RISK MANAGEMENT MAGAZINE

Vol. 16, Issue 2
May – August 2021

EXCERPT

https://www.aifirm.it/newsletter/progetto-editoriale/

Why segmentation matters: a Machine Learning approach for predicting loan defaults in the Peer-to-Peer (P2P) Financial Ecosystem

Adamaria Perrotta, Georgios Bliatsios

Why segmentation matters: a Machine Learning approach for predicting loan defaults in the Peer-to-Peer (P2P) Financial Ecosystem

Adamaria Perrotta (UCD - University College Dublin), Georgios Bliatsios (UCD - University College Dublin – AIB Bank ROI),

Article submitted to double-blind peer review, received on 12th June 2021 and accepted on 5th July 2021

Abstract

Peer-to-Peer (P2P) lending is an online lending process allowing individuals to obtain or concede loans without the interference of traditional financial intermediaries. It has grown quickly the last years, with some platforms reaching billions of dollars of loans in principal in a short amount of time. Since each loan is associated with the probability of loss due to a borrower's failure, this paper addresses the borrower's default prediction problem in the P2P financial ecosystem. The main assumption, which makes this study different from the available literature, is that borrowers sharing the same homeownership status display similar risk profile, thus *a model per segment should be developed*. We estimate the Probability of Default (PD) of a borrower by using Logistic Regression (LR) coupled with Weight of Evidence encoding. The features set is identified via the Sequential Feature Selection (SFS). We compare the forward against the backward SFS, in terms of the Area Under the Curve (AUC), and we choose the one that maximizes this statistic. Finally, we compare the results of the chosen LR approach against two other popular Machine Learning (ML) techniques: the k Nearest Neighbors (k-NN) and the Random Forest (RF).

Keywords: P2P lending, risk prediction, machine learning, loan defaults.

1. Introduction

Peer-to-Peer lending (P2P) is a new financial ecosystem - rapidly growing in the last years - that allows individuals to obtain or concede loans outside the traditional financial intermediary system. In fact, as a consequence of the Global Financial Crisis (GFC), credit from banks has become difficult to access for many consumers and small businesses, so the need for speed of turnaround with available credit has highly increased. P2P lending has emerged in this setting as an innovation in microfinance. Individuals are directly connected through online lending platforms, which can provide loans with lower intermediation costs. Also, borrowers with good credit score can be offered on average lower interest rates than in banks (Namvar, 2013), while lenders with well-diversified portfolio usually target returns that outperform saving accounts or certificates (Serrano-Cinca, Gutierrez-Nieto, & Lopez-Palacios, 2015).

Since online P2P lending platforms allow individual lenders to aggregate their funds to finance loan requests from individuals and businesses, it can be also considered as a crowdfunding debt form. As a consequence, investors, businesses, and regulators are all closely monitoring this new business model since it is a prime example of how information and Internet technologies are transforming the finance industry.

As a consequence of its popularity in microfinance, P2P has recently attracted the interest of many researchers with three main streams of studies involving: the reason behind peer-to-peer lending development, the factors that affect the likelihood of default or funding success (qualitative studies) and the assessment of credit risk prediction of individual loans (quantitative studies).

In relation to the first two streams (qualitative ones) Berger and Gleisner in (Berger & Gleisner, 2009) analyze the role of intermediaries in the development of the P2P market using approximately 14,000 observations from the real-world lending platform Prosper. They find that borrowers operating on these platforms have easier access to financing compared to the standard intermediaries. Wei and Lin in (Wei & Lin, 2016) examine the matching mechanisms of supply and demand in the P2P market, whether the obtained equilibrium of interest rates is optimal and whether the choice of matching mechanism is associated with the default rates. They report that the likelihood of loan approval increases and that the offered interest rates are higher when Fintech lenders impose a matching mechanism.

Furthermore, it is well known that information economics assume that information asymmetry could cause adverse selection and moral hazard which provides theoretical base for the cause of credit risk. This also applies to the situation of P2P lending, because lenders generally do not have much information about the borrower's ability to repay the loan or his credibility. Furthermore, most of the borrowers are individuals or small private business who are normally under stressed economic conditions. In this setting, Freedman and Jin in (Freedman & Jin, 2014) and Lin, Prabhala and Viswanathan in (Lin, N. Prabhala, & Viswanathan, 2013) perform a study on real-world online P2P lending platform, Prosper.com. A unique feature of Prosper is its use of social networking through groups and friends. A non-borrowing individual may set up a group on Prosper and become a group leader. The group leader does not have any legal responsibility. Rather, the group leader is supposed to foster a "community" environment within the group so that the group members feel social pressure to pay the loan on time. Group leaders can also provide an "endorsement" on a member's listing and bids by group leaders and group members are highlighted on the listing page. Freedman and Jin find that such kind of social network information available on Prosper help lenders make good judgments about borrowers, while Lin, Prabhala and Viswanathan prove that friendships increase the probability of successful funding, lower interest rates on funded loans, and are associated with lower ex post default rates. Iyer et al. in (Iyer, Khwaja, Luttmer, & Shue, 2009) mainly study how lenders in the P2P lending markets judge the creditworthiness of borrowers. They find that although lenders consider more the standard banking "hard" information, like credit score or loan repayment stream, the "soft" information available on Prosper like communication with borrowers belonging to a group, maximum interest rate the borrower is willing to pay, or the number of words used in the listing text descriptions also play a role in the success of borrowing.

Michels in (Michels, 2012) uses data originated through Prosper.com to investigate the relationship between voluntary disclosures and the cost of debt, showing that these voluntary disclosures made in loan listings do affect the loans' interest rates. He proves that more unverifiable disclosures are associated with a lower interest rate on a loan. Additionally, more unverifiable disclosures increase the bidding activity on a loan listing.

Finally, Dortfleitner et al. in (Dorfleitner, et al., 2016) investigate the relationship between a set soft factors that are derived from the description texts with the probability of successful funding and the probability of default. The study is based on two leading European P2P platforms located in Germany, Smava and Auxmoney. They find that spelling errors, text length and words indicating positive emotion are associated with the probability of successful funding on the less restrictive of the platforms, Auxmoney, which does not require credit scores and leaves more room for voluntary information. Also, they show that conditional on being funded, text-related factors hardly predict default rates in peer-to-peer lending for both platforms.

In relation to the quantitative studies, a big number of researchers investigated the borrower's default prediction problem in the P2P financial ecosystem; currently this is in fact the most popular research stream in this sector. For example, an analysis conducted by Emekter, Tu, Jirasakuldech and Lu (Emekter, Tu, Jirasakuldech, & Lu, 2015) shows that in reality borrowers with high income and potentially high credit scores do not participate in these types of markets. Moreover, other studies show that higher interest rates for higher risk borrowers usually fail to work in this system, which means that the P2P loan grades are not accurate enough to estimate the potential risk lenders are facing. As a consequence, understanding the application of proper credit management techniques across various platforms is fundamental to evaluate P2P loans. In the literature, there exists a big number of classification algorithms for assessment of the borrowers' creditworthiness and some comparison studies have been already published (see as an example (Baesen, et al., 2003) and (Lessmann, Baesens, Seow, & Thomas, 2015)). However, such theoretical studies have been conducted on small databases with mostly unknown origin. It is worth noting that there are only three real-world platforms that make their data about social lending available: Bondora, Prosper and Lending Club. At the moment, there are no qualitative studies comparing ranking algorithms based on Bondora and Prosper. The papers based on Prosper focus on social and economical studies related to P2P lending and have been cited above. The majority of literature tackling the problem of predicting credit risk uses data originated by Lending Club data and for this reason in the following of the section we refer only to research studies based on this dataset.

In (Kumar, Natarajan, Keerthana, Chinmayi, & Lakshmi, 2016) the Lending Club dataset is used to verify which features are fundamental to determine which individuals are more likely to repay their debts with interest and on time. They use precision and accuracy as measures of performance and conclude that Random Forest is the most appropriate classifier to identify which borrowers would not pay their debts on time, while a single Decision Tree is the best for identifying creditworthy customers.

In (Fu, 2017) the author proposes a method to combine Random Forest and Neural Network, while in (Jin & Zhu, 2015) the authors use Random Forest for selecting the features entering into the modeling phase. In particular, here the target is classified into three categories (default, need attention, not default) rather than just two (default, not default). The authors compare five machine learning models: two Decision Trees, two Neural Networks and one Support Vector Machine and use two metrics: average percent hit rate and area of the lift cumulative curve, to evaluate the prediction results.

A few papers based on Lending Club data are also dedicated to comparing classification methods. In (Wu, 2014) the author compares Logistic Regression and Random Forest.

The research presented in (Tsai, Ramiah, & Singh, 2019) is aimed to avoid as many false positive predictions as possible and therefore precision is used as a performance measure. Moreover, the authors here use a modified version of Logistic Regression with a penalty factor to avoid false positive predictions.

In (Malekipirbazari & Aksakalli, 2015) the authors propose a Random Forest for predicting a borrower's status. Out of all features available, they used only 15 of them. Starting from a 5-fold cross-validation procedure, the parameters of the classifiers are tuned and different metrics are reported. As a result, the authors conclude that the Random Forest obtains superior results when compared to Support Vector Machine, Linear Regression and k-Nearest Neighbors.

Chang, Kim and Kondo in (Chang, Kim, & Kondo, 2015) compare the performance of different naïve Bayes distributions and kernel methods for a Support Vector Machine. They find that naïve Bayes with Gaussian distribution and a Support Vector Machine with linear kernel have the best performance. Finally, in (Teply & Polena, 2020) the authors use a 5-fold cross validation approach, ten different classification techniques (divided into three groups based on the type of algorithm they use: linear, non-linear or rule-based algorithms), and six different performance measures. According to their ranking, logistic regression is placed as the best and artificial neural network as the second-best classification method.

Our contribution is placed in this setting, since we address the borrower's default prediction problem based on Lending Club Data between 2007 and 2014 and we compare some classification algorithms. Our main assumption, which makes this study different from the available literature, is that borrowers sharing the same homeownership status display similar risk profile, thus *a model per segment should be developed* instead of a unique model for all borrowers, as done in all the above cited papers. We adopt the Logistic Regression as our modeling method combined with Weight of Evidence encoding (WoE), which adds a great value to our findings, as shown in the result section. The set of features used in our model has been identified via the Sequential Feature Selection (SFS) process. Specifically, we compare Forward against the Backward SFS in terms of the Area Under the Curve (AUC) and we choose the forward one since it maximizes this statistic. We show that in this setup, our model achieves excellent predictive power using nine features. Finally, in light of the various feature sets generated by the Forward SFS, we compare the Logistic Regression modeling method against the k-NN and Random Forest in terms of the AUC at each iteration. The remainder of the paper is organized as follows. In Section 2 we describe the available Lending Club dataset, the preprocessing and cleansing of the data and the features selection procedure. Section 3 is dedicated to the description of the mathematical background of our methodology. In Section 4 we show the experimental results. Finally, the Section 5 concludes the paper and states final remarks.

2. Data Preprocessing, Cleansing and Feature Analysis

As pointed out in the Introduction, we implemented our methodology on a real-world P2P lending dataset, containing 466,285 records. The dataset used to develop and validate our model contains consumer loans issued in the U.S. by Lending Club between 2007 and 2014. It includes 57[1] features of account, financial and demographic nature (see Table in the Appendix section for

[1] In total 74 of which 17 have no data (empty columns).

details) and is available for free download on Kaggle[2]. In the following we will use the word “feature” to refer to the variable of the Lending Club dataset. In literature, the word “feature” is equivalent to “factor” or “variable” or “characteristic”. Features are the basic building blocks of datasets. The quality of the features in a dataset has major impact on the quality of the insights one will gain especially when Machine Learning (ML) algorithms are employed.

In order to address the borrower's default prediction problem, we define our target variable as a function of the ***loan status***. Specifically, we generate a binary variable in such way where all consumer loans having loan status equal to any of the following *"Default", "Charged Off", "Late (31 - 120 days)", "Does not meet the credit policy. Status Charged Off"* are assigned the value of 1, indicating that are in default**,** otherwise they are assigned a value of 0. Our main assumption is that all borrowers sharing the same ***home ownership*** status, be it "Mortgage", "Rent" or "Own", are likely to display a similar risk behavior. Therefore, our proposal is that a separate model per ***segment*** (i.e. home ownership status) is developed since the more homogenous the population is, the better the stability and predictive power of the model will be. We note that three additional home ownership types exist in the data, however they account for less than 0.05% of the population; and for this reason, they were not included in the study. The breakdown of the population distribution is given in Figure 1.

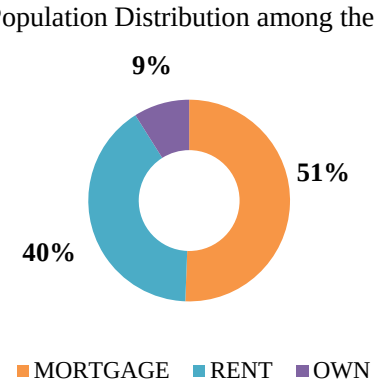


Figure 1: Home Ownership Segments

Figure 2 and 3 below show the quarterly default rate and population distribution of each segment. As can be observed, all three population distributions are left skewed with more than 85% of the data to be concentrated between 2012 and 2014. In terms of the default rate, the high volatility in the early periods is directly related to the low population figures. As the number of borrowers gradually increases, the volatility decreases and the default rate is trending downwards towards its true value. Once the defaulted population per segment had been identified, we pair wise compared the segments by conducting a two-sided Kolmogorov - Smirnov test at 5% significance level. We chose this test firstly because this is a robust non-parametric test which allows us to compare the segments based solely on their empirical cumulative distribution (thus we don’t have to make any assumption on the underlying distribution as other tests would require; such assumptions potentially would not properly fit the skewed distribution of our dataset and introduce a bias in the study). Moreover, the Kolmogorov - Smirnov test is one of the most popular choices across various industries to address the same type of problem, which further supported such decision. In our study, the null hypothesis of the test is that the defaulted borrower samples between the segments come from the same distribution. As can be observed in Table 1, in each scenario the null hypothesis is rejected, which further supports the assumption of using a different model per each segment.

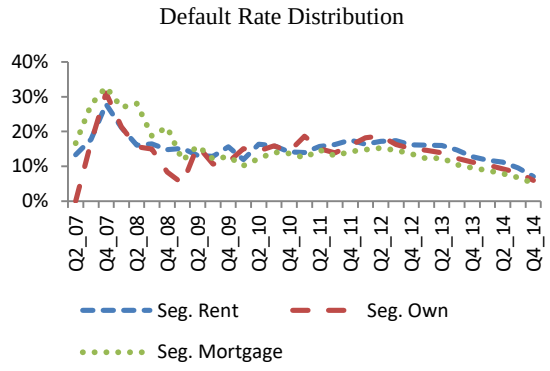


Figure 2: Default Rate Distribution

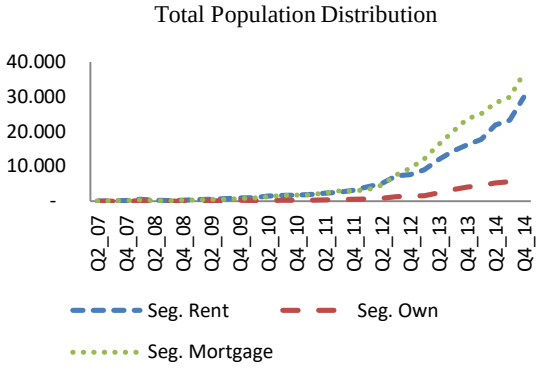


Figure 3: Population Distribution

[2] https://www.kaggle.com/wordsforthewise/lending-club, version 3 (2019 - 04 - 10)

Scenario	KS - Statistic	KS – Critical
Mortgage vs Own	0.0468	0.0221
Own vs Rent	0.0573	0.0220
Rent vs Mortgage	0.0327	0.0126

Table 1: Distribution Comparison

Starting from the assumption of using a different model per segment, we focus our analysis only on the one having home ownership status as "Mortgage"; it is in fact the largest segment and contains ***235,875*** consumer loans, accounting for approximately ***51%*** the data.

We begin our study by performing a data preprocessing and cleansing analysis. Firstly, we keep only those features for which the missing rate is less than or equal to 50% (Siddiqi, 2017). For the features having less than 5% of missing rate, we perform data imputation by substituting the nulls with the most frequent value of each feature (see (Siddiqi, 2017), (Joenssen & Bankhofer, 2012)). The rest of the features has been treated individually, as described in the following steps.

Next to the cleansing analysis, a visual data exploration is performed to understand the nature of each feature and to find out potential trends or grouping options. For example, when controlling for the feature "Purpose" we observe that 80% of the data is concentrated into two categories, "Debt Consolidation" and "Credit Card". Thus, in this case we group the values of this feature in three categories, "Debt Consolidation", "Credit Card" and a third one called "Other" containing all other possible options within the feature. For what concerns the numerical features, we detect outliers based on the Z-scores method and replace them with the median of the feature (in essence, this method measures how many standard deviations an observation is far from the mean of the feature it belongs. We define as an outlier any observation which lies more than 3 standard deviations of the mean of the feature it belongs to). Furthermore, features having the same value for more than 95% of the observations are being dropped since they bear no substantial predictive power as they can be considered almost constant (see (Al-Jabery, Obafemi-Ajayi, Olbricht, & Wunsch, 2020), (Teply & Polena, 2020)). The same applies for features that by construction are highly correlated between each other (see for example "Grade" and "Sub - Grade" in Table 7 in the Appendix) in which case, we only keep one. Simply put, in the case where strongly related features would be chosen for modeling, it would have a negative impact on the accuracy of the model's prediction (see Results section for further discussion on the topic). Finally, variables that contain information which cannot be utilized in any meaningful way, such as "URL", are also being dropped. Once the preprocessing and cleansing was completed, the remaining number of candidate features was reduced from 57 to 26 (see Table 7 in the Appendix section for details).

As second step, we converted all remaining variables to categorical ones allowing for smooth implementation of Weight of Evidence encoding (WoE) (we refer to (Baesens, Roesch, & Harald, 2016) for details). The WoE is a univariate measure describing the relationship between a predictor variable and the dependent one. Assuming that we have consolidated the values of a feature into *m* distinct bins (or equivalently "buckets" or "groups"), the WoE is given by:

$$WoE_i^{Variable} = ln\left(\frac{Distribution\ of\ Performing\ cases}{Distribution\ of\ Defaulted\ cases}\right), i \in \{1,2,\dots m\} \quad (1)$$

As mentioned earlier, when it was meaningful during the data preprocessing phase, we further merged the bins of the categorical variables in which we observed heavily imbalanced distribution of the population. When that was not necessary, we kept the original number of bins intact (i.e., we used the distinct values of the feature as our bins). We calculated the WoE for each feature and we fine-tuned further the bins until strict monotonicity, in WoE terms, was achieved. In addition to improving model performance, this encoding process allows for smooth resolution of the missing values problem since, when applicable, a separate bin was created in order to group them together. Following the fine tuning of the bins, for each feature we computed the Information Value (IV); it is measure of the predictive strength of a feature, given by the following formula:

$$IV_{Variable} = \sum_{i=1}^{m}(Distribution\ of\ Performing\ Cases_i - Distribution\ of\ Defaulted\ Cases_i)WoE_i^{Variable} \quad (2)$$

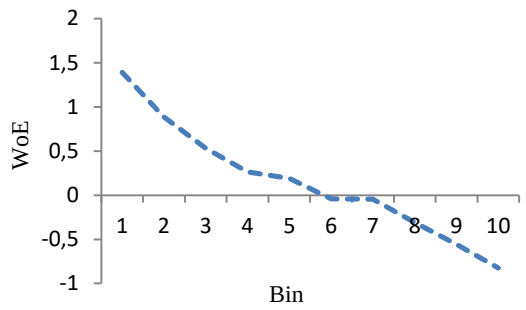
At this step, only the features having IV greater than 0.01 proceed to the next phase. For the numeric features, an equal-size binning approach (i.e, quantiles) was employed. Specifically, each feature was grouped recursively into three, four and all the way up to ten bins and for each iteration, the IV of the features was calculated. The ultimate number of bins was then given by the iteration for which the IV attained its maximum value subject to the condition that the WoE was strictly monotonous. Similarly, as in the case of the categorical features, missing values were grouped together and only the features for which the IV was greater than 0.01 proceed to the next phase. Once the feature engineering was completed, the remaining number of features was further reduced from 26 to 15. In Table 2 we provide a binning example using the numeric feature "Debt to Income". In this case, four bins were created based on the algorithm described previously.

Feature: Debt to Income - Weight of Evidence						
Bin	**Observations**	**Defaults**	**Default Rate**	**Min (%)**	**Max (%)**	**WoE**
1	58,969	4,514	7.7%	0	11.49	0.247514
2	58,969	5,173	8.8%	11.49	16.82	0.09907
3	58,968	6,166	10.5%	16.82	22.57	-0.095178
4	58,969	6,787	11.5%	22.57	39.99	-0.202948

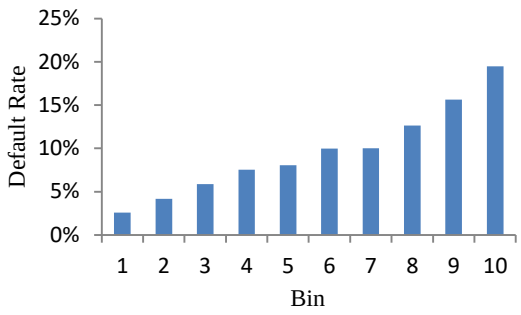
Table 2: Debt to Income - WoE Grouping

As a final step, a Chi-Square test was conducted in order to control whether all of the fifteen remaining features were significant with respect to the target variable, which was indeed the case. The graphs below display the WoE and default rate of each feature based on the "tailor-made" binning that we applied as discussed above.

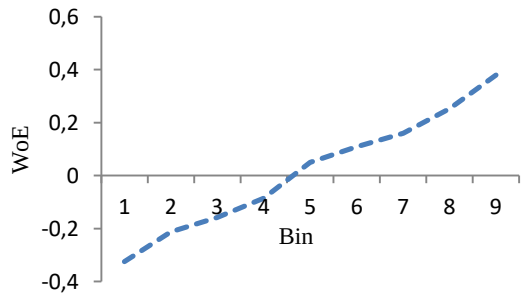
(Numeric) Interest Rate - WoE



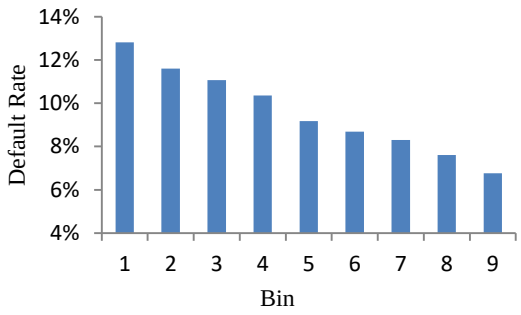
(Numeric) Interest Rate - Default Rate



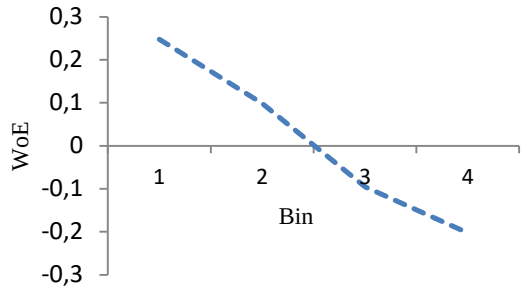
(Numeric) Annual Income - WoE



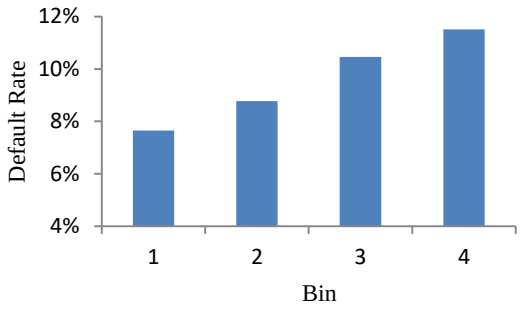
(Numeric) Annual Income - Default Rate



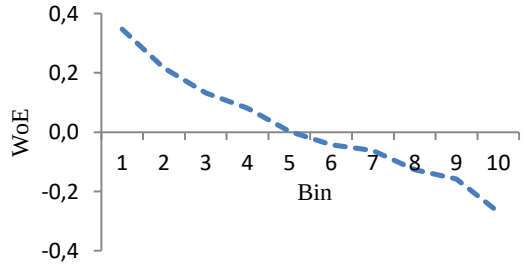
(Numeric) Debt to Income - WoE



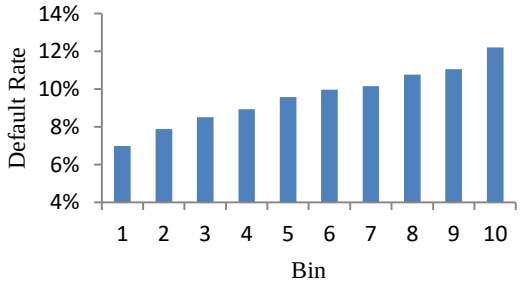
(Numeric) Debt to Income - Default Rate



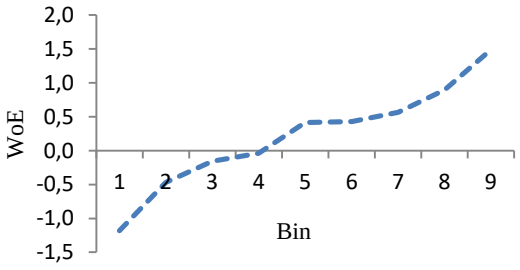
(Numeric) Revolving Line Utilization Rate - WoE



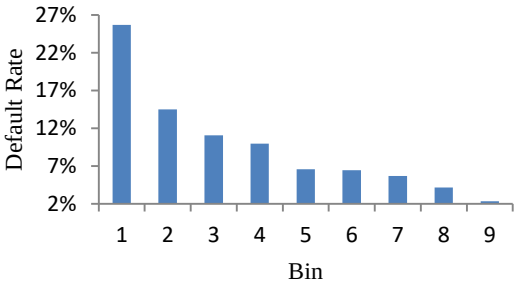
(Numeric) Revolving Line Utilization Rate - Default Rate



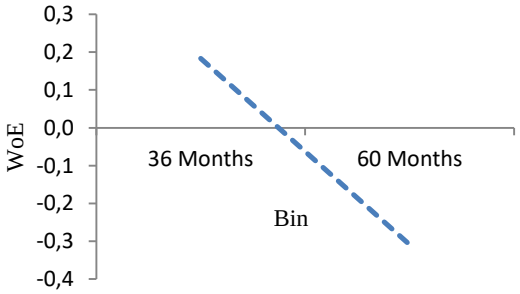
(Numeric) Payment Received for Funded Amount



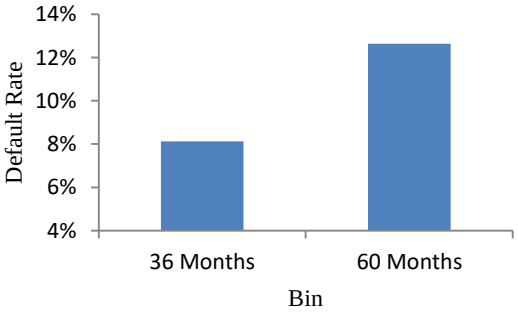
(Numeric) Payment Received for Funded Amount- Default Rate



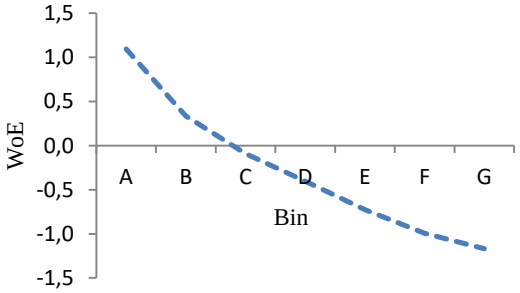
Term - WoE



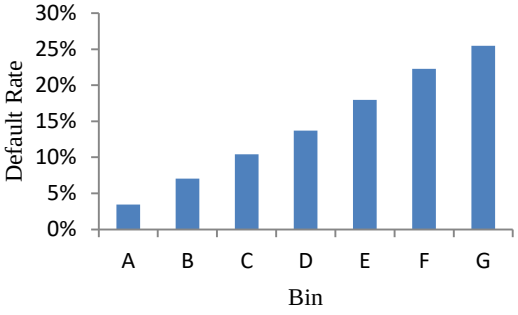
Term - Default Rate



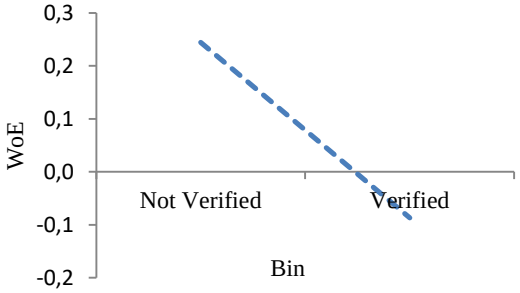
LC Grade - WoE



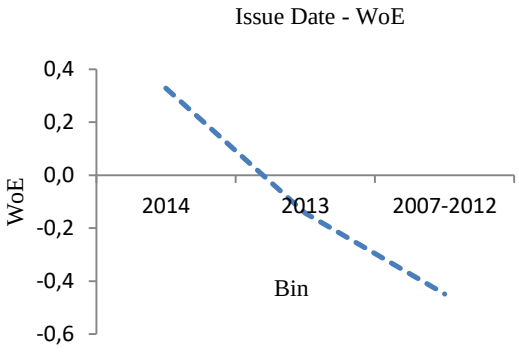
LC Grade - Default Rate

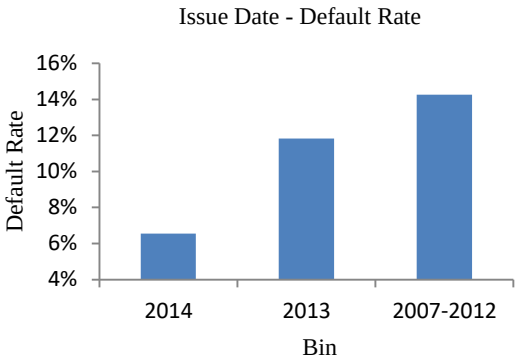


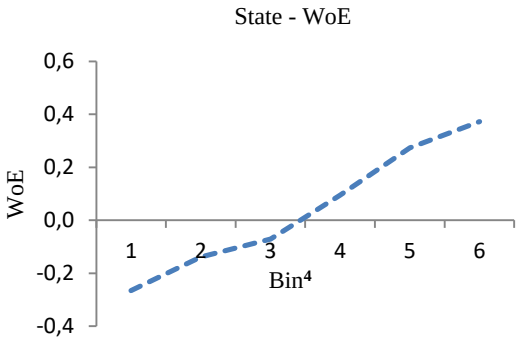
Verification Status - WoE

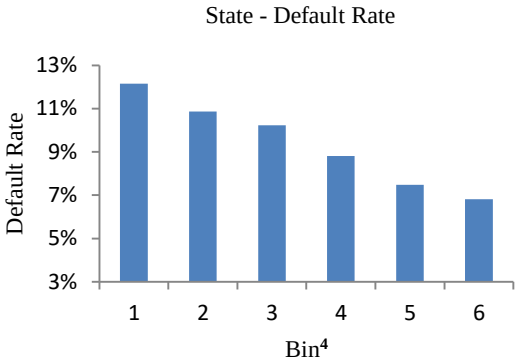


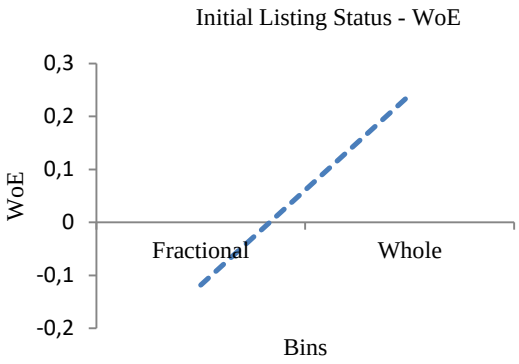
Verification Status - Default Rate

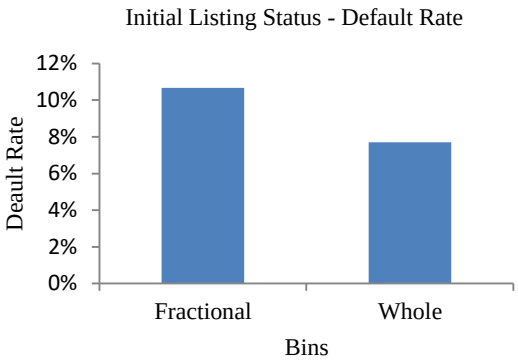


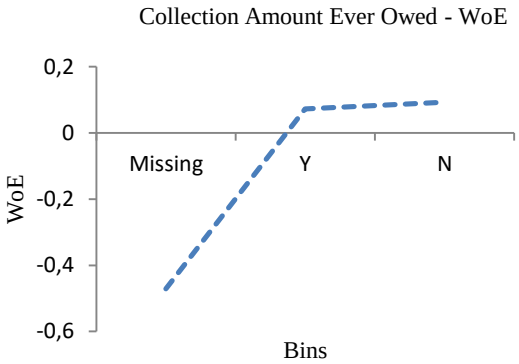




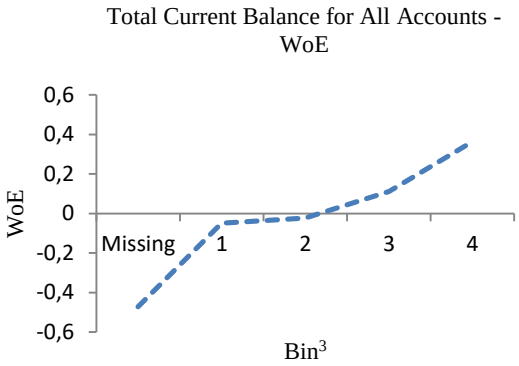


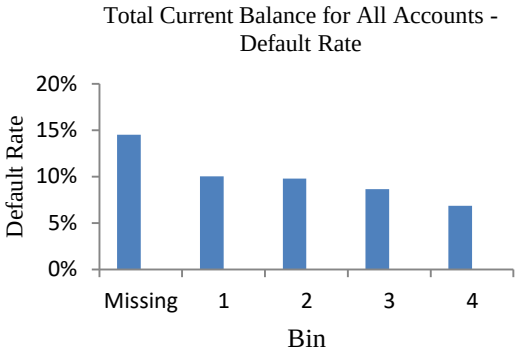


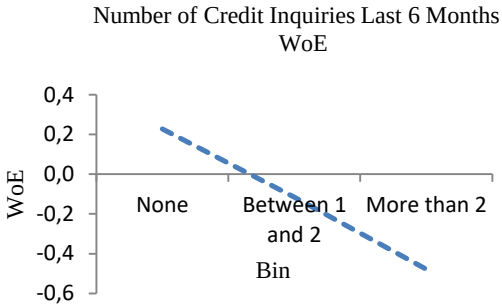


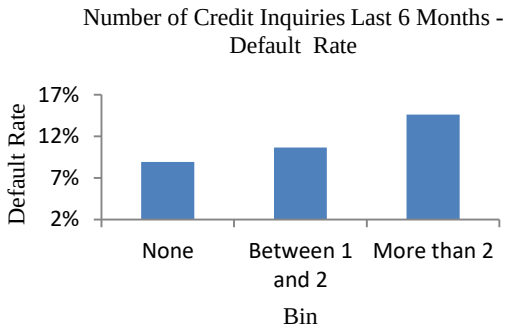


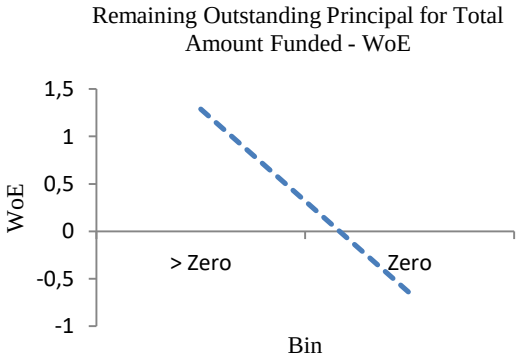
[3] The following bins include the acronyms of the U.S. states. For example, "NE" stands for Nebraska, "IA" for Iowa etc.
Bin 1: NE,IA,ID,NV,AL,ND;
Bin 2:NM,MO,NC,LA,FL;
Bin 3: SD,MD,MI,NJ,NY,KY,OK,OH,AR,PA,MT,VA,UT,RI,IN,DE,TN;
Bin 4: MN,WI,MA,AZ,GA,IL,WA,TX,KS,CA,VT,SC,CT,HI;
Bin 5:OR,MS,CO,NH**;**
Bin 6: WV,WY,AK,DC

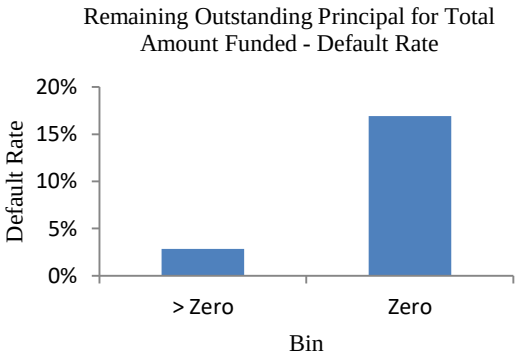












3. The Probability of Default Model

The task of this paper is to address the borrower's default prediction problem via estimating the Probability of Default (PD) of each individual loan. The approach we propose in this article consists of a *regression problem*, i.e., we aim to predict the value of a depended variable (the PD) via modeling its relationship with one or more independent variables (the features).

We chose Logistic Regression (LR) as our preferred modeling method because: it is well suited for such type of PD problems since it limits the output in the (0,1) space (the target is of binary type), it is straight forward to implement, it has low computational cost and it is widely used in the banking sector for such type of problems.

Moreover, the majority of the above cited papers investigating machine learning methods for credit risk addresses the same method, so we wanted to fit our results in that setting. The PD is computed as:

$$PD = \frac{1}{1+ e^{-(a+\sum_i b_i WoE_{Feature_i})}}, i \leq 15 \qquad (3)$$

Here the regression coefficients a and b_i are estimated through the non-linear least square method while the WoE of each feature is calculated as described in the previous section.

In order to identify the optimal feature set to be used for fitting the model into the data, we propose to leverage the Forward and Backward Sequential Feature Selection (SFS) approaches.

To provide the mathematical background of the SFS approach, let us consider a feature set $X = \{x_1, x_2, \dots, x_N\}$; we want to find a subset $Y_M = \{x_{i1}, x_{i2}, \dots, x_{iM}\}$, with $M < N$, that optimizes an objective function $J(Y)$, usually associated with the predictive power

[4] **Bin 1**: Balance <=109,000; **Bin 2**: 109,000< Balance <=188,000**; Bin 3**: 188,000< Balance <=285,000; **Bin 4**: Balance>285,000

or "goodness of fit" of a model. Thus, the SFS requires a search strategy to select candidate feature subsets and an objective function that evaluates them and returns a feedback signal. The feedback signal is processed by the search strategy which results in adjusting the feature selection accordingly.
Sequential algorithms add or remove features successively.
The Forward and Backward selection fall in this category. Specifically, the Sequential Forward algorithm starts from the empty set and sequentially add the feature x^+ that results in the highest objective function $J(Y_k + x^+)$, when combined with the features $J(Y_k)$, that have already been selected. In scheme:

1. Start with an empty set $Y_0 = \emptyset$
2. Select the next best feature $x^+ = \underset{x \, not \, in \, Y_k}{argmax} [J(Y_k \cup \{x\})]$
3. Update $Y_{k+1} \leftarrow Y_k + x^+, \; k \leftarrow k + 1$
4. Go to 2

The Backward Selection works in the opposite direction of the Forward one, starting from the full set, sequentially remove the feature x^- that results in the smallest decrease in the value of the objective function $J(Y_k - x^-)$. In scheme:

1. Start with the full set $Y_0 = X$
2. Remove the worst feature $x^- = \underset{x \; in \, Y_k}{argmax} [J(Y_k \setminus \{x\})]$
3. Update $Y_{k+1} = Y_k \leftarrow x^-, \; k \leftarrow k + 1$
4. Go to 2

At this stage, we need to choose an objective function for our modeling purposes. We decided to use the Area Under the Curve (AUC), since it is a robust performance measure for a large number of classifiers (including LR) and it is extensively used in the literature related to ML methods applied to credit risk (see for example (Tasche, 2008), (Tang & Chi, 2005), (Fantazzini & Figini, 2009), (Kruppa, Schwarz, Arminger, & Ziegler, 2013), (Addo, Guegan, & Hassani, 2018)).
The AUC is a performance measurement statistic describing the strength of the classifier in terms of assigning a lower PD to a true random performing observation than a true random defaulted observation.
In general, the performance of a classifier like the AUC can be described through a confusion matrix of the following form:

		Observed	
		Default	**Performing**
Predicted	**Default**	True Positive	False Positive
	Performing	False Negative	True Negative

Table 3: Confusion Matrix

Here, True Positive (TP) represents the number of obligors that the model classified as in default and were actual defaults, False Positive (FP) represents the number of borrowers that the model classified as in default but were in performing status, False Negative (FN) represents the number of obligors that the model classified as performing but were in default status and finally, True Negative (TN) represents the number of obligors that the model classified as performing and were in performing status. Starting from the values contained in the confusion matrix, one can calculate the following two rates:

$$True\ Positive\ Rate\ (TPR) = \frac{TP}{TP + FN} \quad (4)$$

$$False\ Positive\ Rate\ (FPR) = \frac{FP}{FP + TN} \quad (5)$$

Based on this set-up, the Receiver Operating Characteristic (ROC) curve is then defined as the set of all points (FPR, TPR) across all possible cut-off probability thresholds, i.e., the level representing the probability of a true prediction. The AUC statistic, is then simply defined as the area between the FPR axes and the ROC curve.
Having introduced the mechanics of the SFS algorithm and our preferred performance measurement statistic, the next step in our process is to compare the Forward and Backward algorithms in terms of AUC by implementing five-fold cross validation.
Since the cross-validation method segments the dataset into 5 blocks where each time one is used for testing while the remaining four for training the model, it is natural that an average AUC value per feature set is used as a comparison measure between the two algorithms.
Considering that the higher the average AUC, the better the model performance, it follows that the optimal feature set will then be the one implied by the maximum average AUC across all iterations of both approaches.
In this paper, we define the "best" SFS approach as the one producing the maximum AUC. Once we have identified the features that producing the maximum AUC for the LR, we decided to further investigate the default prediction problem using also a non-linear classifier.
The LR is in fact a classification technique based on linear algorithm (see (Kuhn & Johnson, 2013) and (Wendler & Griottrup, 2016)).

Classifiers using non-linear algorithm are as an example the support vector machine (SVM), artificial neural network (ANN), k-nearest neighbor (k-NN), naïve Bayes (NB), random forest (RF) and Bayesian network (B-Net) (see (Breiman, 2001), (Karatzoglou, Meyer, & Hornik, 2006), (Kuhn & Johnson, 2013), (Wendler & Griottrup, 2016) and (Arora & Kaur, 2020)). Among the non-linear classifiers, we have referred to k-NN and RF.

The k-NN approach to classification is a relatively simple approach that is also completely nonparametric. The basic principle behind this method is that a given instance within a data set will generally exist in close proximity with other instances sharing similar properties.

Hence, additional information about an instance can be obtained by observing other instances that are close to it, that is, the Nearest Neighbors (NNs).

If the instances within a data set are tagged with a classification label, then the class of a new instance can be determined by observing the classes of its NNs.

The advantage of nearest-neighbor classification is its simplicity. There are only two choices the modeler must make: the number of neighbors *k* and the distance metric to be used.

Common choices of distance metrics include Euclidean distance, Mahalanobis distance, and city-block distance. The number of neighbors is usually selected by either cross-validation or by testing the quality of the classifier on a second, test data set.

The use of RF in classification problems oftentimes produces strong results in terms of predictive accuracy; however, it is a computationally intensive and complex non-parametric approach.

Technically, an RF is a supervised learning algorithm defined as a collection of decision trees, each one generated based on a random partition of the underlying dataset.

The output of an RF can be based either on "voting", i.e., each tree "votes" separately and the final prediction is then defined by the majority vote or by averaging the predictions (in terms of probability) across the trees. Given the non-parametric nature of the RF, any trends and patterns in the data are inferred during the fitting process automatically without any external intervention.

However, the specifications of fitting process are defined during the model design phase through a set of parameters (commonly referred to as "hyperparameters") which cannot be inferred a priori by the data, are not predetermined and subject to the modeler's discretion and experimental outputs.

Such parameters, for example, are the number of trees that the forest will contain, the splitting rules, the leaf node size etc. (see (Probst & Boulesteix, 2018)).

The advantage of an RF is that it does not require intensive data preprocessing and performs well in large datasets.

In light of the feature sets produced by the forward ("best") SFS approach, we have implemented the k-NN algorithm for k =3, 5 & 7 using the Euclidean distance (see (Sun & Huang, 2010)) and the RF algorithm by setting the underlying decision trees and leaf node size parameters as 100 and 0.5% of the size training dataset respectively (see (Song & Lu, 2015) and (Probst & Boulesteix, 2018)).

Then, we calculated the respective set of AUC values based on a 70% / 30% train - test split of the dataset (no cross validation applied) and finally compared the results between the Logistic Regression versus the k-NN and RF modeling methods.

The following schema provides an illustration of this process. Results are presented in Section 4.

	Logistic Regression			RF / k-NN	
Iteration	**Resulting Features**	**Average AUC**	**Iteration**	**Input Features**	**AUC'**
1	{Feature 1}	AUC_1	1	{Feature 1}	AUC'_1
2	{Feature 1, Feature 2}	AUC_2	2	{Feature 1, Feature 2}	AUC'_2
⋮	⋮	⋮	⋮	⋮	⋮
15	{Feature 1, …, Feature 15}	AUC_{15}	15	{Feature 1, …, Feature 15}	AUC'_{15}

Table 4: Methodological Schema

4. Results

In this section we present the numerical results obtained by fitting equation (3) in our dataset in view of the optimal feature set produced by the Forward SFS approach and the AUC values of the RF and k-NN algorithms when taking as inputs the resulting selected features of each iteration as defined by the Forward SFS.

Table 5 provides the model coefficients of the Logistic Regression model based on the forward feature selection when 9 features are chosen ("best" model) for a 70% / 30% train - test split.

In this table we are adding the p-value and the Variance Inflation Factor (VIF) (Ron Johnston, Jones, & Manley, 2018) computed as:

$$VIF_{variable_i} = \frac{1}{1 - R^2_{variable_i}} \quad (6)$$

where $R^2_{variable_i}$ is the R - squared value resulting from regressing the i^{th} variable against all other available variables. The VIF statistic is a measure of multi-collinearity among the features used in the regression. If present, it can negatively impact the regression coefficients (increasing their sensitivity to small changes in the data) and ultimately the reliability and stability of the output.

As can be seen, VIF and p-values are below the standard threshold hinting acceptable level of multicollinearity and all chosen features are statistically significant.[5]

Feature	Coefficient	Standard Error	Variance Inflation Factor	p-Value
Constant	-2.1902	0.0107	-	<0.001
Interest Rate (WoE)	-1.2245	0.0199	1.2517	<0.001
Annual Income (WoE)	0.7489	0.0484	1.2288	<0.001
Debt to Income (WoE)	-1.0357	0.059	1.1151	<0.001
Revolving Line Utilization Rate (WoE)	-0.1949	0.0591	1.2012	0.001
Payments Received to Date (WoE)	-1.5705	0.0146	1.2482	<0.001
Verified Status (WoE)	-2.2445	0.0719	1.1390	<0.001
Issue Date (WoE)	-0.4025	0.0334	1.3263	<0.001
State (WoE)	-0.8508	0.0787	1.0068	<0.001
Remaining Outstanding Principal (WoE)	-1.2818	0.0138	1.2892	<0.001

Table 5: Logistic Regression, SFS, 9 Features

Figure 4 shows that the AUC of the model described by Table 5 is **0.8603**, indicating an excellent predictive power.

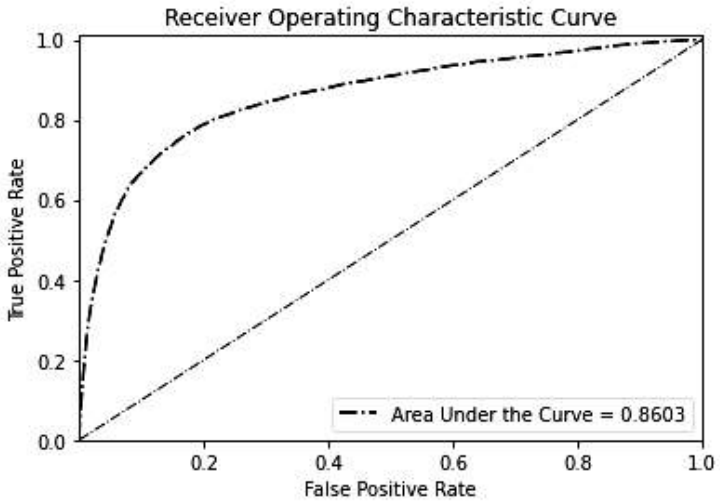


Figure 4: Logistic Regression, SFS, 9 Features, AUC

With respect to the comparison between the LR versus the k-NN and RF algorithms, as can be seen in Figure 6 the LR constantly outperformed the k-NN across all iterations while its performance was marginally inferior when compared to the RF approach. Note that, as discussed in the previous section, the LR values are the outcome of a 5-fold cross-validation approach.
Therefore, they represent average AUC values as opposed to the k-NN and RF ones which are derived based on a simple 70% / 30% train - test split of the dataset.
With respect to the comparison between the Forward and Backward SFS approaches employed to identify the optimal feature set for building the LR model, we note that the maximum average AUC value was attained at 9 features in view of both approaches, although the resulting chosen features were slightly different.
The "best" SFS was the forward one which produced an average AUC value of 0.8528 as opposed to 0.8521 produced by the backward approach (see Table 6 for 9 "Number of Features") although any AUC differences between them were found to be negligible.
Furthermore, observe that the AUC of the model described in Table 5 is, as expected, very close to both figures and within their respective 95% confidence intervals.
In the k-NN case, all three different models achieved their maximum when all fifteen features were used. Among them, the 7-NN model consistently outperformed its peers across all but in the case of two features in which the 5-NN model was found to be marginally better.
With respect to the RF case, the maximum AUC was produced when five features were used and it is approximately 0.73% higher than the value produced by the model as described in Table 5.
See Table 6 and figures 5 and 6 for details.

[5] A variable will remain in the model if it is statistically significant (p-value ≤ 0.05) and its Variance Inflation Factor (VIF) is less than 5 since as this is a commonly used cut-off threshold (see (Menard, 2002)).

Number of Features	LR (FSFS)	LR (BSFS)	RF	3NN	5NN	7 NN
1	0.7038	0.7038	0.6999	0.4795	0.4549	0.5440
2	0.8050	0.8050	0.8211	0.6597	0.7125	0.7120
3	0.8461	0.8461	0.8591	0.7040	0.7259	0.7688
4	0.8498	0.8498	0.8630	0.7273	0.7666	0.7932
5	0.8505	0.8504	**0.8666**	0.7638	0.7920	0.8046
6	0.8515	0.8510	0.8632	0.7504	0.7841	0.8068
7	0.8522	0.8515	0.8608	0.7500	0.7718	0.7771
8	0.8528	0.8519	0.8597	0.7298	0.7451	0.7499
9	**0.8528**	**0.8521**	***0.8641***	***0.7451***	***0.7697***	***0.7828***
10	0.8528	0.8519	0.8615	0.7510	0.7803	0.7967
11	0.8527	0.8516	0.8592	0.7552	0.7861	0.8011
12	0.8513	0.8512	0.8629	0.7686	0.7955	0.8102
13	0.8500	0.8505	0.8600	0.7712	0.8001	0.8152
14	0.8485	0.8492	0.8602	0.7731	0.8026	0.8181
15	0.8482	0.8482	0.8617	**0.7735**	**0.8030**	**0.8187**
Average	**0.8379**	**0.8376**	**0.8482**	**0.7268**	**0.7527**	**0.7733**
Relative AUC Difference	-	**-0.03%**	**1.23%**	**-13.26%**	**-10.17%**	**-7.71%**

Table 6: Model Comparison[6]

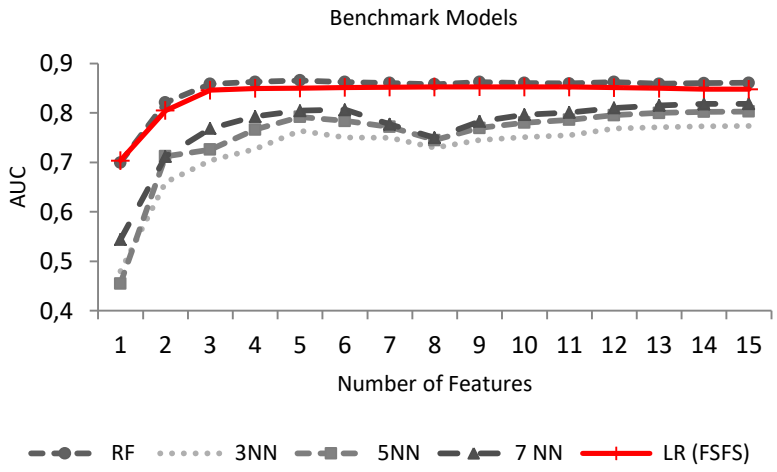


Figure 5: Benchmark Models

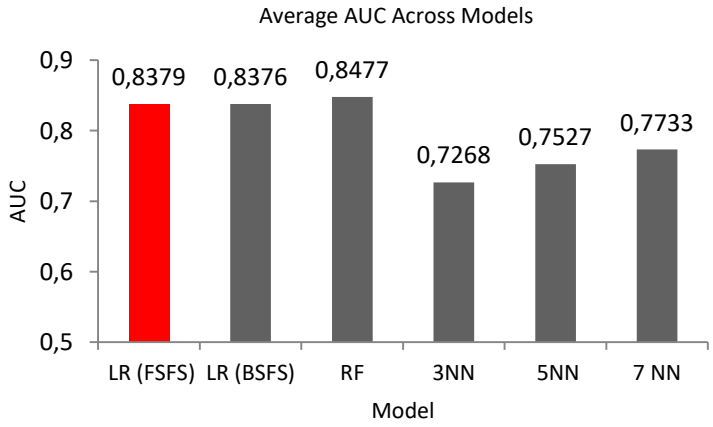


Figure 6: Average AUC Across Models

[6] Results are rounded to four decimal places. We stress in "Red" the maximum AUC value of each approach.

5. Conclusion

In this paper we develop and validate a PD model based on publicly available data provided by Lending Club between 2007 and 2014. Our main assumption is that borrower sharing the same homeownership status are likely to display similar risk behaviour. This assumption has been coupled with thorough data pre-processing, cleansing and the application of the Weight of Evidence Encoding, a powerful technique well suited for Logistic Regression problems since it introduces a monotonic relationship between the target variable and the predictors.

Our main assumption and the application of the WoE encoding make our paper unique and different from the cited literature and add value to our findings.

We have based the selection of the features for the LR model on the Sequential Feature Selection (SFS) algorithm. Specifically, in order to identify the optimal feature set, we have compared the Forward against the Backward method, finding out that the former achieves excellent model performance when 9 features are used.

The same result is true for the Backward approach however, the maximum AUC between the two was produced by the Forward one, which justifies our choice to proceed with it.

We note that any differences between these methods both in terms of the final feature set and the AUC statistic are minimal. Finally, we have compared the LR model against two non-linear classifiers, the k-NN algorithm (for k= 3, 5 & 7) and the RF across all iterations of the feature set as produced by the Forward SFS. The results indicate that on the one hand, the LR classifier coupled with WoE outperformed the k-NN while it displayed very similar, although inferior, predictive strength when compared to the RF approach.

Given the interpretability and simplicity that the LR method offers coupled with less computational effort and complexity as opposed to the RF one, we conclude that, if we develop a model per segment, the choice of an LR model leveraging the WoE technique is very well suited and produced excellent predictive power in comparison to the reference literature.

As future development of this work, the authors expect more data to become available in the near future; with a bigger dataset available, it would be of interest to investigate whether the LR and RF approaches will display significant differences in terms of AUC (convergence / divergence) or remain the same.

Moreover, the authors are planning to develop a set of models for the other two homeownership segments in order to cover all borrower types and compare the results in each segment.

6. Appendix

The following table provides the feature name, description and information regarding whether it is included or not in the model development phase.

It has been compiled based on the available information provided in the "LCDataDictionary" document accompanying the dataset and is also available to download on Kaggle.

In bold you can find the features that have been included in the development process.

#	Feature Name	Description	Exclusion Phase / Included in the Development Process
1	Accounts Now Delinq	The number of accounts on which the borrower is now delinquent.	Preprocessing & Cleansing
2	**Annual Income**	**The self-reported annual income provided by the borrower during registration.**	**Included in the Development Process**
3	Application Type	Indicates whether the loan is an individual application or a joint application with two co-borrowers.	Preprocessing & Cleansing
4	Collection Recovery Fee	Post charge off collection fee.	Preprocessing & Cleansing
5	Collections 12 Month excl Med	Number of collections in 12 months excluding medical collections.	Preprocessing & Cleansing
6	Delinquency (2yrs)	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years.	Feature Analysis
7	Description	Loan description provided by the borrower.	Preprocessing & Cleansing
8	**DTI**	**A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.**	**Included in the Development Process**
9	Earliest Credit Line	The month the borrower's earliest reported credit line was opened.	Preprocessing & Cleansing
10	Employment Length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.	Feature Analysis
11	Employment Title	The job title supplied by the Borrower when applying for the loan.	Preprocessing & Cleansing
12	Funded Amount	The total amount committed to that loan at that point in time.	Preprocessing & Cleansing
13	Funded Amount Inv	The total amount committed by investors for that loan at that point in time.	Preprocessing & Cleansing
14	**Grade**	**LC assigned loan grade.**	**Included in the Development Process**
15	Home Ownership	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER *(Used in population segmentation).*	Preprocessing & Cleansing
16	ID	A unique LC assigned ID for the loan listing.	Feature Analysis

17	**Initial Listing Status**	**The initial listing status of the loan. Possible values are – Whole (W), Fractional (F).**	**Included in the Development Process**
18	**Inquiries Last 6 Months**	**The number of inquiries in past 6 months (excluding auto and mortgage inquiries).**	**Included in the Development Process**
19	Installment	The monthly payment owed by the borrower if the loan originates.	Preprocessing & Cleansing
20	**Interest Rate**	**Interest Rate on the loan.**	**Included in the Development Process**
21	**Issue Date**	**The month which the loan was funded.**	**Included in the Development Process**
22	Last CR Pull Date	The most recent month LC pulled credit for this loan.	Preprocessing & Cleansing
23	Last Payment Amount	Last total payment amount received.	Preprocessing & Cleansing
24	Last Payment Date	Last month payment was received.	Preprocessing & Cleansing
25	Loan Amount	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.	Feature Analysis
26	Loan Status	Current status of the loan *(Used in the construction of the target variable)*.	Preprocessing & Cleansing
27	Member ID	A unique LC assigned Id for the borrower member.	Feature Analysis
28	Months since Last Delinquency	The number of months since the borrower's last delinquency.	Preprocessing & Cleansing
29	Months Since Last major derogatory	Months since most recent 90-day or worse rating.	Preprocessing & Cleansing
30	Months since Last Record	The number of months since the last public record.	Preprocessing & Cleansing
31	Next Payment Date	Next scheduled payment date.	Preprocessing & Cleansing
32	Open Accounts	The number of open credit lines in the borrower's credit file.	Feature Analysis
33	Outstanding Principal Inv	Remaining outstanding principal for portion of total amount funded by investors.	Preprocessing & Cleansing
34	**Outstanding Principal**	Remaining outstanding principal for total amount funded.	Included in the Development Process
35	Payment Plan	Indicates if a payment plan has been put in place for the loan.	Preprocessing & Cleansing
36	Policy Code	Publicly available policy_code=1 New products not publicly available policy_code=2	Preprocessing & Cleansing
37	Public Records	Number of derogatory public records.	Feature Analysis
38	Purpose	A category provided by the borrower for the loan request.	Feature Analysis
39	Recoveries	Post charge off gross recovery.	Preprocessing & Cleansing
40	Revolving Balance	Total credit revolving balance.	Feature Analysis
41	**Revolving Utilization**	**Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.**	**Included in the Development Process**
42	**State**	**The state provided by the borrower in the loan application.**	**Included in the Development Process**
43	Sub – Grade	LC assigned loan sub-grade.	Preprocessing & Cleansing
44	**Term**	**The number of payments on the loan. Values are in months and can be either 36 or 60.**	**Included in the Development Process**
45	Title	The loan title provided by the borrower.	Preprocessing & Cleansing
46	**Tot Collection Amount**	**Total collection amounts ever owed.**	**Included in the Development Process**
47	Total Accounts	The total number of credit lines currently in the borrower's credit file.	Preprocessing & Cleansing
48	**Total Current Balance**	**Total current balance of all accounts.**	**Included in the Development Process**
49	**Total Payments**	**Payments received to date for total amount funded.**	**Included in the Development Process**
50	Total Payments Inv	Payments received to date for portion of total amount funded by investors.	Feature Analysis
51	Total Rec Interest	Interest received to date.	Feature Analysis
52	Total Rec Late Fee	Late fees received to date.	Preprocessing & Cleansing
53	Total Rec Principal	Principal received to date.	Preprocessing & Cleansing
54	Total Rev hi limit	Total revolving high credit/credit limit.	Preprocessing & Cleansing
55	URL	URL for the LC page with listing data.	Preprocessing & Cleansing
56	**Verification Status**	**Indicates if income was verified by LC, not verified, or if the income source was verified**	**Included in the Development Process**
57	Zip Code	The first 3 numbers of the zip code provided by the borrower in the loan application.	Preprocessing & Cleansing

Table 7: Variables Description

Bibliography

Addo, P., Guegan, D., & Hassani, B. (2018). Credit Risk Analysis Using Machine and Deep Learning Models. University Ca' Foscari of Venice, Dept. of Economics Research Paper Series No. 08/WP/2018.

Al-Jabery, K., Obafemi-Ajayi, T., Olbricht, G., & Wunsch, D. C. (2020). Computational learning approaches to data analytics in biomedical applications. Academic Press.

Arora, N., & Kaur, P. (2020). A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment. Applied Soft Computing, 86, 105936.

Baesen, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Van-Thienen, J. (2003). Benchmarking state-of-art classification algorithm for credit scoring. Journal of the Operational Research Society, 54, 627-635.

Baesens, B., Roesch, D., & Harald, S. (2016). Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS. Wiley.

Berger, S., & Gleisner, F. (2009). Emergence of Financial Intermediaries in Electronic Markets: The Case of Online P2P Lending. BuR Business Research Journal, 39-65.

Breiman, L. (2001). Random Forest. Machine Learning, 45(1), 5-32.

Chang, S., Kim, S., & Kondo, G. (2015). Predicting default risk of lending club loans. Machine Learning, 1-5.

Dorfleitner, G., Priberny, C., Schuster, S., Stoiber, J., Weber, M., de Castro, I., & Kammler, J. (2016). Description-text related soft information in peer-to-peer lending – Evidence from two leading European platforms. Journal of Banking & Finance, 64, 169-187.

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

Fantazzini, D., & Figini, S. (2009). Random Survival Forests Models for SME Credit Risk Measurement. Methodology and Computing in Applied Probability volume, 11, 29-45.

Freedman, S., & Jin, G. (2014). The signaling value of online social networks: lessons from peer-to-peer lending. NBER Working Paper, 19820.

Fu,Y.(2017).Combination of random forest and neural network in social lending.Journal of Financial Risk Management,6,418-426

Iyer, R., Khwaja, A., Luttmer, E., & Shue, K. (2009). Screening in new credit markets: can individual lenders infer borrow creditworthiness in peer-to-peer lending? NBER Working Paper, 15252.

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.

Joenssen, D., & Bankhofer, U. (2012). Hot Deck Methods for Imputing Missing Data. Machine Learning and Data Mining in Pattern Recognition. MLDM 2012. Lecture Notes in Computer Science. 7376, p. 63-75. Berlin: Springer.

Karatzoglou, A., Meyer, D., & Hornik, K. (2006). Support vector algorithm in R. Journal of Statistical Software, 15, 1-28.

Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. Expert Systems with Applications, 40(13), 5125-5131.

Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling . Springer.

Kumar, V., Natarajan, S., Keerthana, S., Chinmayi, K., & Lakshmi, N. (2016). Credit risk analysis in peer-to-peer lending system. Knowledge Engineering and Applications (ICKEA) IEEE International Conference, (p. 193-196).

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

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

Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in social lending via ramdom forests. Expert Systems with Applications, 42(10), 4621-4631.

Menard, S. (2002). Applied Logistic Regression Analysis. Sage University.

Michels, J. (2012). Do Unverifiable Disclosures Matter? Evidence from Peer-to-Peer Lending. The Accounting Review, 87(4).

Namvar, E. (2013). An introduction to peer to peer loans as investments. Journal of Investment Management, 12, 1-18.

Probst, P., & Boulesteix, A. (2018). To Tune or Not to Tune the Number of Trees in Random Forest. Journal of Machine Learning Research, 18, 1-18.

Ron Johnston, R., Jones, K., & Manley, D. (2018). Confounding and collinearity in regression analysis: a cautionary tale and an alternative procedure, illustrated by studies of British voting behaviour. Quality and Quantity, 52, 1957–1976.

Serrano-Cinca, C., Gutierrez-Nieto, B., & Lopez-Palacios, L. (2015). Determinants of default in P2P lending. PLoS ONE,10, 1-22.

Siddiqi, N. (2017). Intelligent Credit Scoring: Building and Implementing Better Credit Risk Scorecards, 2nd Edition. Wiley.

Song, Y. Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130-135.

Sun, S., & Huang, R. (2010). An adaptive k-nearest neighbor algorithm. 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, 91-94.

Tang, T., & Chi, L. (2005). Predicting multilateral trade credit risks: comparisons of Logit and FuzzyLogic models using ROC curve analysis. Expert Systems with Applications, 28(3), 547-556.

Tasche, D. (2008). Validation of internal rating systems and PD estimates. The Analytics of Risk Model Validation, 169-196.

Teply, P., & Polena, M. (2020). Best classification algorithms in peer-to-peer lending. The North American Journal of Economics and Finance, 51, 100904.

Tsai, K., Ramiah, S., & Singh, S. (2019). Peer Lending Risk Predictor. Stanford University CS229 Project Report.

Wei, Z., & Lin, M. (2016). Market mechanisms in online peer-to-peer lending. Manag. Sci., 63(12), 4236-4257.

Wendler, T., & Griottrup, S. (2016). Data Mining with SPSS modeler. Springer.

Wu, J. (2014). Loan default prediction using lending club data. http://www.wujiayu.me/assets/projects/loan-default-predictionJiayu-Wu.pdf.