Data Mining Techniques to Predict Default in LendingClub

Jiaying Sun

Abstract—**This study aims to build a predictive model for default in LendingClub using Artificial Neural Networks, and to compare its performance to the Logistic Regression model. The dataset was downloaded from LendingClub on Kaggle and the files contain complete loan data for all loans issued from 2007-2015, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. There were 1,147 defaults out of 201,350 transactions. The dataset is highly unbalanced, and the positive class (defaults) accounts for 0.570% of all transactions. The records were randomly assigned into one of two groups: training sample and testing sample. We used the two models to predict the risk of default in LendingClub in the testing sample. Receiver operating characteristics (ROCs) were calculated and compared for these two models and a curve measuring predicted probability versus observed probability was plotted to demonstrate the calibration measure for these two models. A ROC of 0.73 in the training sample showed that the Logistic Regression clearly performed better. In the testing sample, the ROC was 0.75 for the Logistic Regression and 0.66 for the Artificial Neural Network. When compared to the Artificial Neural Network model, Logistic Regression had a better discriminating capability and was a better model in estimating credit defaults.**

Index Terms—**Artificial neural network, default in peer-to-peer lending, logistic regression, predictors.**

I. Introduction

Prediction of default risk has been a critical topic for banks and individual lenders for centuries. In an improvement from the past, we now have more availability of large datasets; with our data mining techniques, we are able to avoid unnecessary default risks when choosing whom to lend to [1]. Nowadays, as people are more engaged in investment transactions, they need loans. Small businesses grasp this opportunity to serve as an intermediate and gain profits in between.

Crowdlending is the practice of lending money to individuals or businesses through online services that match lenders with borrowers. It is also called peer-to-peer lending, abbreviated as P2P lending. This method allows investors to lend money through the form of loans to individual borrowers in return for proceeds based on the corresponding interest rates. In Europe, this investing method can produce an "average return on investment of 12-14% per year" [2].

People choose P2P lending over traditional banks because borrowers and lenders both benefit from it. From the borrowers' perspective, they can access the loans with lower interest rates than banks or other traditional financial institutions; or sometimes, a sense of community might form on a P2P lending platform [3]. A potential borrower with a low credit score can choose to share his or her sympathetic story, making a lender amenable to potentially forgoing a higher interest rate and be willing to take the greater risk to fund the loan.

However, these also lead to plenty of downsides for the lenders. They face default risks, especially when they are talked into a worthless and very risky investment by a good sob story. The investors should be aware that borrowers might sometimes remit late payments or even not pay back at all. It is also possible that the loan originator (the loan platform) might dissolve and the investor would not be able to recover any of the money that he or she had invested. Moreover, in the European regions, there is no legal framework for regulating P2P lending when the loan is provided to a business, although the Consumer Credit Directive service provides some regulation for loans processed for consumers.

In recent years, predicting default risk in lending has become an important research theme for P2P lending companies and banks. Although investors see new opportunities on these P2P lending platforms, they should also be aware of the potential uncertainties. With increasingly easier access for investors to lend to individual borrowers on platforms such as LendingClub, Prosper, and Upstart, a default risk predicting model is needed.

LendingClub, the world's largest P2P lending platform, had claimed that $15.98 billion in loans had originated through its platform as of December 31, 2015 [4]. LendingClub enables borrowers to create unsecured personal loans in amounts between $1,000 and $40,000. The standard loan period is three years. Investors can search and browse the loan listings on LendingClub's website and select loans they want to invest in based on the information supplied about the borrower, amount of loan, loan grade, and loan purpose [5]. Investors gain profits from the interest, and LendingClub makes money by charging an origination fee for borrowers and a service fee for investors.

Predicting default risk in lending has been an important theme for LendingClub. To reduce default risk, it focuses on high-credit-worthy borrowers, having declined approximately 90% of the loan applications received as of 2012 and assigning higher interest rates to riskier borrowers within its credit criteria [4]. Only borrowers with a FICO score of 660 or higher can be approved for loans. At its inception, LendingClub's default rate ranged from 1.4% for top-rated three-year loans to 9.8% for the riskiest loans in 2012 [6]. P2P lending's future is dependent on the successful

Manuscript received November 29, 2021; revised January 21, 2022.

Jiaying Sun is with Miss Porter's School, USA / Ivy Analytics LLC, United States (e-mail: jsun22@missporters.org).

doi: 10.18178/joebm.2022.10.1.674

management of the default rate in companies such as LendingClub.

This study aims to 1) examine the predictors of Default in LendingClub 2) build a predictive model for Default in LendingClub using Artificial Neural Networks and Logistic Regression 3) compare its performance to the Logistic Regression model. The following paper is organized as follows: we introduce the topic of P2P, provide a description of the dataset, introduce the methodology that we used, display the results from both training and testing samples, compare our results with previous studies, and draw a conclusion from those graphs.

II. Dataset and Methods

A. Dataset Overview

We worked with a dataset from LendingClub that was found on Kaggle (https://www.kaggle.com).

The files contain complete loan data for all loans issued from 2007 to 2015, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. The file containing loan data through the "present" contains complete loan data for all loans issued through the previous completed calendar quarter. Additional features include credit scores, numbers of financial health inquiries, addresses including zip codes, states of residence, and collection histories, among others. The file is a matrix of about 890 thousand observations and 75 variables. A data dictionary is provided in a separate file.

The outcome of interest in this study is the default (versus fully paid) of payment in LendingClub. The 21 variables that we included in the model are identified in Table I. Overall, other variables were presented in the dataset, such as the month in which the loan was funded, the upper boundary range of the borrower’s most recently pulled FICO score, or the number of months since the last public record was available. We decided not to include all of them as predictors in the model due to their irrelevance.

TABLE I: Variables Included in the Model as Predictors

loan_amnt	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.
int_rate	Interest Rate on the loan
installment	The monthly payment owed by the borrower if the loan originates.
annual_inc	The self-reported annual income provided by the borrower during registration.
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.
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
open_acc	The number of open credit lines in the borrower's credit file.
pub_rec	Number of derogatory public records
revol_bal	Total credit revolving balance
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
total_acc	The total number of credit lines currently in the borrower's credit file
Grade_AB	LC assigned loan grade
Grade_CD	LC assigned loan grade
Grade_EF	LC assigned loan grade
emp7y	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.
emp10y	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.
owned	Home ownership
verified	Indicates if income was verified by LC, not verified, or if the income source was verified
debt_conso	Loan Purpose: Debt consolidation
credit_c	Loan Purpose: Credit card payment

B. Methodology

We will use two models: Logistics Regression and Artificial Neural Network. Artificial Neural Network is “a comparison with a black box having multiple input and multiple output which operates using a large number of mostly parallel connected simple arithmetic units” [7]. Logistic Regression is a model that is used to predict and analyze binary dependent variables using a Logistic function in statistics. All eligible data were randomly assigned into 2 groups: training sample and testing sample. These two models were built using training samples. In the testing sample, we used these two models to predict the risk of default in LendingClub.

Receiver operating characteristics (ROCs) were calculated and compared for these two models based on their discrimination capability, and a curve using predicted probability versus observed probability was plotted to demonstrate the calibration measure for the two models. The ROC curve shows the trade-off between sensitivity (True Positive Rate) and specificity (1 - False Positive Rate). Any increase in sensitivity will be accompanied by a decrease in specificity [8]. Curves that are closer to the top-left corner indicate a better performance, or a higher accuracy. As a baseline, a random classifier is expected to result in points lying on the diagonal, where the False Positive Rate equals the True Positive Rate.

We used a total of 233 lines of R language code to achieve the entire implementation process. We were able to produce the graphs (Fig. 1, Fig. 2…), Table II, and efficiently classify the numbers in the R studio.

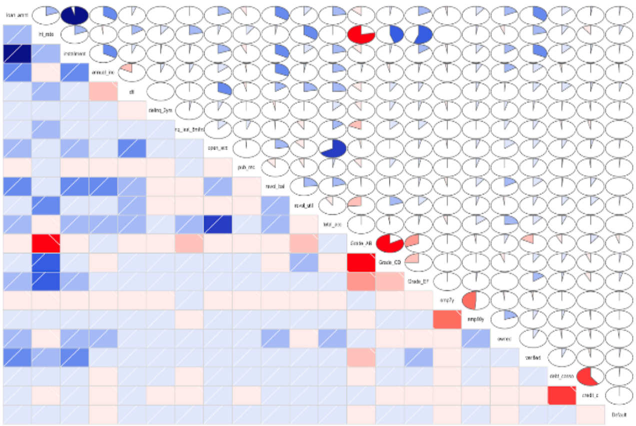

Fig. 1. Matrix of correlations between variables.

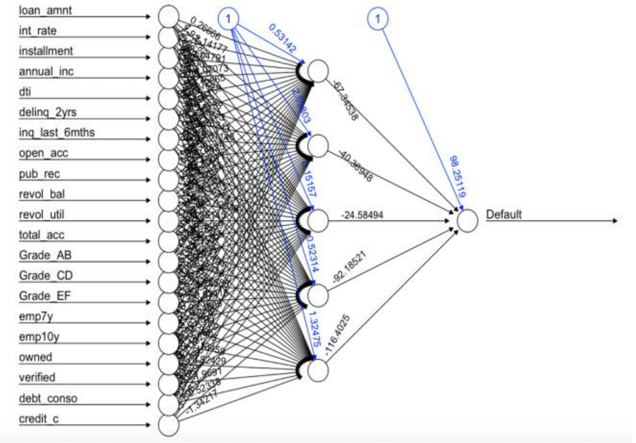

Fig. 2. Artificial Neural Network in training sample.

TABLE II: LOGISTIC REGRESSION FOR DEFAULT IN LENDINGCLUB

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-6.015	0.447	-13.454	< 2e-16	***
loan_amnt	0.000	0.000	7.578	0.000	***
int_rate	0.038	0.015	2.495	0.013	*
installment	-0.002	0.000	-6.134	0.000	***
annual_inc	0.000	0.000	-2.011	0.044	*
dti	0.046	0.004	10.656	< 2e-16	***
delinq_2yrs	0.148	0.026	5.615	0.000	***
inq_last_6mths	-0.060	0.030	-1.998	0.046	*
open_acc	0.048	0.008	5.976	0.000	***
pub_rec	0.253	0.050	5.060	0.000	***
revol_bal	0.000	0.000	-1.059	0.290	
revol_util	0.000	0.001	0.296	0.768	
total_acc	-0.021	0.004	-5.606	0.000	***
Grade_AB	-1.254	0.318	-3.947	0.000	***
Grade_CD	-0.398	0.265	-1.502	0.133	
Grade_EF	-0.073	0.238	-0.308	0.758	
emp7y	-0.032	0.080	-0.399	0.690	
emp10y	0.098	0.071	1.369	0.171	
owned	-0.404	0.064	-6.339	0.000	***
verified	0.133	0.076	1.738	0.082	
debt_conso	0.121	0.086	1.397	0.162	
credit_c	0.140	0.106	1.320	0.187	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

III. RESULTS AND DISCUSSION

A. Results

There were 1,147 defaults out of 201,350 transactions. The dataset is highly unbalanced, and the positive class (defaults) accounts for 0.570% of all transactions.

Basically, a corrgram as shown in Fig. 1 is a graphical representation of the cells of a matrix of correlations. The idea is to display the pattern of correlations in terms of their signs and magnitudes using visual thinning and correlation-based variable ordering. Moreover, the cells of the matrix can be shaded or colored to show the correlation value. The positive correlations are shown in blue, while the negative correlations are shown in red; the darker the hue, the greater the magnitude of the correlation.

According to the Logistic Regression in Table II, the loan amount, interest rate and installment were important predictors for default in LendingClub. So were the debt/income ratio, incidences of delinquency in the past 2 years, inquiries in the past 6 months, number of open credit lines, number of derogatory public records and total number of current credit lines.

The equation of Logistic Regression is defined as below:

$$Ln\left(\frac{P}{1-P}\right)=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_kX_k$$

In the above plot, the line thickness represents weight magnitude and line color is a weight sign (black = positive, grey = negative). The net is essentially a black box so we cannot say that much about the fitting, the weights and the model. Suffice to say that the training algorithm has converged and therefore the model is ready to be used. The loan grade and ownership of the home were also significant predictors.

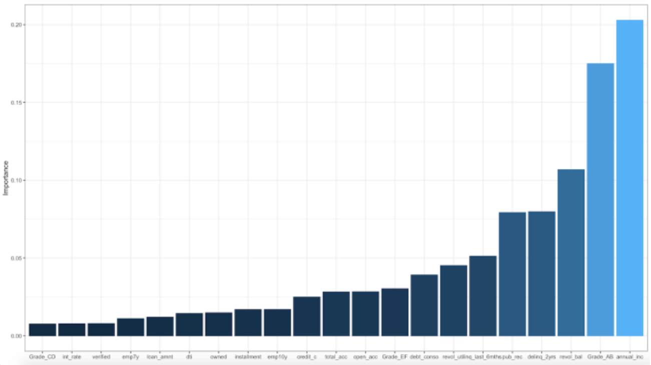

Fig. 3. Variable importance in artificial neural network.

In Fig. 3, the top 5 most important predictors were annual income, total credit revolving balance, number of derogatory public records, revolving line utilization rate and loan grade. All five together accounted for more than 60% of the weights in the model.

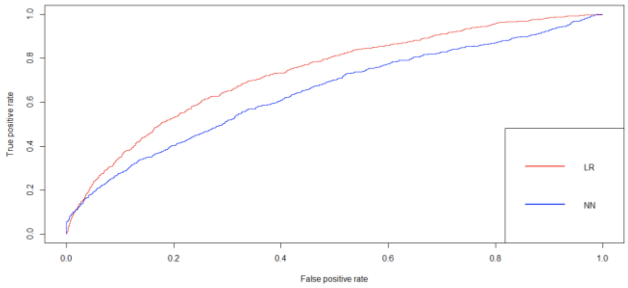

Fig. 4. ROC in training sample for Logistic Regression (Red) vs Neural Network (Blue).

For the training sample displayed in Fig. 4, the ROC was 0.73 for the Logistic Regression (red) and 0.65 for the Artificial Neural Network (blue). Logistic Regression clearly performed better.

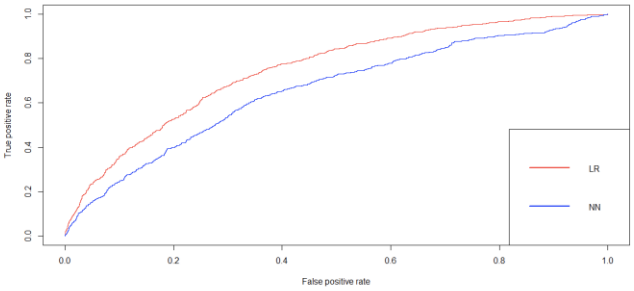

Fig. 5. ROC in testing sample for Logistic Regression (Red) vs Neural Network (Blue).

In the testing sample, the ROC was 0.75 for the Logistic Regression (red) and 0.66 for the Artificial Neural Network (blue), as shown in Fig. 5. Similarly, in Fig. 4, Logistic Regression performs at a better rate of accuracy.

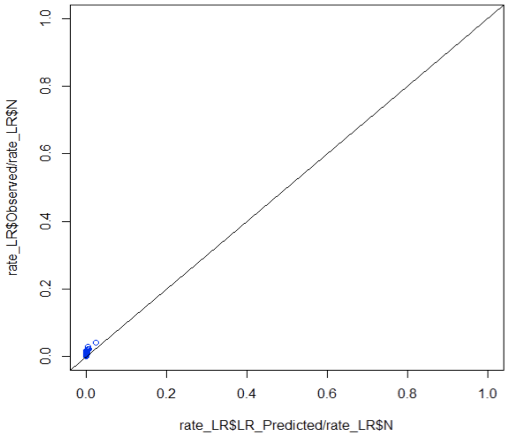

Fig. 6. Predicted Probability vs. Observed Probability in testing sample for Neural Network, sorted by predicted probability.

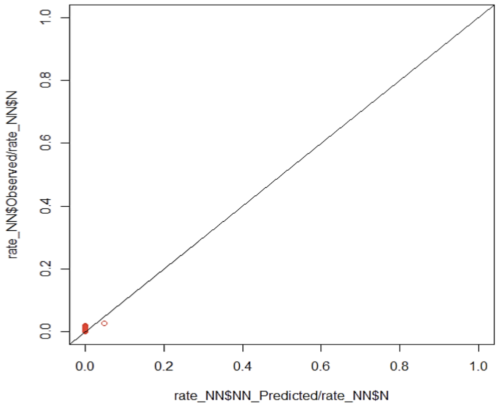

Fig. 7. Predicted Probability vs. Observed Probability in testing sample for Logistic Regression, sorted by predicted probability.

By visually inspecting and comparing the plots of Fig. 6 and Fig. 7, we can see that the predictions made by the neural network and the Logistic model appeared roughly the same.

B. Discussion

We graphed two ROCs for both the training sample and testing sample. As one of the characteristics of the ROC curve, the closer the curve comes to a 45-degree diagonal of space, the less accurate it is. As shown in both Fig. 4 and 5, the blue curve which stands for the Artificial Neural Network performed closer to a 45-degree measurement, indicating that the Logistic Regression model is a better prediction model.

There has been a considerable number of previous studies on loan evaluations using machine learning models of LendingClub datasets. Some of these studies concluded with results similar to our own.

Chang *et al.* [9] experimented with Logistic Regression, Naïve Bayes, and SVM for default prediction. They concluded that Naïve Bayes with Gaussion outperformed all of the others with 80.1% of sensitivity. As Li *et al.* [10] posited in their paper, if the current loans are considered as positive examples, they may become the default in the future, which labels some true negatives as positives. They then decided to only test with finalized loans, which course we followed, as we only used the 1,147 defaults out of all 201,350 transactions. This would increase our accuracy rate and avoid the error of mistaking current status loans for finalized ones.

Tsai *et al*. [11] used algorithms from machine learning to optimize P2P lending risk. They focused on optimization for loans classified as good as the primary metric and compared the return rate of LendingClub to theirs given the same default risk rate. The four algorithms they used were: Logistic Regression, LibSVM, Naïve Bayes, and Random Forest. They found that Logistic Regression achieved the best accuracy out of all four. This was similar to our result because we found that Logistic Regression outperformed Artificial Neural Network.

Pujun *et al.* [12] included a combination of classification, regression, and clustering methodologies to explore the loan application process in LendingClub. Their objective was to ascertain which were the best predictors for a loan application to be accepted or not, from the lending platform's perspective. They discovered that the loan grade was a "near perfect predictor of the interest rate" based on the regression results. We found it interesting because loan grade was a relatively important predictor in our Artificial Neural Network results. An interesting revelation of their study was the finding that applicants should state their purpose of the loan as credit card consolidation in order to optimize the chance of its acceptance.

Overall, our results had some similarities to those of previous studies, and we had a lot of takeaways from reading them.

IV. Conclusion

Known as crowd-lending, P2P features many transactions that are unsecured personal loans, though some of the largest amounts are lent to businesses. Unsecured personal loans demand better management of risk of default to make a legitimate investment.

In this study, we built a predictive model using Artificial Neural Networks as well as Logistic Regression to provide a tool for predicting default in LendingClub. The difference in the two models highlights the need to employ different tools to understand the predictors of payment default in LendingClub, to better manage the business risk.

According to Logistic Regression, the loan amount, interest rate and installment were important predictors for default in LendingClub. So were the debt/income ratio, incidences of delinquency in the past 2 years, inquiries in the past 6 months, number of open credit lines, number of derogatory public records and total number of current credit lines.

According to the Artificial Neural Network, the top 5 most important predictors were annual income, total credit revolving balance, number of derogatory public records, revolving line utilization rate and loan grade. All five together accounted for more than 60% of the weights in the model.

We did not test the external validity of Logistic Regression or the Artificial Neural Network. However, we performed a comprehensive split-sample validation with both strategies. When compared to Artificial Neural Network models, Logistic Regression had better discriminatory capability. It follows that P2P lending companies can use Logistic Regression as a model to more efficiently predict loan defaults in the future.

V. Future Studies

Our study can be useful for future scholars because they can choose Logistic Regression over Artificial Neural Networks to predict defaults using LendingClub datasets.

Future studies could use external data and test the performance of the outputs from these two models in this study. A predictive model would be an extremely useful tool to timely identify default in LendingClub. Different datasets might generate a different result due to the variance.

We've read papers from scholars who've done studies on dataset models for other large lending platforms; it appears that future studies could compare our paper to theirs when exploring this issue. Other statistical techniques, such as the bootstrap, could be used in comparisons with the two models we included. New findings might reveal different conclusions as to which one is the best model to use in predictive studies.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] J. D. Turiel and T. Aste, "Peer-to-peer loan acceptance and default prediction with artificial intelligence," *R. Soc. Open Sci*., p. 7191649, 2020.

[2] P2P lending, crowdlending. (Aug. 11, 2019). P2P lending risks - Is Peer to peer lending safe? [Online]. Available: https://www.crowdfunding-platforms.com/p2p-lending-risks-is-peer-to-peer-lending-safe

[3] B. Schneider. (Aug. 26, 2020). Peer-to-peer lending breaks down financial borders. *Investopedia*. [Online]. Available: www.investopedia.com/articles/financial-theory/08/peer-to-peer-lending.asp#:~:text=The%20major%20benefits%20of%20P2P,why%20they%20need%20the%20money

[4] Lending club statistics. [Online]. Available: https://www.lendingclub.com/

[5] Online personal loans at great rates. LendingClub. [Online]. Available: https://www.lendingclub.com/

[6] C. Barth. (June 6, 2012). Looking for 10% yields? Go online for peer to peer lending. *Forbes*. [Online]. Available: https://www.forbes.com/sites/chrisbarth/2012/06/06/looking-for-10-yields-go-online-for-peer-to-peer-lending/

[7] J. Zupan, "Artificial Neural Networks (ANNs)," *Chemoinformatics*, pp. 438–452, 2018.

[8] T. Tape. Plotting and interpreting an ROC curve. [Online]. Available: https://www.gim.unmc.edu/dxtests/ROC2.htm

[9] S. Chang, S. D.-O. Kim, and G. Kondo, "Predicting default risk of lending club loans," *CS229: Machine Learning*, 2015.

[10] P. Li and G. Han, "LendingClub loan default and profitability prediction," *CS229: Computer Science*, 2018.

[11] K. Tsai, S. Ramiah, and S. Singh, "Peer lending risk predictor," *CS229: Business*, 2014.

[12] B. Pujun, C. Nick, and L. Max, "Demystifying the workings of lending club," *CS229 Stanford*, 2016.

Jiaying Sun was born in Shanghai, China. She studies at Miss Porter's School. Jiaying Sun is focused on the field of data science and machine learning.

She is a research assistant for Betty Wang from Ivy Analytics LLC in Delaware County in Pennsylvania. She is also working on a project with Jayson Lynch, a postdoctoral researcher at University of Waterloo. She is currently working on a project related to reversible computing on machine learning algorithms.