Scientific Research Publishing

Journal of Financial Risk Management, 2023, 12, 28-49
https://www.scirp.org/journal/jfrm
ISSN Online: 2167-9541
ISSN Print: 2167-9533

Predictive Analysis of Default Risk in Peer-to-Peer Lending Platforms: Empirical Evidence from LendingClub

Rocheny Sifrain

Independent Researcher, Port-au-Prince, Haiti
Email: rsifrain@gmail.com

How to cite this paper: Sifrain, R. (2023). Predictive Analysis of Default Risk in Peer-to-Peer Lending Platforms: Empirical Evidence from LendingClub. *Journal of Financial Risk Management, 12,* 28-49.
https://doi.org/10.4236/jfrm.2023.121003

Received: December 26, 2022
Accepted: March 3, 2023
Published: March 6, 2023

Abstract

In recent years, the expansion of Fintech has speeded the development of the online peer-to-peer lending market, offering a huge opportunity for investment by directly connecting borrowers to lenders, without traditional financial intermediaries. This innovative approach is though accompanied by increasing default risk since the information asymmetry tends to rise with online businesses. This paper aimed to predict the probability of default of the borrower, using data from the LendingClub, the leading American online peer-to-peer lending platform. For this purpose, three machine learning methods were employed: logistic regression, random forest and neural network. Prior to the scoring models building, the LendingClub model was assessed, using the grades attributed to the borrowers in the dataset. The results indicated that the LendingClub model showed low performance with an AUC of 0.67, whereas the logistic regression (0.9), the random forest (0.9) and the neural network (0.93) displayed better predictive power. It stands out that the neural network classifier outperformed the other models with the highest AUC. No difference was noted in their respective accuracy value which was 0.9. Besides, in order to enhance their investment decision, investors might take into consideration the relationship between some variables and the likelihood of default. For instance, the higher the loan amounts, the higher the likelihood of default. The higher the debt to income, the higher the likelihood of default. While the higher the annual income, the lower the probability of default. The probability of default has a tendency to decline as the number of total open accounts rises.

Keywords

Peer-to-Peer Lending, Default Risk, LendingClub, Machine Learning Methods, Predictive Analysis

1. Introduction

During the past years, the rapid growth of the financial technology (Fintech) has accelerated the emergence of the modern, online peer-to-peer lending (P2P) or social lending. Fintech is technology used to improve the delivery of financial services. Peer-to-peer lending is one type of crowdfunding. The latter is defined as a method of raising capital through the collective effort of friends, family, customers and individual investors. This method draws on the collective efforts of a large pool of individuals—primarily online via social media and crowdfunding platforms—and leverages their networks for greater reach and exposure (Jahangir, 2020). Peer-to-peer lending is then described as a new decentralized model for lenders to lend money and for potential borrowers to access funds. In this approach, individuals, i.e. lenders and borrowers, work in partnership on common online platforms without the use of traditional financial intermediaries.

In P2P lending, borrowers are seeking better terms than they can have through their local banks, while individual investors are expecting to earn higher rates of return than the ones offered by traditional financial intermediaries. This can be explained by the fact that social lending platforms have the opportunities to cut operating costs compared to traditional financial institutions, such as banks. Unlike banks, this innovative approach doesn't need a physical branch to be purchased or built and maintained. In addition, there is no need to staff the operation with employees, which implies paying multiple salaries, as well as related employee social benefits. The acquisition and maintenance of costly equipment are not either necessary. Social lending platforms are generally more efficient than traditional banks. Thus, P2P lending is not only known as a suitable way to connect individual lenders to borrowers, but also as a potentially lucrative investment opportunity.

Online P2P lending began in 2005 in the United Kingdom, with the launch of Zopa. The achievements of the first ever British P2P lending platform in providing people with access to loans and investments led to the spread of the alternative lending method across the world, as a response to the failure of banks to lend to individuals. The six top world leading P2P lending platforms in 2022 are: Upstart, LendingClub, Prosper, Solo Funds and Kiva based in the United States, and Funding Circle, based in the United Kingdom (Serio, 2022). Founded in 2007 and headquartered in San Francisco, LendingClub might be the most recognizable P2P lending platform with over $70 billion borrowed as of October 2022 (https://www.lendingclub.com/). It offers a range of diverse types of financing, including personal loan, business loan and auto loan refinancing. Their personal loans range from $1000 to $40,000, with annual percentage rate (APR) varying from 8.30% to 36.00% and terms from 3 to 5 years. The platform applies a fixed interest rate and its turnaround time is 2 business days. Their business loans range from $5000 to $500,000, with repayment terms varying from 1 year to 5 years. Their APR is not disclosed.

Online P2P lending platforms offer a huge opportunity for investment by di-

rectly connecting borrowers to lenders. These provide the lending market with speed, efficiency and ease of access to data and their processing (Pokorná & Sponer, 2016). However, as an emerging method of loan financing, online P2P has become progressively challenging, with regard to risk exposure, because of its lack of comprehensive mechanism of risk control and early warning (Ma, Hou, & Zhang, 2021). In the last years, many P2P lending platforms have coped with bad loan portfolio quality, and many of them have not been able to earn profits, because of their high loan default rates. Other online P2P lending platforms have even failed. This is the case with many Chinese peer-to-peer lending platforms. As a result, millions of investors have lost their life savings after founders liquidated the platform or made off with their money (He & Li, 2020).

However, in order to minimize their risk exposure, lots of online P2P lending platforms claim to have established an effective credit risk management system that could not only help control their default rates and reduce cost operations, but also boost investors' confidence. One of the key components of their credit risk management is the usage of credit scoring models with machine learning methods that would be effective in predicting the probability of borrower default. For instance, LendingClub uses FICO 8 and VantageScore 2.0, in addition to a proprietary scoring system, to assess the credit risk of borrowers (Millerbernd & Choudhuri-Wade, 2022).

In this paper, we focus on the evaluation of predictive capacity power of credit scoring models in online P2P lending platforms, using LendingClub's dataset, which is available on their website, applying statistical performance metrics, including the Kolmogorov Smirnov (KS) test. In addition, the paper proposes the development of three new statistical models, using machine learning methods, with logistic regression, random forest, and neural network in response to the poor performance of the credit scoring model in use at LendingClub. The construction of the models allows comparison and helps to select the one that would contribute to enhancing the applicant credit risk analysis of LendingClub. This paper is different from previous studies, in that it proceeds to the assessment of the existing model used by the online P2P lending platform under study before starting to build new credit scoring models.

The remainder of this paper is organized as follows: Section 2 reviews the theory and the relevant literature related to our research topic; Section 3 describes the research methods; Section 4 describes the data used to carry out our statistical analysis; Section 5 presents and discusses the main findings of the research; and Section 6 concludes this study.

2. Theoretical and Literature Review

2.1. Modern Portfolio Theory

This article uses the Modern Portfolio Theory (MPT), which is one of the most considerable economic theories in finance and investment. Its application is mostly extended in portfolio and risk management. This theory was developed

by Harry Markowitz (1952). The MPT is known as an investment theory that enables investors to select and construct assets portfolio that maximizes expected return for a given level of risk. The theory is based on the assumption that investors are risk-averse. For a given level of expected return, they always seek the least risky portfolio. The selection and the building of assets portfolio are then founded on maximizing the expected return while minimizing the investment risk (Fabozzi et al., 2002). According to the MPT, the investors optimize their portfolio by diversifying them (Pfaff, 2012). This can be realized by using different amounts of investments that are cautiously selected while taking into consideration how the investment is probable to be affected by the other elements of the portfolio rather than picking individual securities (Francis & Kim, 2013). Each security has its own risks, which are higher than that of a portfolio containing various securities (Pfaff, 2012). The risk component of MPT can be estimated, with different mathematical formulations, and condensed through the concept of diversification which aims to suitably select a weighted collection of investment assets that together show lower risk factors than investment in any individual asset. Diversification is the key concept of the MPT (Mangram, 2013).

Online P2P lending platforms work without the traditional financial intermediaries, which increases the asymmetry information between lenders and borrowers. As a result, the credit risk, which is the possibility of loss due to a borrower's defaulting on a loan, is more likely to rise. The investors whose purpose is to maximize their returns need to find mechanisms to predict the probability of default of the borrower, by constructing their loan portfolio. This can help them reduce the credit risk associated with their investment. Building sound credit scoring models using machine learning methods can help individual investors increase their portfolio by intelligently allocating funds to P2P lending marketplaces. Many platforms, including LendingClub, are considered leaders in employing credit scoring models to assess credit risk of the borrower.

2.2. Related Works

In the last decade, scholars researched on credit risk modelling in online peer-to-peer lending companies, using machine learning methods to predict the probability of defaults of borrowers. Data from LendingClub platform were commonly used. Serrano-Cinca et al. (2015) investigated the factors explaining loan default in P2P lending platforms, using a sample of 24,449 loans collected from LendingClub over the period of 2008-2014. Their univariate means tests and survival analysis indicated that loan purpose, current housing situation, indebtedness, annual income and credit history are the factors influencing loan default. In addition, the authors built a statistical model with logistic regression technique to predict the probability of default of the borrower. The results of the model showed that the subgrade assigned by LendingClub, based on FICO credit score and other attributes, is the most significant factor. But the accuracy of the model is enhanced by adding other information, mainly the debt level of the borrower. The paper concluded that the usage of mathematical models, including means

test, survival analysis and logistic regression, can enhance loan selection by individual investors.

Fu (2017) investigated on combination of random forests and neural networks in online peer-to-peer lending to predict the borrower's default, in order to prevent the risks in LendingClub. The results indicated that the method used by the paper outperformed the LendingClub good borrowers' grades. In their research paper, Vinod Kumar et al. (2016) analyzed the credit risk in P2P lending system of LendingClub, using a sample of 235,629 loan applications, from 2013 to 2015. Machine learning methods (decision tree, random forest, and bagging) and preprocessing techniques were employed to explore, analyze and determine the most significant factors in predicting the default risk. The output of their study showed that random forest predictor is better in identifying the defaults, while decision tree is more powerful in finding good credits. The study concluded that overall return on investment will be high as the model identifies most of the good loans at the same time as identifying potential defaults.

Namvar et al. (2018) invest aged on credit risk prediction in an imbalanced social lending environment, using a dataset of 66,376 loan applications of LendingClub, over the period of 2016-2017. To achieve this goal, their paper displayed an empirical comparison of different combinations of classifiers and resampling methods within a novel risk assessment methodology that integrates imbalanced data. The credit predictions from each combination are assessed with a G-mean measure to avoid bias towards the majority class. The study concluded that combining random forest and random under-sampling may be an effective strategy for calculating the credit risk associated with loan applicants in online peer-to-peer lending platforms.

Authors like Wan et al. (2019) researched on influencing factors of peer-to-peer network loan prepayment risk, through cox proportional hazards, using data collected from LendingClub. The study used a dataset of 655,007 personal loans, collected from September 2016 to March 2018. The results indicated that the loan interest rate is the most significant variable of early repayment risk. The protection factors include the loan amount, the verification status, and so forth. The hazard factors are the number of inquiries in the past 6 months, the number of open credit times in the borrower's credit life, the number of installment accounts open in the past 12 months, and so on.

Hou (2020) researched on identifying and predicting the online P2P borrower default, using recursive feature elimination (RFE) method to select key variables, combined with the classification model, including Logistic regression, CART decision tree and BP neural network, to predict the borrower's default behavior. This research employed a sample of 122,703 loans provided by LendingClub, in the third quarter of 2017. The study results indicated that the borrower's latest repayment amount, loan amount and loan interest rate have a great impact on the borrower's default status. Thus, the recursive feature elimination method can screen the key variables influencing the borrower default. In addition, the classification model has high accuracy, suggesting that it has significant classification

effect.

Chen et al. (2021) used an imbalanced dataset containing 269,668 loans collected from LendingClub, from 2007 to 2015, to predict default risk in online P2P lending. Their paper employed several machine leaning schemes, including random forest, logistic regression and neural network. Besides, re-sampling and cos-sensitive mechanisms to process imbalanced dataset are utilized. The research results showed that random under-sampling displayed the best performance in different classifiers. Thus, the proposed scheme can successfully increase the prediction accuracy for default risk.

More recently, Chang et al. (2022) used LendingClub data to build P2P lending credit scoring models, based on machine learning and artificial neural networks The methods used include logistic regression, support vector machine, decision tree, random forest, XGBoost, LightGBM and 2-layer neural networks. Through a performance comparison, the research results showed that the GBDT methods, including XGBoost and LightGBM, are the most suitable P2P credit-scoring models, outperforming 2-layer neural networks and the traditional approach of logistic regression. In addition, XGBoost had the best performance, with accuracy of around 88%.

From the previous works presented above, we can observe that the focus was basically on building credit scoring models to predict the likelihood of borrower's default, when it comes to assess the creditworthiness of loan applicants. This paper instead aimed to assess the predictive power of the LendingClub's model using borrowers' grades which correspond to their default probability through statistic metrics, prior to constructing credit scoring models that should show better predictive power.

3. Research Methods

This section describes succinctly the different methods used to build the credit scoring models, i.e. the logistic regression, the random forest and neural network. In addition, it presents the statistical metrics employed to evaluate the models performance (KS statistic, ROC Curve, AUC and Accuracy).

3.1. Logistic Regression

The dependent variable of our study is a binary variable, which takes the value 1 (default or bad loans) or the value 0 (non-default or good loans). In statistics, this kind of issue is usually solved by using probit or logit models, assuming that the probability of event occurrence follows a certain probability distribution. The probit model is typically employed if the probability of event occurring follows the cumulative standard normal distribution. When the probability of event occurring obeys a logistic distribution, the logit model (or logistic model) is used (Aldrich & Nelson, 1984).

So as to predict the probability of occurrence of the event, which is the default in our research, we assume that the probability of default to be a linear combina-

tion of the independent variables:

$$P\left(y=1\mid x_1,\cdots,x_j\right)=G\left(\beta_0+\beta_1x_1+\cdots+\beta_jx_j+\epsilon\right)$$

where y ($y\in[0,1]$) is the dependent variable, $x_1,\cdots,x_j$ are the independent variables, ϵ is the error term, and G is the logistic cumulative distribution function that maps $x_1,\cdots,x_j$ to the real number space from 0 to 1, ensuring that the estimated value falls in [0, 1] as follows:

$$G(x)=\frac{e^x}{1+e^x}$$

Then, we define $Q(x)=p(y=1\mid x)$ and we obtain

$$Q(x)=\frac{e^{\beta_0+\beta_1x_1+\cdots+\beta_jx_j}}{1+e^{\beta_0+\beta_1x_1+\cdots+\beta_jx_j}}$$

Or

$$\ln\left(\frac{Q(x)}{1-Q(x)}\right)=\beta_0+\beta_1x_1+\cdots+\beta_jx_j$$

The coefficients are estimated by maximizing the log-likelihood function, which is

$$\log L(\beta)=\sum_{i=1}^{N}y_i\log\left(F\left(x_i'\beta\right)\right)+\sum_{i=1}^{N}\left(1-y_i\right)\log\left(1-F\left(x_i'\beta\right)\right)$$

where $F\left(x_i'\beta\right)=p\left(y=1\mid x_i;\beta\right)$.

3.2. Random Forest

The Random Forest algorithm is a supervised classification algorithm (Breiman, 2001). This is an ensemble of decision trees. It uses self-learning decision trees. These trees automatically define rules at each node, based on a training dataset. It creates a multitude of different models by bagging selection training data and randomly selection features. The Random Forest has in this case two layers of randomness. First, it employs a random sample of the training dataset (with replacement, i.e. a bootstrapped sample) for growing each individual decision tree. Second, it applies a random selection of the features (for instance, spectral bands) considered at each node to determine the best rule for splitting the data and finally determining a class label. The final result is decided by using majority voting (**Figure 1**).

The Random Forest seeks to minimize the heterogeneity of the two resulting subsets of the data created by the respective rule. Heterogeneity is then expressed as the Gini Impurity index and the rule which creates the least heterogeneous subsets of the data is used for the respective node.

This paper uses the methodology of Classification and Regression Tress (CART) to construct the trees of the Random Forest. A CART is a predictive algorithm applied in learning machine which explains how the values of a target variable can be predicted based on other values. CART algorithm is a binary decision tree using Gini Index for impurity measurement. If all of the elements are accurately

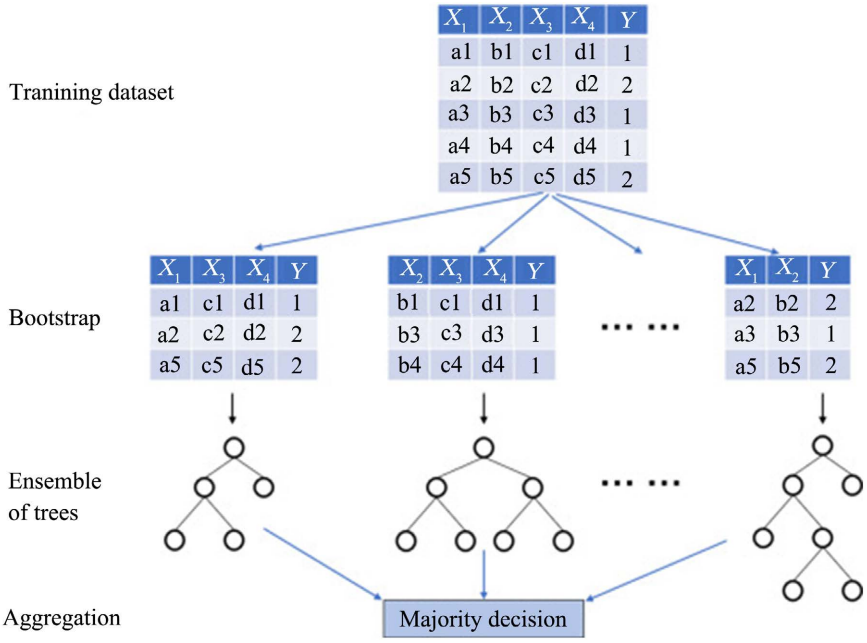

Figure 1. A schematic structure of random forest algorithm.

separated into different classes (an ideal scenario), the division is called pure. The Gini impurity is used to predict the likelihood that a randomly selected example would be incorrectly classified by a specific node. It is known as "impurity" metric, because it indicates how the model differs from a pure division.

The Gini impurity is given by the following formula (Breiman et al., 1984):

$$\text{Gini}(S) = 1 - \sum_{i=1}^{n} (p_i)^2$$

where S represents a dataset containing samples from n classes, p_i denotes the probability of samples belonging to class i at given node.

The Gini impurity indicates the probability of misclassifying an observation. Its values range from 0 to 1. The lower the Gini the better the split, i.e. the lower the likelihood of misclassification.

3.3. Neural Network

Neural Network (or Artificial Neural Network) is an information processing model inspired by biological neuron system. This is comprised of a huge number of highly interconnected processing elements known as the neuron to solve problems. It obeys the non-linear path and processes information in parallel throughout the nodes. Neural Network is known as a complex adaptive system, i.e. it has the capability to alternate its internal structure by adjusting weights of inputs (Navlani, 2019).

Figure 2 displays the three types of layers of neural networks. The input layer transmits different features and interacts with one or more hidden layers. The node is named neuron presenting an activation function. Every connection indicates a weight. The weight value is different from one to another. These weights and non-linear activation function produce complex relationships.

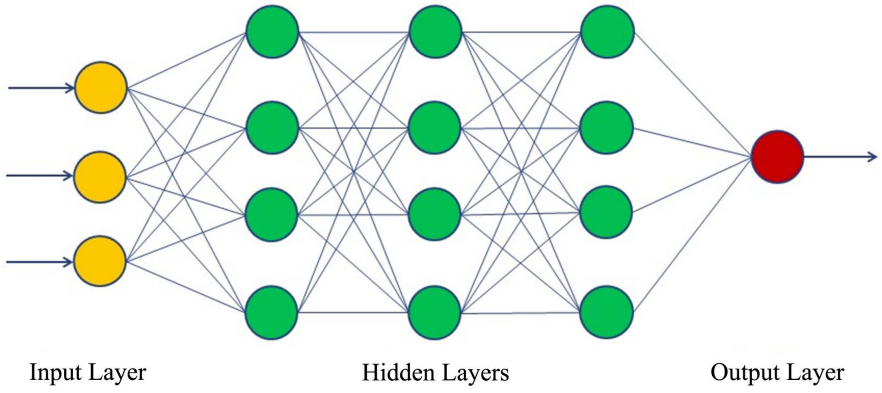

Figure 2. Demonstration of neural network classifier.

3.4. Models Performance Measures

3.4.1. Kolmogorov Smirnov (K-S) Test

The Kolmogorov Smirnov (K-S) test is a non-parametric test which measures a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution function of two samples. In credit risk modeling, the K-S statistic calculates the maximum vertical separation between two cumulative distributions (default and non-default), indicating the degree of discrimination between good and bad loans. The output of the test can be between 1 and 100, where the higher the K-S, the better the discrimination.

The empirical distribution function F_n for n i.i.d (independent and identically distributed) ordered observations X_i is given as follows:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{[-\infty,x]}(X_i)$$

where $I_{[-\infty,x]}(X_i)$ is the indicator function, equals to 1, if $X \le x_i$.

The K-S statistic for a given cumulative distribution function $F(X)$ is:

$$D_n = \sup_x \left|F_n(x) - F(x)\right|$$

where $\sup_x$ is the supremum of the set of distances.

3.4.2. Receiver Operating Characteristic (ROC)

A Receiver Operating Characteristic Curve (ROC curve) is a graph exhibiting the performance of a classification model as its discrimination threshold fluctuates. The ROC curve plots the True Positive Rate (TPR) or sensitivity against the False Positive Rate (FPR) or (1-specificity) at different threshold settings.

The True Positive Rate (TPR) is given as follows:

$$\text{TPR} = \frac{\text{True Positive}(\text{TP})}{\text{True Positive}(\text{TP}) + \text{False Negative}(\text{FN})}$$

The False Positive Rate (FPR) is obtained as follows:

$$\text{FPR} = \frac{\text{False Positive}(\text{FP})}{\text{False Positive}(\text{FP}) + \text{True Negative}(\text{TN})}$$

For the perfect model, the graph for the ROC curve moves through the upper left corner, where the share of the false positive outcomes equals zero. The closer is the curve to the upper left corner; the higher is the predictive power of the model. The diagonal line (line of no discrimination or random guess) indicates the bad.

The Area under the Curve (AUC) is a measure of a model's discriminatory power. It is employed as a summary of the ROC curve. The AUC is between 0.5 and 1. For a random model (useless model), the AUC is 0.5, for a perfect model, the AUC is 1. A model with greater power presents a larger AUC.

In general, AUC values are interpreted as follows (Abdou et al., 2016):

1) $0.5 \leq AUC < 0.6$ = fail;

2) $0.6 \leq AUC < 0.7$ = poor;

3) $0.7 \leq AUC < 0.8$ = fair;

4) $0.8 \leq AUC < 0.9$ = good;

5) $0.9 \leq AUC \leq 1.0$ = excellent.

3.4.3. Accuracy

Accuracy is one of the performance metric used to evaluate classification models, which represents the percentage of cases correctly classified. The higher the accuracy, the better the model performance. Mathematically, for a binary classifier, it is represented as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

4. Data

This section presents the data source and data cleaning process, including removing irrelevant and redundant variables, dealing with missing data and scaling, converting variables, variables selection and undersampling.

4.1. Data Source

This paper uses the accepted loan application data from 2007 to 2018 provided by the LendingClub platform. The full dataset contains 2,260,701 observations and 151 variables. The main variables include the borrower's personal information, loan characteristics, credit history and others. Since this study focuses only on the individual loans, all of the joint applications were excluded from the dataset. The loans with current as debt status and those in grace period were not considered. Likewise, when the data were collected in 2018, the loans disbursed after 2014 were not fully expired. It was then difficult to gauge them as good or bad loans. As a consequence, they were not included in the sample. Since the data before 2013 contain too many missing values, they were removed from the dataset. As a result, the study covers only the data collected from January 2013 to December 2014, corresponding to a total of 388,518 loans with 151 variables each.

This paper used the loan status as the reference variable to define default and

non-default. "Default", "Charged Off", "Does not meet the credit policy. Status: Charged Off", "Late (31 - 120 days)", "Does not meet the credit policy. Status: Late (31 - 120 days)" are considered as default, whereas "fully Paid" denotes non-default. The dataset of 388,518 loans contains 295,995 non-defaulted loans (82.6%) against 62,523 defaulted loans (17.4%).

4.2. Data Cleaning

4.2.1. Removing Redundant and Irrelevant Variables

Before selecting the final variables that were used in predictive modeling, we removed some redundant variables. We started with the variables that are not associated with loan information, for instance the borrower's membership ID in LendingClub, the descriptive variables including a paragraph indicating the loan description provided by the borrower, showing the reason the hardship plan was offered; the variables that were generated after the loan had been approved, including the date of the previous loan repayment; the credit variables marked by LendingClub. Overall, the variables presenting too many missing values were also removed.

4.2.2. Converting Variables

The categorical variables were converted into numerical ones, in order to make them suitable for the model training. First of all, in the dependent variable (Loan status), the "non-defaulted" category was set to 0 and the "defaulted" category to 1. For the variable Employment length representing the number of working years of the borrower, since it is a variable with sequential meaning, ordinal encoding was used to convert it, making this variable as orderly numbers (von Eye & Clogg, 1996). So, the borrowers with more than 10 years are set as 10, those with less than 1 year as 0 and the ones between 1 to 10 years as their numerical values in the dataset. Regarding the remainder of the categorical variables, including the loan purpose, housing ownership and others, we used one-hot encoding to convert them, indicating that each one of them was defined as independent binary variable with only two values 0 and 1 (Lantz, 2013).

4.2.3. Dealing with Missing Data and Scaling

Before processing the models training data, we dealt with the missing values. Overall, we processed the variables by replacing the missing information by the mean of that variable. In addition, the data were standardized, in order to make sure that each variable will only affect the prediction result proportionately, because with algorithms, such as logistic regression, which employ the mean square error as the loss function, the scale of the variables can easily affect the prediction performance, since the models tend to be sensitive to variables with large scales.

4.2.4. Variables Selection

After our analysis and studies upon the variables, using mainly bivariate analysis and correlation analysis, we selected 18 of them to be used as independent va-

riables for the predictive models. These are described in Table 1.

4.2.5. Undersampling

Our dataset contains 295,995 non-defaulted loans (82.6%) against 62,523 defaulted loans (17.4%). The non-defaulted represents then almost 5 times of the defaulted loans. So, there is imbalanced classification in our dataset, implying that the predictive model will overtrain the category with more samples and consider the category with few samples as noise, or even ignore it. Therefore, bias will be easily produced and poor predictive performance of the model will be observed. The reliability might be reduced with a fake high accuracy (Madasamy & Ramaswami, 2017). Oversampling and undersampling are the two most common methods used to deal with imbalanced classifications. According to Shelke et al. (2017), oversampling duplicates the minority sample, while undersampling picks part of the majority sample to attain the balance. Oversampling can induce overfitting and undersampling can eliminate important information regarding the overall pattern of the data (Abd Elrahman & Abraham, 2013). Each of the two approaches has its own strengths and weaknesses. Thus, it is not easy to conclude which of the two approaches is better or worse. In order to reduce the training time of the random forest and the neural network algorithms, this paper adopted the undersampling method. To do so, we randomly sampled the same amount of data for non-defaulted loans as defaulted loans, i.e. 62,523 loans. And the total defaulted loans remained the same (62,523 loans). The final size of the data for training (75%) and validation (25%) was 125,046 loans.

5. Empirical Results

5.1. Performance Evaluation of the LendingClub Model

Before proceeding to the development of the predictive models, this paper assessed the predictive power of the LendingClub model, measuring the ability of the model to correctly classify loans. In other words, this study analyzed the capability of the model to separate good loans from bad loans. The distribution of the grades assigned by LendingClub was analyzed (Table 2), based on the same definition of non-defaulted and defaulted loans adopted for the models building, as described previously. K-S and AUC are two statistical measures used for this assessment.

In Table 2, the column (% of non-defaulted) should present the following pattern: the percentage of non-defaulted loans has to decrease as the grades fall and rise as the grades increase. As observed, the column (% of non-defaulted) obeys the pattern as expected. From A to G, the percentage of non-defaulted loans decreases, varying between 94.8% and 56.6%. Similarly, the percentage of defaulted loans increases from A to G. As indicated in Table 2, the maximum K-S is 25.9% representing the maximum difference between the cumulative distribution of non-defaulted loans and defaulted loans. In addition, the Gini calculated for the LendingClub model is equal to 0.35, which is of average quality (Abdou et al., 2016). The AUC of 0.67 is obtained, using the following formula:

Table 1. Description of the independent variables.

Variable	Variable declaration	Definition	Category
		Socio-economic	
annual_inc	Annual income	The annual income provided by the borrower during registration.	Numeric
verification_status	Verification status	Indicates if income was verified by LendingClub, not verified, or if the income source was verified.	Categorical
emp_length	Employment length	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LendingClub loan, divided by the borrower's self-reported monthly income.	Numeric
home_ownership	Home ownership	The length of time in years that borrowers have been working for their current company. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.	Categorical
		The home ownership status provided by the borrower during registration or obtained from the credit report. Possible values are: Rent, Own, Mortgage, and Others.	Categorical
		Loan Characteristics	
purpose	Loan purpose	A category provided by the borrower for the loan request. 14 possible loan purposes: credit card, car, educational, debt consolidation, house, home improvement, major purchase, moving, medical, small business, renewable energy, vacation, wedding and others.	Categorical
loan_amnt	Loan amount	The listed amount of the loan applied for by the borrower.	Numeric
initial_list_sta	Initial list status	The initial listing status of the loan. Possible values are: "W" for Whole, "F" for Fractional.	Categorical
		Credit History	
open_acc	Open accounts	The number of open credit lines in the borrower's credit file.	Numeric
total_acc	Total accounts	The total number of credit lines currently in the borrower's credit file	Numeric
delinq_2 yrs	Delinquency 2 years	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years	Numeric
inq_last_6mths	Inquiries last 6 months	Number of credit inquiries in the past 6 months.	Numeric
revol_util	Revolving utilization	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.	Numeric
revol_bal	Revolving balance	Total credit revolving balance.	Numeric
pub_rec	Public records	Number of derogatory public records.	Numeric
chargeoff_within_12_mths	Charged-off within 12 months	Number of charge-offs within 12 months	Numeric
recoveries	Recoveries	Post charge off gross recovery	Numeric
collections_12_mths_ex_med	Collections within 12 months	Number of collections in 12 months excluding medical collections	Numeric

Table 2. Distribution of the grades of the borrowers according to the definition of the defaults.

Grade	Non-defaulted	Defaulted	Grand Total	% Non-defaulted	% Cum Non-defaulted	% Defaulted	% Cum Defaulted	K-S
A	50,779	2779	53,558	94.8%	100.0%	5.2%	100.0%	0.0%
B	92,771	11,326	104,097	89.1%	82.8%	10.9%	95.6%	12.7%
C	81,819	18,969	100,788	81.2%	51.5%	18.8%	77.4%	**25.9%**
D	44,629	15,449	60,078	74.3%	23.9%	25.7%	47.1%	23.2%
E	18,277	9155	27,432	66.6%	8.8%	33.4%	22.4%	13.6%
F	6358	3801	10,159	62.6%	2.6%	37.4%	7.7%	5.1%
G	1362	1044	2406	56.6%	0.5%	43.4%	1.7%	1.2%
Total	**295,995**	**62,523**	**358,518**	**82.6%**		**17.4%**		

Source: Author's own calculations in Excel.

Gini = (AUC × 2) − 1. This value indicates a poor performance, suggesting the low discriminatory power of the online lending platform under study (Abdou et al., 2016). Therefore, this paper aimed to build models that will outperform the LendingClub model.

5.2. Results and Discussion

5.2.1. Logistic Regression Model

We used the Information Value (IV) statistic to measure the importance of the independent variables. IV is a numerical value to quantify the predictive power of an independent variable x in capturing the binary dependent variable y. IV is known as a common screener for selecting predictive variables for binary logistic regression (Lund & Brotherton, 2013). The larger the IV, the more predictive is the independent variable. Figure 3 shows the importance of the variables importance regarding the definition of defaults and non-defaults adopted by this study.

As observed in Figure 3, the top five predictive variables are: recoveries, debt to income (dti), verification status, annual income and loan amount, with the highest information values.

Furthermore, we trained the logistic regression model using the logit link function in R with all 18 variables. The Akaike Information Criterion (AIC) was employed to select the best model. The best-fit model is the one with the lowest AIC. Its parameters are shown in Table 3.

As indicated in Table 3, out of 18 variables used 13 are retained by the logistic regression model. All of them are statistically significant, with the *p*-value (Pr (>|z|)) smaller than 0.05. The lower the *p*-value, the more significant are the variables. The coefficient estimate of most of variables is positive. This means an increase in them is associated with an increase in the probability of being defaulted. Whereas an increase in the variables with the negative coefficient estimate is associated with a decreased probability of being defaulted. The standard

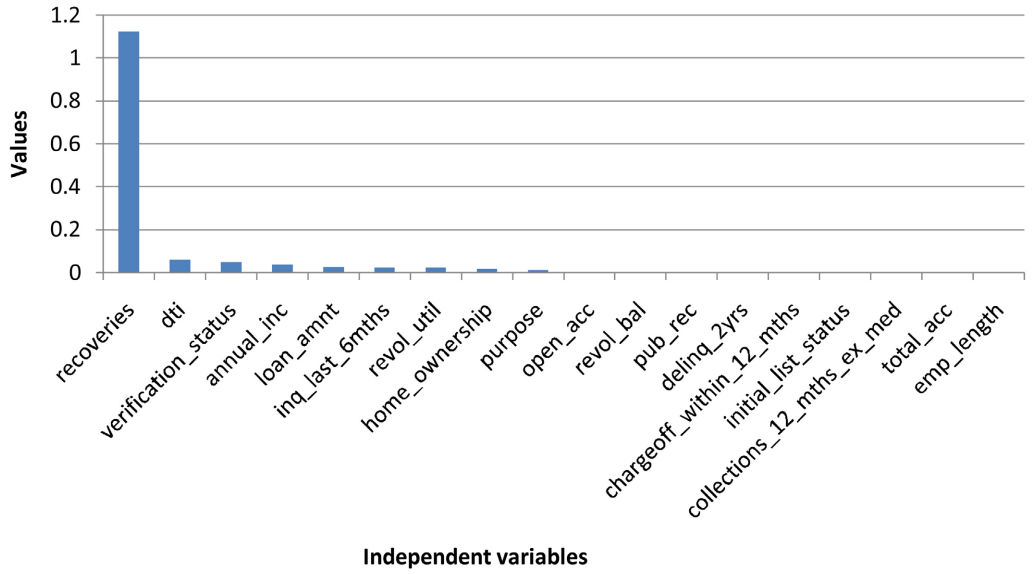

Figure 3. Variables importance according to the information value.

Table 3. Parameters of the selected logistic regression model.

Variable	Estimate	Std. Error	z value	Pr (>|z|)
Intercept	3.35E+04	3.06E+03	10.97	<2e−16
emp_length	1.90E−02	3.25E−03	5.85	4.91E−09
purpose	3.14E−02	6.32E−03	4.961	7.02E−07
home_ownership	1.65E−01	1.28E−02	12.852	<2e−16
annual_inc	−5.83E−01	2.59E−02	−22.486	<2e−16
dti	2.15E−01	1.31E−02	16.412	<2e−16
inq_last_6mths	1.28E−01	1.14E−02	11.21	<2e−16
open_acc	1.38E−01	1.61E−02	8.556	<2e−16
revol_bal	−1.08E−01	2.17E−02	−5.001	5.70E−07
revol_util	1.01E−01	1.29E−02	7.792	6.58e−15
total_acc	−5.75E−02	1.65E−02	−3.481	5.00E−04
recoveries	1.35E+05	1.23E+04	10.971	<2e−16
loan_amnt	3.33E−01	1.47E−02	22.696	<2e−16
pub_rec	7.47E−02	1.16E−02	6.459	1.05E−10

Source: Author's own calculation.

error (Std. Error) of the coefficients measures the precision of the coefficients. The smaller the standard error, the more precise the estimate.

5.2.2. Random Forest Model

We used 75% of the dataset to train the model and 25% for validation, which are

respectively 93,784 and 31,262 observations. The best random forest classifier is the model with 500 trees, 2 as the number of variables at each split. This is the best combination that produces the smallest OOB (Out of Bag) estimate of error rate (10.18%). So, the train dataset model accuracy is around 90%, indicating that around 10% of the total observations are misclassified.

Figure 4 shows the variable importance using the mean decrease in Gini coefficient. This is a measure of how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest. The higher the value of the mean decrease Gini score, the higher the importance of the variable in the model.

As shown in Figure 4, the top five important variables are recoveries, debt to income (dti), annual income, revolving balance, revolving utilization and loan amount. It is convenient to underscore the huge gap between recoveries and the other independent variables in terms of importance.

5.2.3. Neural Network Model

In order to fit the neural network model, we used the package nnet in R with 1 hidden layer containing 18 neurons. We run a maximum of 500 iterations, with logistic activation function. The model was trained using 75% of the retained sample.

We used Garson's algorithm to determine the variable importance for the neural network model (Figure 5).

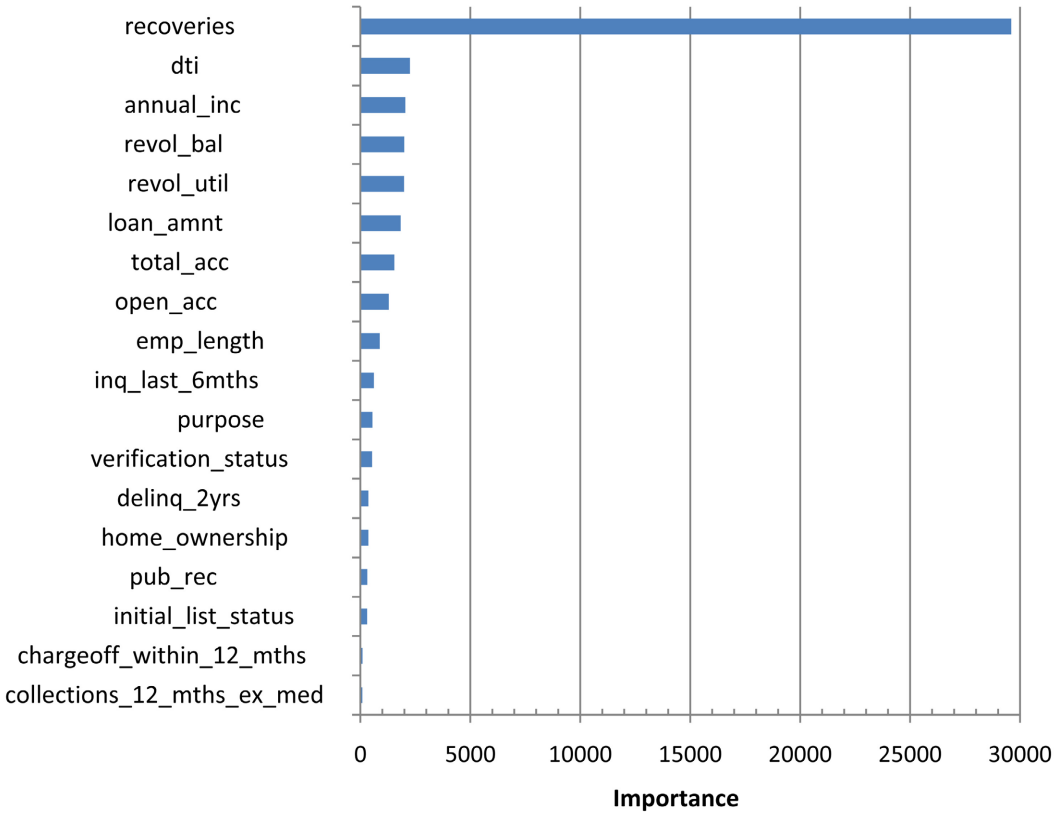

Figure 4. Variables importance according to the mean decrease Gini.

It stands out that the variable employment length has the lowest importance on the output, whereas the variable recoveries have the highest importance, followed by verification status, annual income. It is convenient to indicate that the variable recoveries remain the top variable for all of the three models used by this paper (logistic regression, random forest and neural nework).

5.2.4. Models Comparison

We evaluated the predictive power of our models using 25% of the dataset corresponding to 31,262 observations. We compared their results with two main measures: Receiver Operating Characteristic (ROC) and Confusion matrix.

As shown in **Table 4**, the neural network classifier appears to be the best model with the highest AUC value, 0.936, which indicates an excellent model according to Abdou et al. (2016); whereas the performance of the logistic regression and the one of the random forest are the same. However, all three models outperform the LendingClub model whose evaluation indicates an AUC of 0.67.

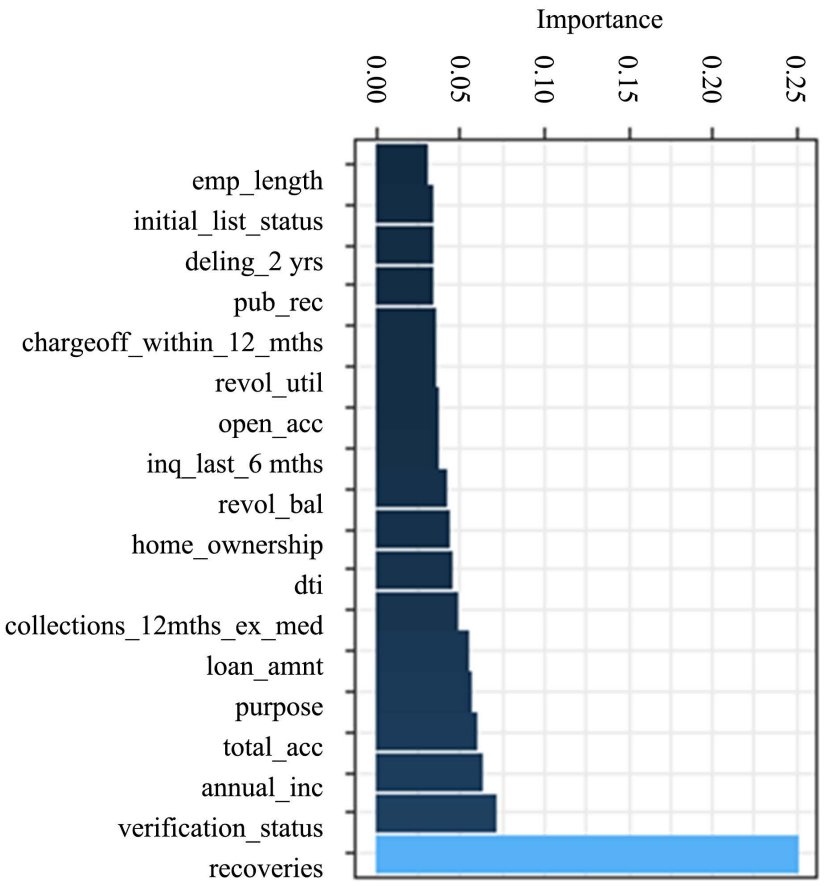

Figure 5. Variable importance using Garson's algorithm.

Table 4. Receiver operating characteristic-AUC.

Measure	Logistic regression	Random forest	Neural network
AUC	0.899	0.899	0.936

These performances are reflected in the ROC curve of the classifiers (Figure 6).

As observed in Figure 6, the neural network model displays a better discriminatory power. Its ROC curve is more slightly distant to the diagonal line (which is referred to as a model without discriminatory power) than the one of the other models.

With regard to the confusion matrix as performance metric, no difference is really observed in the models accuracy (Table 5). The results show an accuracy of approximately 0.90, meaning that the models make around 90% of correct predictions for the test data, which is relatively high.

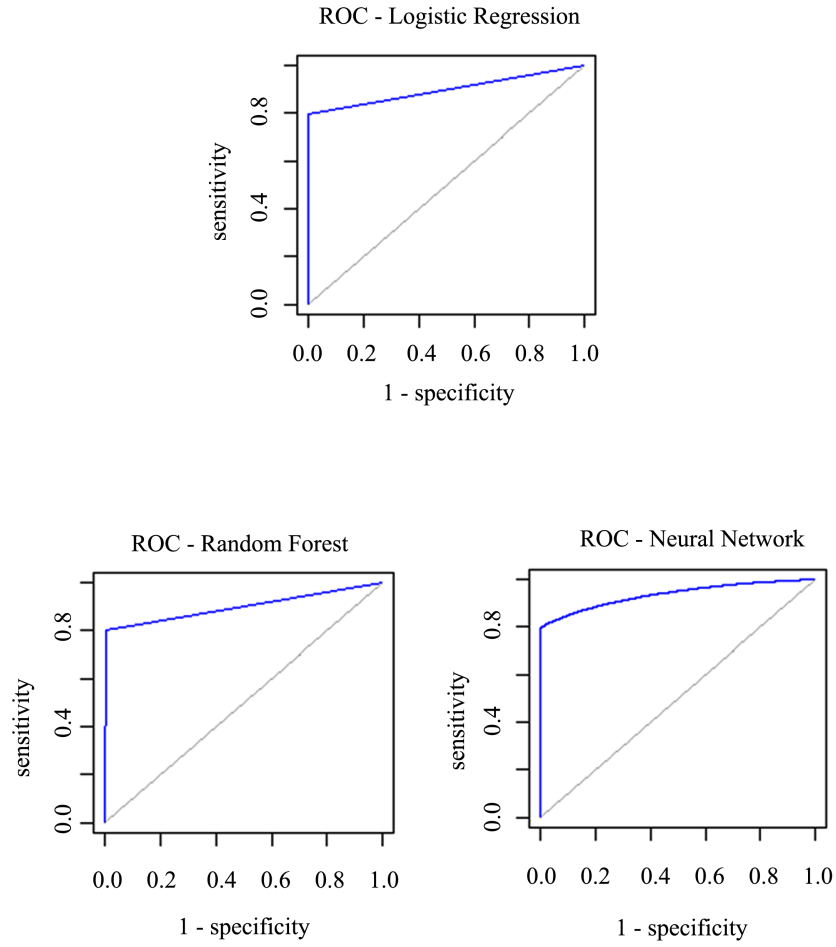

Figure 6. ROC curve of the three classification models.

Table 5. Confusion matrix of the models.

Parameter	Logistic regression	Random forest	Neural network
Accuracy	0.899	0.899	0.894
Sensitivity	1.000	0.995	0.790
Specificity	0.799	0.803	0.825

When it comes to the sensitivity, the logistic regression and the random forest models display the best performance, whose values indicate that they correctly classify all of the defaulted loans; whereas the neural network model correctly classifies 79% of the defaulted loans. On the other hand, the neural network model shows the highest specificity indicating that 82.5% of the non-defaulted loans are correctly classified. Around 80% of the non-defaulted loans are correctly classified by the logistic regression and the random forest models.

Overall, the logistic regression, the random forest and the neural network models display high AUC and high accuracy indicating good ability to discriminate between defaulted loans and non-defaulted loans. So, there is no need for further tuning.

6. Conclusion

In order to help investors make informed decisions by minimizing default risks and reducing information asymmetry relating to the peer-to-peer lending platforms, this paper aimed to build three credit scoring models using data from the leading American peer-to-peer lending platform, LendingClub, employing three machine learning methods: logistic regression, random forest and neural network. The study used the accepted loan application data from 2007 to 2018 containing 2,260,701 observations and 151 variables including the borrower's personal information, loan characteristics and credit history. The analysis window considered by the study was January 20213 to December 2014, which is the most suitable and relevant period, mainly because loans disbursed after 2014 were not fully expired when the data were collected in 2018. Therefore, it might be difficult to classify them as defaulted or non-defaulted loans. The study was interested only in the individual loans. As a result, the dataset was reduced to a sample of 388,518 loans with 151 variables each. Furthermore, the definition of default and non-default was based on the variable loan status, where "full paid" represents non-default and the rest of the attributes stands for default. Approximately 83% of the loans (295,995) were classified as non-defaulted against 17% as defaulted loans (62,523). After data processing and exploratory data analysis, we selected 18 significant variables. We adopted the undersampling method to deal with the imbalanced data. Thus, the same amount of data were randomly sampled for non-defaults (62,523) and defaults (62,523), corresponding to a total of 125,046 loans whose 75% was used to develop the models and 25% to validate the models.

The empirical results show that the independent variables recoveries, debt to income, annual income and the loan amount had the strongest relationship with the response variable, mainly with regard to the logistic regression and the random forest classifiers. We can also notice that the independent variables verification status and annual income gained in importance when it comes to determine variable importance for the neural network model. The independent variable recoveries were the most significant ones with the strongest association with the response variable for the three models under study. Using the Receiver Operat-

ing Characteristic as a performance metric, the neural network model outperformed the logistic regression (0.899) and the random forest (0.899) with an AUC of 0.936, while the assessment of the LendingClub model showed a low performance (0.67). On the other hand, in terms of confusion matrix, no difference is noted in the accuracy of the models, since their respective values are essentially 0.9. The neural network displayed the best specificity (over 0.8), while its sensitivity is the smallest one (less than 0.8). Overall, the neural network classifier displayed a better predictive power compared to the logistic regression and the random forest, even though no significant differences are overall observed in the values of the performance metrics used.

In addition to the credit scoring models, investors might take into account in their investment decision the relationship between some variables and the likelihood of default. For instance, the higher the loan amounts, the higher the likelihood of default. The higher the debt to income, the higher the likelihood of default. Whereas the higher the annual income, the lower the likelihood of default. The likelihood of default tends to decrease as the number of total open accounts increases.

Finally, this paper shed the light on the relatively low predictive power of the LendingClub model through an assessment using the distribution of the grades of the borrowers provided in the dataset and the response variable we defined. The low performance observed served as foundation to develop and propose three better credit scoring models, adopting undersampling method. However, there is still room for improvement. Future studies might collect data from different periods, with oversampling technique or both oversampling and undersampling for comparison, in order to select the model with the best predictive power that would help investors construct a much more profitable investment loan portfolio.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

Abd Elrahman, S. M., & Abraham, A. (2013). A Review of Class Imbalance Problem. *Journal of Network and Innovative Computing, 1,* 332-340. http://ias04.softcomputing.net/jnic2.pdf

Abdou, H. A., Dongmo Tsafack, M. D., Ntim, C. G., & Baker, R. D. (2016). Predicting Creditworthiness in Retail Banking with Limited Scoring Data. *Knowledge-Based Systems, 103,* 89-103. https://www.journals.elsevier.com/knowledge-based-systems https://doi.org/10.1016/j.knosys.2016.03.023

Aldrich, J. H., & Nelson, F. D. (1984). Linear Probability, Logit, and Probit Models. In *Quantitative Application in the Social Science.* SAGE Publications. https://doi.org/10.4135/9781412984744

Breiman, L. (2001). Random Forests. *Machine Learning, 45,* 5-32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984) *Classification and Regression Trees.* Chapman and Hall.

Chang, A.-H., Yang, L.-K., Tsaih, R.-H., & Lin, S.-K. (2022). Machine Learning and Artificial Neural Networks to Construct P2P Lending Credit-Scoring Model: A Case Using Lending Club Data. *Quantitative Finance and Economics, 6,* 303-325. https://doi.org/10.3934/QFE.2022013

Chen, Y.-R., Leu, J.-S., Huang, S.-A., Wang, J.-T., & Takada, J.-I. (2021). Predicting Default Risk on Peer-to-Peer Lending Imbalanced Datasets. *IEEE Access, 9,* 73103-73109. https://doi.org/10.1109/ACCESS.2021.3079701

Fabozzi, F. J., Gupta, F., & Markowitz, H. M. (2002). The Legacy of Modern Portfolio Theory. *The Journal of Investing, 11,* 7-22. https://doi.org/10.3905/joi.2002.319510

Francis, J. C., & Kim, D. (2013). *Modern Portfolio Theory: Foundation, Analysis and New Development.* John Wiley & Sons, Inc.

Fu, Y. (2017). Combination of Random Forests and Neural Networks in Social Lending. *Journal of Financial Risk Management, 6,* 418-426. https://doi.org/10.4236/jfrm.2017.64030

He, Q., & Li, X. (2020). *The Failure of Chinese Peer-to-Peer Lending Platforms: Finance and Politics.* BOFIT Discussion Paper No. 27/2020. The Bank of Finland Institute for Emerging Economies. https://ssrn.com/abstract=3764783 https://doi.org/10.2139/ssrn.3764783

Hou, X. Y. (2020). P2P Borrower Default Identification and Prediction Based on RFE-Multiple Classification Models. *Open Journal of Business and Management, 8,* 866-880. https://doi.org/10.4236/ojbm.2020.82053

Jahangir, R. (2020). *Peer-to-Peer Lending.* https://www.researchgate.net/publication/341445894_PEER_TO_PEER_P2P_LENDING

Lantz, B (2013). *Machine Learning with R.* Packt Publishing.

Lund, B., & Brotherton, D. C. (2013). *Information Value Statistic.* https://mwsug.org/proceedings/2013/AA/MWSUG-2013-AA14.pdf

Ma, Z., Hou, W., & Zhang, D. (2021). A Credit Risk Assessment Model of Borrowers in P2P Lending Based on BP Neural Network. *PLOS ONE, 16,* e0255216. https://doi.org/10.1371/journal.pone.0255216

Madasamy, K., & Ramaswami, M. (2017). Data Imbalance and Classifiers: Impact and Solutions from a Big Data Perspective. *International Journal of Computational Intelligence Research, 13,* 2267-2281. https://www.ripublication.com/ijcir17/ijcirv13n9_09.pdf

Mangram, M. E. (2013). A Simplified Perspective of the Markowitz Portfolio Theory. *Global Journal of Business Research, 7,* 59-70. https://ssrn.com/abstract=2147880

Markowitz, H. (1952). Portfolio Selection. *The Journal of Finance, 7,* 77-91. https://doi.org/10.1111/j.1540-6261.1952.tb01525.x

Millerbernd, A., & Choudhuri-Wade, R. (2022). *LendingClub Personal Loans: 2023 Review.* NerdWallet. https://www.nerdwallet.com/reviews/loans/personal-loans/lendingclub-personal-loans

Namvar, A., Siami, M., Rabhi, F., & Naderpour, M. (2018). Credit Risk Prediction in an Imbalanced Social Lending Environment. *International Journal of Computational Intelligence Systems, 11,* 925-935. https://doi.org/10.2991/ijcis.11.1.70

Navlani, A. (2019). Neural Network Models in R.

Pfaff, B. (2012). *Financial Risk Modelling and Portfolio Optimization with R.* John Wiley

& Sons. https://doi.org/10.1002/9781118477144

Pokorná, M., & Sponer, M. (2016). Social Lending and Its Risks. *Procedia-Social and Behavioral Sciences, 220,* 330-337. https://doi.org/10.1016/j.sbspro.2016.05.506

Serio, A. (2022). *What Is Peer-to-Peer Lending? How Peer-to-Peer Lending Works and the 6 Best Peer-to-Peer Lending Sites.* https://www.finder.com/peer-to-peer-lending

Serrano-Cinca, C., Gutiérrez-Nieto, B., & López-Palacios, L. (2015). Determinants of Default in P2P Lending. *PLOS ONE, 10,* e0139427. https://doi.org/10.1371/journal.pone.0139427

Shelke, M. S., Deshmukh, P. R., & Shandilya, V. (2017). A Review on Imbalanced Data Handling Using Undersampling and Oversampling Technique. *International Journal of Recent Trends in Engineering & Research.*

Vinod Kumar, L., Natarajan, S., Keerthana, S., Chinmayi, K. M., & Lakshmi, N. (2016). Credit Risk Analysis in Peer-to-Peer Lending System. In *2016 IEEE International Conference on Knowledge Engineering and Applications (ICKEA)* (pp. 193-196). IEEE. https://www.researchgate.net/publication/312038529_Credit_Risk_Analysis_in_Peer-to-Peer_Lending_System

von Eye, A., & Clogg, C. C. (1996). *Categorical Variables in Developmental Research: Methods of Analysis*. Academic Press.

Wan, J., Zhang, H., Zhu, X., Sun, X., & Li, G. (2019). Research on Influencing Factors of P2P Network Loan Prepayment Risk Based on Cox Proportional Hazards. *Procedia Computer Science, 162,* 842-848. https://doi.org/10.1016/j.procs.2019.12.058