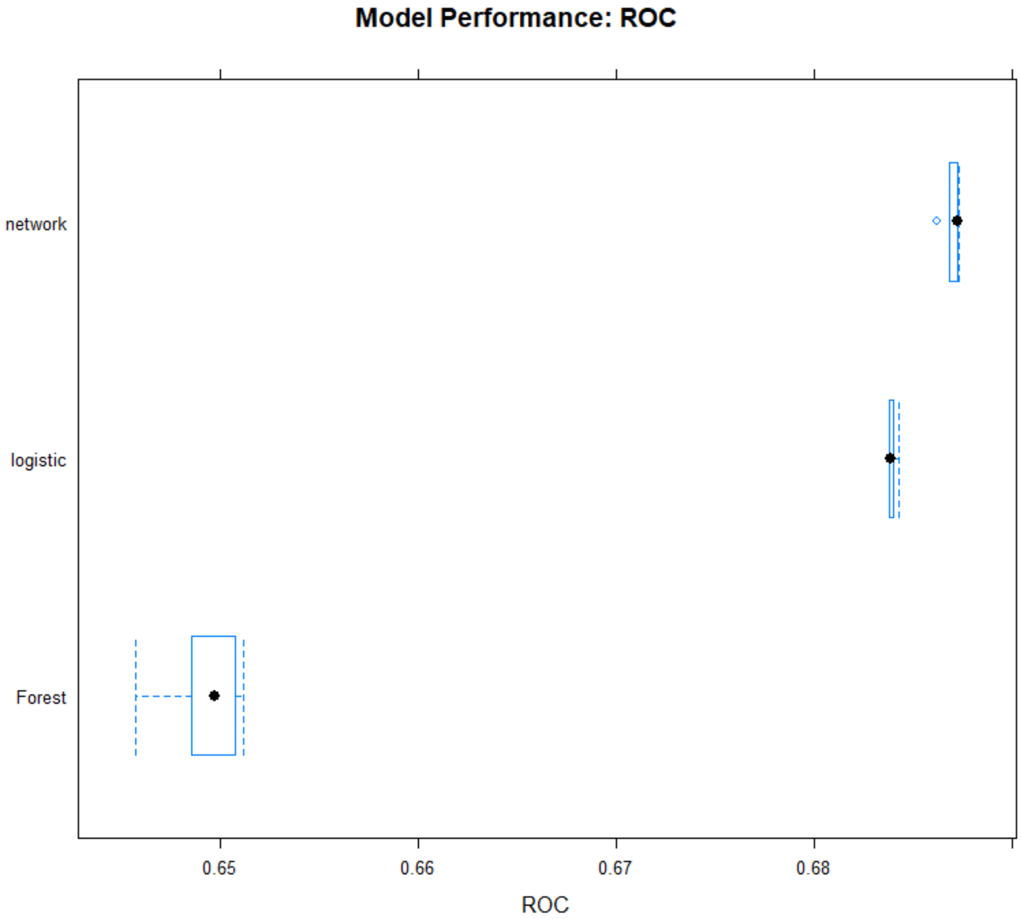


Figure 9: Performance on training data (5 folds cross validation).

Table 7: Model performance on testing data.

Model	ROC
Logistic model 1	0.68
Logistic model 2	0.6831
Logistic model 3	0.6812
Neural Network	0.6848
Random Forest	0.6472

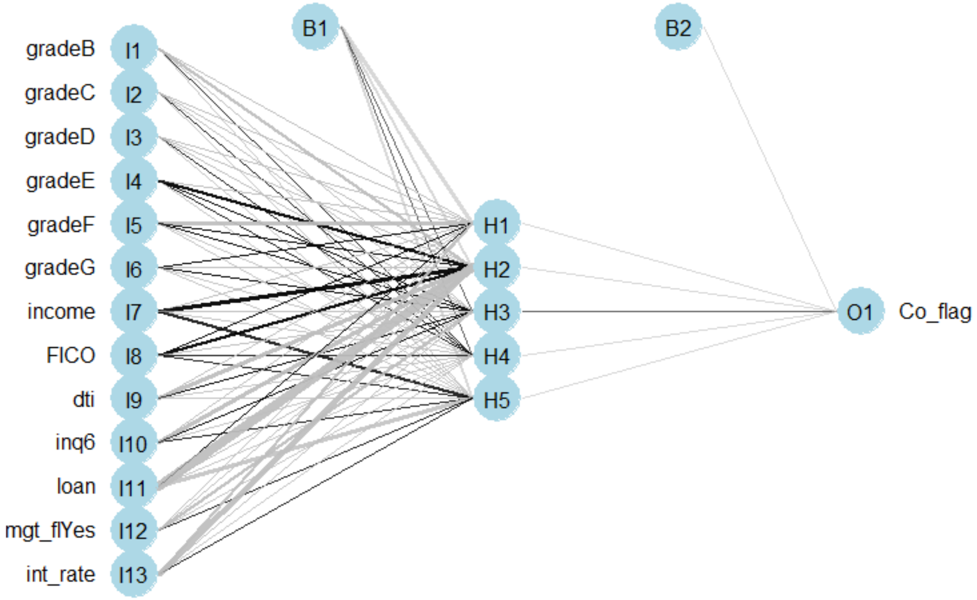


Figure 10: Neural network model with four hidden layers.

Chapter 6: Conclusion and Policy Implications

P2P lending has been increasingly relevant in today's banking and financial markets. There are great amount of opportunities to gain from this novel market and many lessons yet to be learned. This thesis explores the main factors for predicting loan default in P2P setting and compared the performance of three modeling techniques. The study found the credit grade of the borrower, FICO score, DTI, the number of inquiries and annual income to be highly predictive of loan outcomes. Moreover, loan amount, revolving utilization rate and installment amounts were found to be statistically significant. Furthermore, borrowers with mortgages tend to be more creditworthy. Another key finding of the paper is that black box models with hyperparameters such as neural networks do not necessary yield better performance than classical methods such as logistic regression in predicting credit default. There are three main policy recommendations from this paper.

First, the government regulations are needed to harmonize the regulatory regime for banks and P2P lenders. Marketplace lenders are currently benefiting from regulatory arbitrage since they are not subject to the same level of regulation as deposit-taking institutions. Creating a uniform regulatory framework will even the level playing field between platforms and traditional banks.

Second, limit risk taking: as the data shows Fintechs are growing fastest in high risk segments i.e. highly indebted, low credit and low-income segments. Despite this high growth, all the money invested is raise by private investors and the platforms do not assume any risk. A sudden economic downturn could lead to economic collapse since billions of dollars of loans will no longer be serviced. This could be more damaging than the 2008 great recession because most of the investors on P2P platforms do not have enough political clout to access public rescue packages unlike banks in 2008.

Third, expand access to financial literacy, from the data above we can see that as much as 82% of the loans issued on the platforms are reportedly for consolidation purposes. This shows that there is a large segment of the country in debt.

High levels of debts are always dangerous for the economy as shown during the great recession. More needs to be done to raise awareness of financial products and the risk in accumulating debt. Some natural extensions of the paper in the future include: (a) modeling credit default using survival analysis, (b) combining loan-application level data from several platforms to build predictive models that are more robust. Survival analysis is more dynamic since the time to default is a critical piece of credit risk assessment. Some later stage defaults do not incur any losses to the bank while default in the early stages of loan repayment can be extremely costly. However, the current paper only considers instances of default. The combination of data from several platforms can also be more robust and instrumental in assessing P2P credit risk.

References

Addo Martey, P., Guegan, D., & Hassani, B. (2018). Credit risk analysis using machine and deep learning models. *Risks*, p. 38.

Baesens, B., Daniel, R., & Scheule, H. (2016). *Credit risk analytics.* Hoboken: John Wiley & Sons.

Balyuk, T. (2019). *Financial innovation and borrowers: Evidence from peer-to-peer lending.* Retrieved from https://www.fdic.gov/bank/analytical/fintech/papers/balyuk-paper.pdf

Breiman, L. (2001). Random forests. *Machine Learning*, pp. 5-32.

Byanjankar, A., Heikkila, M., & Mezei, J. (2017). Predicting credit risk in peer-to-peer lending: A neural network approach. *Computational Intelligence (SSCI) 2017 IEEE Symposium*, (pp. 1-8).

Calebe, d. R., Loriana, P., & Paolo, T. (2016). *How does P2P lending fit into the consumer credit market?* Bundesbank discussion paper No. 30/2016.

Club, L. (2019, October 3). *LendingClub statistics*. Retrieved from The Lending Club Website: https://www.lendingclub.com/info/statistics.action

Cornaggia, J., Wolfe, B., & Yoo, W. (2018). *Crowding out banks: Credit substitution by peer-to-peer lending*. New York Fed.

Deloitte. (2016). *Credit scoring: Case study in data analytics.* Deloitte.

Demyanyk, Y., Loutskina, E., & Kolliner, D. (2017). *Three myths about peer-to-peer loans.* Federal Reserve Bank.

Dinsdale, L., & Edwards, R. (2019, October 15). *Random forests.* Retrieved from Random Forests: https://dinsdalelab.sdsu.edu/metag.stats/code/randomforest.html

Dore, T., & Mach, T. (2019). *Marketplace lending and consumer credit outcomes: Evidence from prospe.* Washington, DC: Federal Reserve Board.

Emekter, R., Tub, 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.

Hertzberg, A., Liberman, A., & Paravisini, D. (2019, August 24). *Adverse selection on maturity: Evidence from on-line.* Retrieved from https://www.federalreserve.gov/conferences/files/adverse-selection-on-maturity.pdf: https://www.federalreserve.gov/conferences/files/adverse-selection-on-maturity.pdf

Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. *R news*, pp. 18-22.

Milne, A., & Parboteeah, P. (2016). *The business models and economics of Peer-to-Peer lending.* Brussels: European Credit Research Institute.

Schumpeter. (2013). Peer review. *The Economist*.

Serrano-Cinca, C., Gutiérrez-Nieto, B., & López-Palacios, L. (2015). Determinants of Ddfault in P2P lending. *PLOS*.

Stasoft. (2019, October 19). *Random forests*. Retrieved from Statsoft.com: http://www.statsoft.com/Textbook/Random-Forest

Treasury, U. D. (2016). *Opportunities and challenges in online marketplace lending.* Washington: U.S. Department of the Treasury.

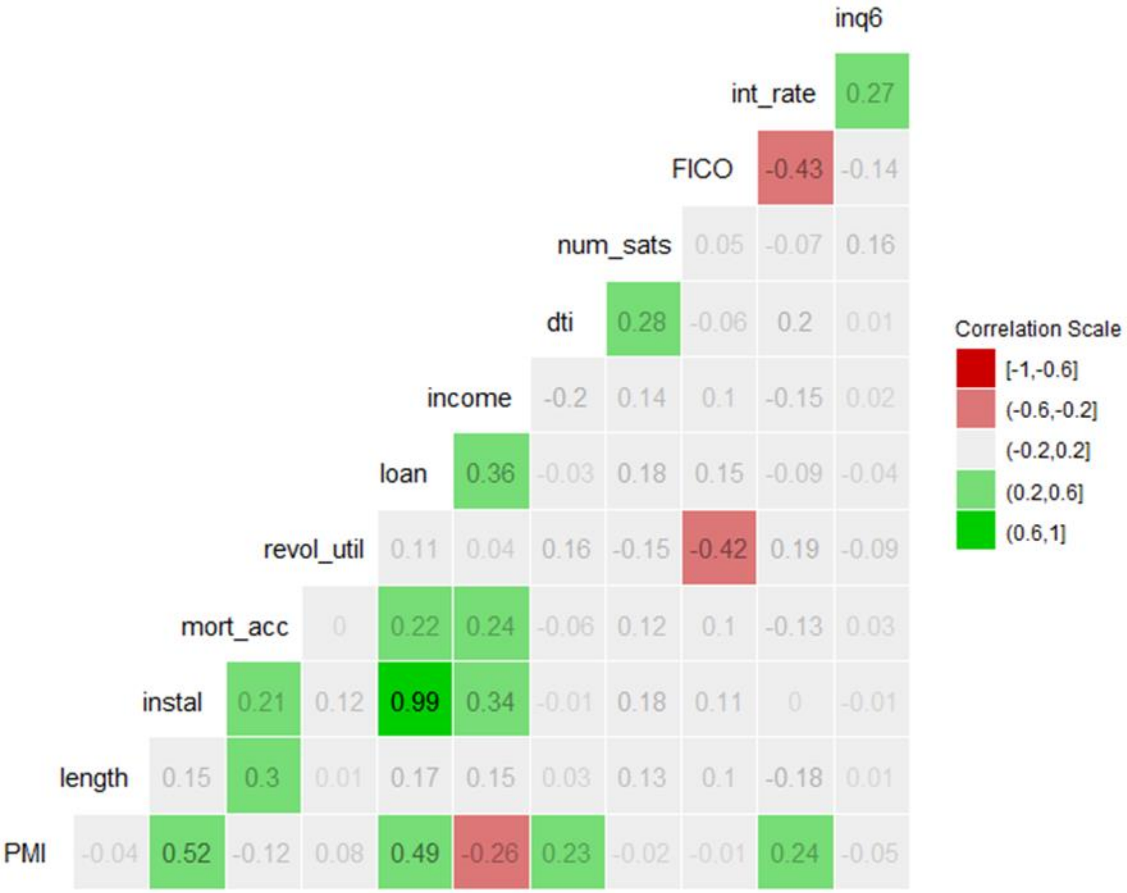
University, R. (2019, November 9). *Calculating ROC curves and AUC scores.* Retrieved from Radboud University: http://www.cs.ru.nl/~tomh/onderwijs/dm/dm_files/roc_auc.pdf

Zou, Z., Huixin, C., & Zheng, X. (2017). A study of non-performing loan behavior in P2P: Lending Under Asymmetric Information. *Transformations in Business & Economics*, pp. 490-504.

Appendix A: Data Dictionary

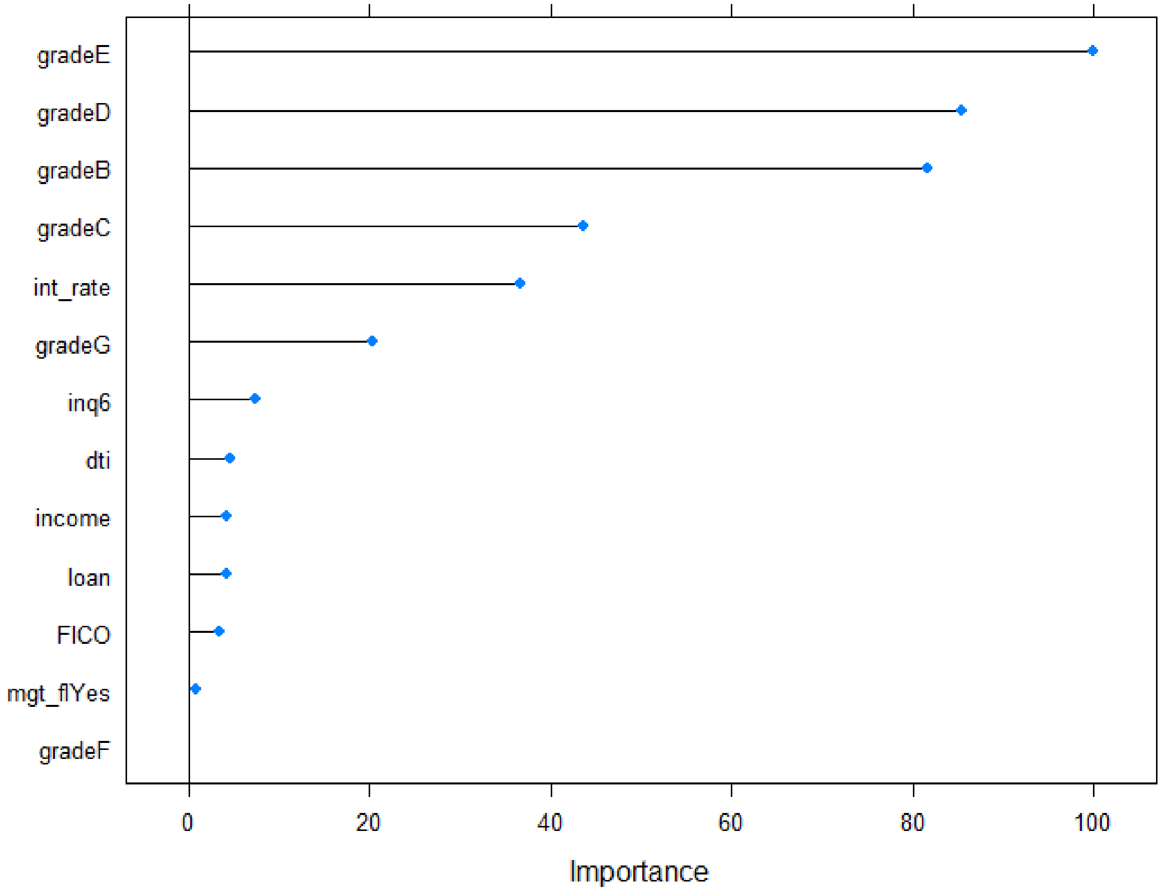
Factor	Alias	Definition
Annual_inc	Income	The self-reported annual income provided by the borrower.
DTI		A ratio of total monthly debt payments on the total debt obligations.
Grade		LC assigned loan grade.
HomeOwnership		The home ownership status provided by the borrower.
MortAcc		Number of mortgage accounts.
RevolUtil		Amount of credit borrower is using relative to all available revolving credit.
purpose		A category provided by the borrower for the loan request.
InqLast6Mths	Inq6	The number of inquiries in past 6 months .
Int_rate		Interest Rate on the loan.
FundedAmnt	Loan	The total amount committed to that loan at that point in time.
Num_sats		Number of satisfactory accounts.
Installment	Instal	The monthly payment owed by the borrower if the loan originates.
PMI		Ratio of payment to income.
Sub_grade		LC assigned loan subgrade.
Mgt_fl		A flag indicating the presence/absence of a mortgage account.

Appendix B: Correlation between Top Variables

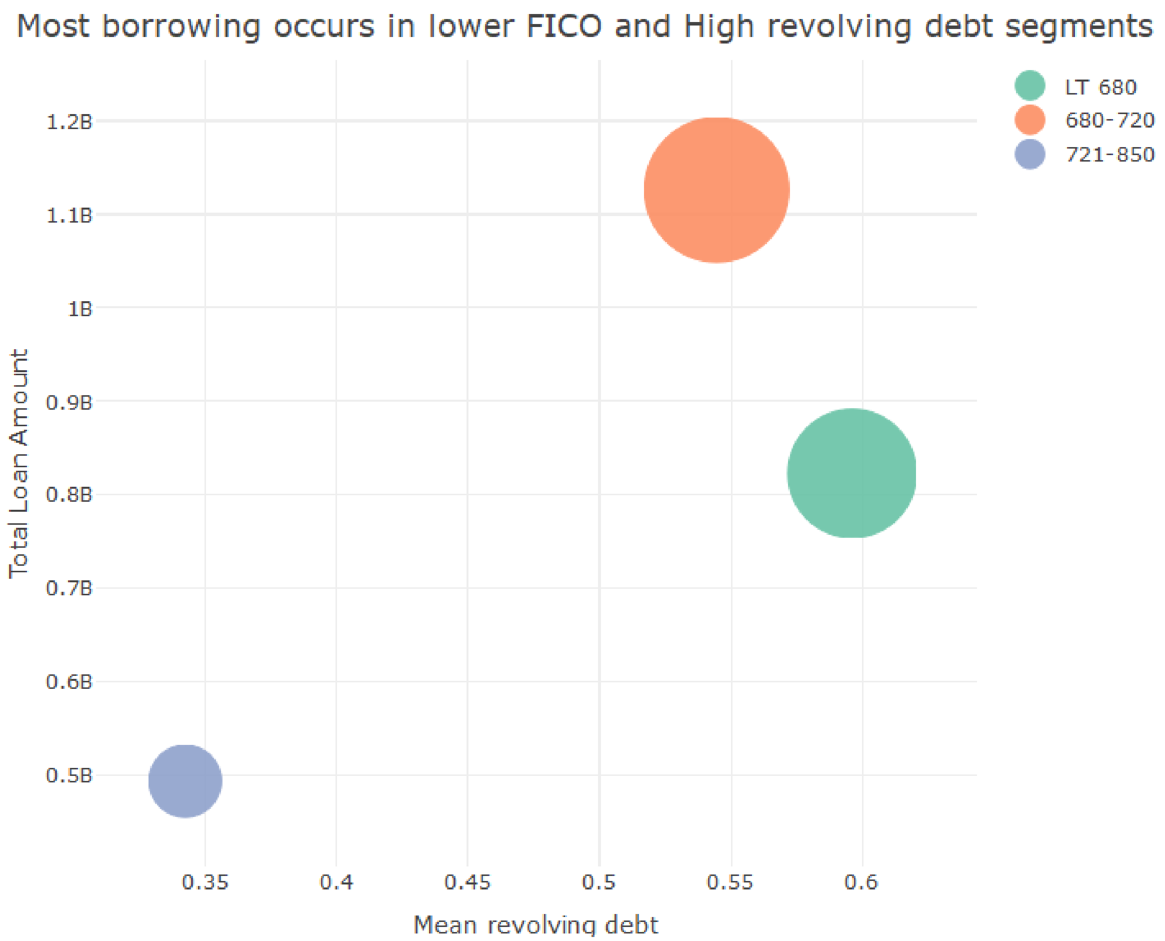


	length	instal	mort_acc	revol_util	loan	income	dti	num_sats	FICO	int_rate	inq6
int_rate											0.27
FICO										-0.43	-0.14
num_sats									0.05	-0.07	0.16
dti								0.28	-0.06	0.2	0.01
income							-0.2	0.14	0.1	-0.15	0.02
loan						0.36	-0.03	0.18	0.15	-0.09	-0.04
revol_util					0.11	0.04	0.16	-0.15	-0.42	0.19	-0.09
mort_acc				0	0.22	0.24	-0.06	0.12	0.1	-0.13	0.03
instal			0.21	0.12	0.99	0.34	-0.01	0.18	0.11	0	-0.01
length		0.15	0.3	0.01	0.17	0.15	0.03	0.13	0.1	-0.18	0.01
PMI	-0.04	0.52	-0.12	0.08	0.49	-0.26	0.23	-0.02	-0.01	0.24	-0.05

Appendix C: Neural Network Model Variable Importance



Appendix D: FICO and Revolting Debt Utilization among Borrowers


Most borrowing occurs in lower FICO and High revolving debt segments
LT 680
680-720
721-850
Total Loan Amount
1.2B
1.1B
1B
0.9B
0.8B
0.7B
0.6B
0.5B
0.35
0.4
0.45
0.5
0.55
0.6
Mean revolving debt