Scientific Research Publishing

Journal of Financial Risk Management, 2017, 6, 418-426
http://www.scirp.org/journal/jfrm
ISSN Online: 2167-9541
ISSN Print: 2167-9533

Combination of Random Forests and Neural Networks in Social Lending

Yijie Fu

Shanghai Foreign Language School, Shanghai, China
Email: jessyfu@163.com

How to cite this paper: Fu, Y. J. (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

Received: October 26, 2017
Accepted: December 26, 2017
Published: December 29, 2017

Abstract

Social lending, also known as peer-to-peer lending, provides customers with a platform to borrow and lend money online. It is now rapidly gaining its popularity for its superior monetary advantage comparing to banks for both borrowers and lenders. Thus, choosing a reliable is very important, whereas the only method most of the platforms use now is a grading system. In order to better prevent the risks, we propose a method of combining Random Forests and Neural Network for predicting the borrowers' status. Our data are from Lending Club, a popular social lending platform, and our results indicate that our method outperforms the lending Club good borrower grades.

Keywords

Peer-to-Peer Lending, Machine Learning Methods, Random Forests, Neural Networks

1. Introduction

Peer-to-peer lending is a rapidly growing competitor of traditional financial institutions. It is the process of lending money to individuals through online services. It has gained popularity in the past few years for lenders earning higher returns than savings and investment products in the banks, while borrowers loaning at lower interest rates. In P2P lending, lenders are able to choose whom to lend to. However, this flexibility also brings out the problem that since the government does not protect risks, lenders will have to take efforts finding the best borrower based on their limit information, which makes our work extremely important.

There is a variety of social lending platforms in use today and the most popular ones are Lending Club, Upstart, Funding Circle, Prosper Marketplace, CircleBack

Lending, Peerform, and SoFi. Among them, we picked the first publicly traded online P2P lending company in the U.S., Lending Club (LC), to finish our research. LC now has more than 1.5 million customers and the total amount of money borrowed is about 28 billion USD (LendingClub.com, 2015). The data from LC is chosen for this research because it is easily accessible and relatively large enough, comparing to other social lending websites.

Most of the popular P2P platforms graded the borrowers based on cooperating credit reporting agencies, and the grade determined a lot. Take LC as an example, the loan grade is generated by credit score and adjusted by client's credit report and loan application (LendingClub.com, 2015). And the possibility to get a loan as well as the interest rate and term is determined by the grade.

An analysis by Emekter, Tu, Jirasakuldech, and Lud suggests that high income borrowers with the highest FICO credit scores don't necessarily borrow from LC. More precisely, the top one third FICO credit score consumers do not create any loan on LC. What's more, higher interest rates for higher risk borrowers, a widely accepted concept, usually fails to work as people imagined it to, which means that LC loan grades are not accurate enough to estimate the potential risk lenders are facing. The above two findings reveal the great importance of a better evaluation method (Emekter et al., 2015).

In order to improve the identification method of good borrowers, in this paper, we compare several machine learning methods including Random Forests, Neural Networks and a specially designed combination of the two classifiers to better identify good borrowers in peer-to-peer lending.

The rest of this essay is organized as follows: Section 2 provides a brief introduction to works that have been done on this topic. In Section 3, we introduce the features we picked among the data and how we process the nominal features. Section 4 includes a further explanation of the machine learning methods we used in our model.

2. Literature Review

FICO score basically reveals a customer's credit on tradition bank-mediated financial markets, while an application of the machine learning methods used in P2P lending reveals that if an individual with low FICO score lives in a trusted social community, he, with a great possibility, is also trustworthy. Thus, FICO scores cannot reveal everything and sometimes overlook the credit of customers (Lopez, 2009).

Other methods are developed to evaluate a borrower instead of simply looking at his or her FICO score. A study (Klafft, 2008) concluded three rules to decrease the risk: invest only in borrowers with no delinquent accounts, debt-to-income (DTI) less than 20% and no credit inquiry in the last half a year.

Other research has been done on using Random Forests and Neural Network to train the lending club data. A RF-based methodology for identification of good borrowers in social lending was conducted. In the research, a comparison

of the machine learning methods Radom Forests, SVM, Linear Regressions, and k-NN for identifying good borrowers in social lending was proposed, and it turned out that Random Forests outperformed the other classifiers as well as the FICO scores and LC grades (Malekipirbazari &Aksakalli, 2015).

The other work, which is close to ours, is the work of Zang, Qi, and Fu. In their work, Neural Network was used to assess whether a borrower is trustworthy. BP Neural Network performs better than traditional statistics model because it reduces assumptions and difficulty in counting parameter of measurement (Zang, Qi, & Fu, 2015).

Current literature provides concrete prediction using several machine learning algorithms. However, few scholars tested the combination of different algorithms. In this paper, the performances of combination of Random Forests and Neural Networks are tested and discussed.

3. Data Exploration and Explanation

3.1. An Overview of Lending Club

As we picked data from the Lending Club as our analyzing subject, the following present how the Lending Club works:

- Borrowers meeting the certain criteria apply for a loan on Lending Club's online platform.
- The Lending Club determines the interest rate based on borrower's LC credit grade.
- Lenders browse the loans and the borrowers' information including their FICO score, LC credit grade, debt to income ratio, home ownership status, and number of delinquent accounts the platform and build a portfolio of loans. (Usually, to prevent losing too much, lenders divide their money to support different programs so as to spread the risks.)
- A loan can stay on the platform for up to 14 days and when loans will be deleted from the platform once they funded enough loans within the limit.
- Borrowers will have to pass a verification to get their loans.
- Once the borrower passes the verification and his or her loans are fully funded, borrowers will receive their money in a couple of business days and begin making payments within 30 days.
- Lending Club receives a certain amount of service fee from the lenders.

3.2. Explanatory Data Analysis

Our data includes approximately 1320 K borrower records, containing 85 features in total, which are commonly used by credit agencies to evaluate the borrowers. However, some of these records have not reached maturity yet or are lack of information. We filtered the data our feature selection process, which is always of cardinal importance in machine learning because these features you chose might determine the accuracy of your model. After our analysis and studies upon these features, we selected 13 of them to be used in predictive modeling.

We also pre-process the 13 features by replacing the missing information by the mean of that feature, as well as converting nominal attributes to corresponding values. We selected our features in the model as shown below.

- Loan Status: Binary variable indicating whether a borrower has fully paid his loan or not. In our model, "fully paid" is referred to good ones and others including "Charged Off" and "Default" are referred to bad ones.
- Annual Income: The annual income information provided by the borrowers.
- Delinquencies: The numbers of delinquencies in the past two years of each borrower.
- DTI (Debt to Income): Ratio of the borrower's monthly debt to monthly income.
- Employment Length: The length of time in years the borrower is employed in a company. Possible values are integers between 1 and 10 as well as "more than 10 years" and "less than a year".
- Grade: The grade ranged from A to G given by the data source based on loan characteristics and risk assessment of the borrowers.
- Home Ownership: The feature presents whether the borrower owns his or her house. Possible answers are "own", "mortgage", "rent", "any" and "none".
- Interest Rate: the proportion of a loan that is charged as interest to the borrower.
- Loan Amount: The total amount of money of a loan.
- Open Accounts: The number of open credit accounts on the borrower's credit file.
- Revolving Utilization: The amount of revolving credit limits that the borrower is currently using.
- Term: Every period of time the borrower pay for his loan. Values can be either 36 months or 60 months.
- Total Accounts: The total number of open credit accounts on the borrower's credit file.

3.3. Data Processing

Out of the 13 features we picked, 9 of them are numeric. For those rows lack of valid data, we replaced them with the mean of that column before any model building process. The remaining 4 features are nominal, which means that we processed them to convert them into corresponding numbers.

Take Home Ownership as an example, possible answers are "own", "mortgage", "rent", "any" and "none". What we did was using values to represent each of the answers. To be more specific, "own" is considered "1", "mortgage" is regarded as "0.5"and the others represent "0". Through this process, our algorithms can analyze the data.

Since the range of values of features varies widely, objective functions may not work properly. To deal with this, we normalized our data by subtracting the mean from each feature and divided the values by variance. This act could efficiently avoid the situation that one feature with a broad range dominating the results.

Parameters are one of the most important factors in algorithms. In order to find out the most suitable parameter, we listed several possible answers and ran all of them to see which one performed the best in Random Forests and Neural Networks.

Besides parameters, data selection was also very important in machine learning. In our research, we made use of only the second half part of the dataset because the first part was too many years ago to be valuable and the assessing system wasn't advanced enough at that time, so using these data might impact our final results in a negative way.

4. Methodology

4.1. Random Forests

Random forests classifier is a popular classification way in machine learning. By constructing a great amount of decision trees, random forests classifier is strengthened. Decision trees, whose basic idea is that groups of weak learners come together and form a stronger learner, start with a root, keep growing its branches, and ultimately reach its terminal node called leaves. The branches imported to the "tree" are features or processed information based on those features. Comparing to other algorithms, Random Forest Classifiers run efficiently on a large database with a relatively high accuracy due to its lower risk of overfitting.

Random Forest is an advanced bagging technique instead of a boosting technique, which can help lead to "improvements for unstable procedures" (Breiman, 2001). By randomly splitting attributes, Random Forests de-correlate the decision trees (**Figure 1**), leading to an improvement in the bagging techniques.

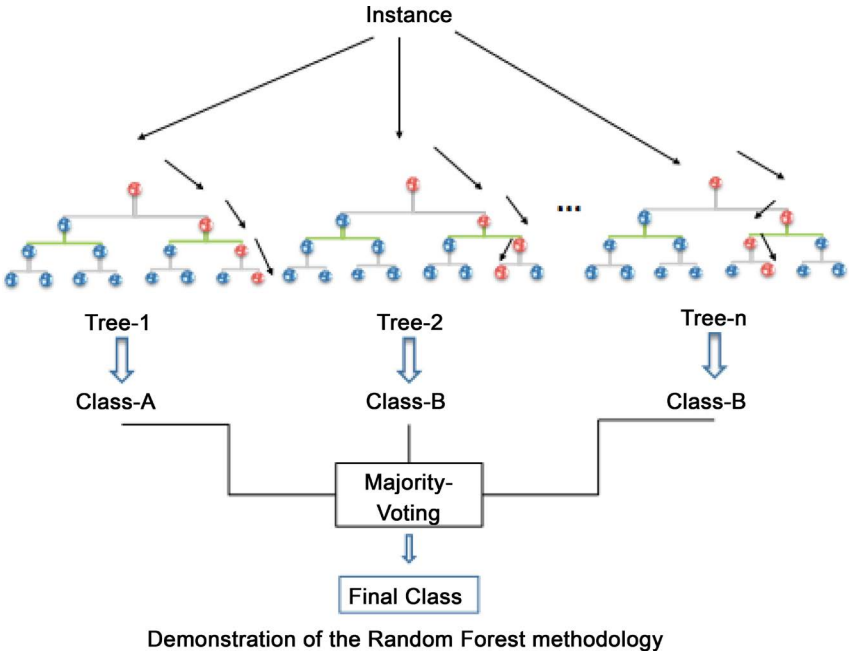

Figure 1. Demonstration of the random Forest methodology.

4.2. Neural Network

Neural Network Classifier (Figure 2) has been widely used in machine learning in the past few years. Each neural network, consisted of neurons that convert an input vector into the output, computes a nonlinear function and passes the output onto the next layer. Neural network can detect complex nonlinear relationships between variables and construct multiple training algorithms. However, Neural Networks have some shortcomings. To have better results, we used 8 epochs, which means that we used and ran our data 8 times to better fit.

4.3. Combination

Both Neural Networks and Random Forests are well-known algorithms, but they differ from each other not only on their proneness to overfitting but also on the basic principles of two algorithms. mathematical modeling. We compare Neural Networks with Random Forests in Figure 3. Random Forests are based on decision trees, while nonlinear models are performed by neural networks. The difference between them made us concerned whether a combination can turn out better results.

What we did was basically using alternative classifiers. We first processed the features as we did in Random Forest and Neural Network algorithms. Then we ran the Random Forest classifier using the most suitable parameters as we tried several combinations of them before.

In our training, we calculated the predicted class, a vote by trees in the forest weighted by their probability estimates, among which we picked the one with the highest mean probability estimate across the trees as one of our features. Training the updated features through Neural Networks, we found a slight improvement in the accuracy. We tried both sequences and it turned out that there were slight improvements after such an operation.

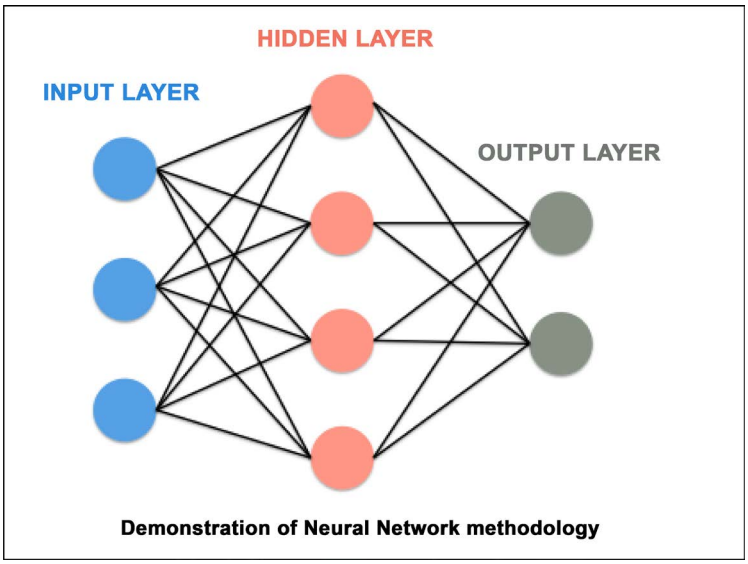

Figure 2. Demonstration of the Neural Network methodology.

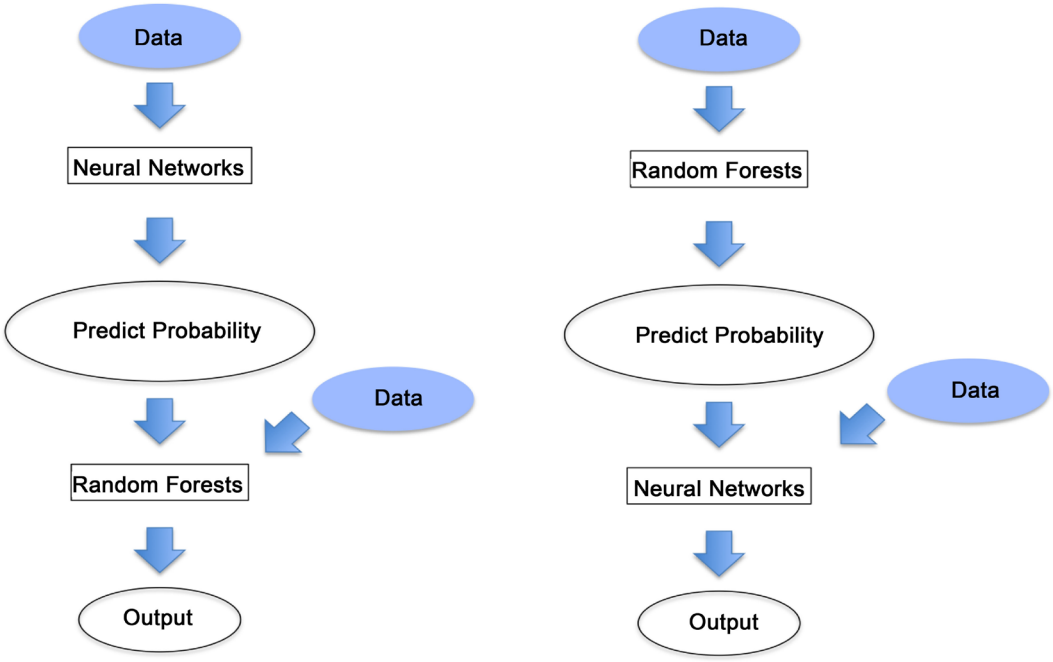

Figure 3. Neural Networks and Random Forests.

5. Results

In our research, whether a model is successful or not is determined by the accuracy. To strengthen the comparison effect, we first used the linear function. Figure x will be presenting two sets of results, differing from each other only on the preprocessing part. As we mentioned earlier in the article, we subtracted the mean from each feature and divided the values by variance. The first set included this act and the second didn't.

In Table 1, after preprocessing, the highest accuracy of the linear function was about 65.6%. The highest accuracy of single Random Forest was 73.3% and the one of Neural Network was 67.8%. There were slight changes using our combination algorithms and different sequence of the combination turns out different results. A combination of first Random Forest than Neural Network had the accuracy of 72.0% while the opposite one turned out 73.5%.

After preprocessing, the highest accuracy of the linear function was about 65.6%. The highest accuracy of single Random Forest was 73.3% and the one of Neural Network was 67.8%. There were slight changes using our combination algorithms and different sequence of the combination turns out different results. A combination of first Random Forest than Neural Network had the accuracy of 72.0% while the opposite one turned out 73.5%.

Comparing to the linear function and Neural Network, Random Forest exhibited superior advantages of more than 5% in its results. We can also see improvements in our "Neural Network + Random Forest" combination. Though the "Random Forest + Neural Network" combination turned out lower accuracy than single Random Forest, it was still far higher than single Neural Network and the average accuracy of Random Forest and Neural Network.

Table 1. Linear Function.

Classifier	Accuracy (%)	
	Preprocessed	Raw
Linear function	65.6%	49.7%
Single Random Forest	73.3%	73.3%
Single Neural Network	67.8%	51.6%
Random Forest + Neural Network	72.0%	50.3%
Neural Network + Random Forest	73.5%	73.3%

The next part of the table showed the second result set in which we excluded the process of subtracting the mean of features and dividing the variance. We can see that all the results are equal to or worse than the one with preprocessing, which proved the necessity of the preprocessing. And also without the preprocessing, there were no advantages of our combination and we concluded that it was because the repeating training led to overfitting.

6. Summary and Conclusion

Given that true creditworthiness is significant to social lending markets, in order to help reduce the situation of mismatching, we conducted our research and drew the following conclusions:

1) The preprocessing makes great differences when using our combination.

As shown in the results, the accuracy with and without the preprocessing of subtracting the mean of features and dividing the variance differs from each other a lot.

The reason of this would be even each model captures nonlinearity in a different manner, they are both adequate to describe the nonlinearity for the task of 2-class classification problem with a small number of features.

2) Our combination does help when the network is relatively simple while making not much difference when it is complicated.

According to our results, we can see that our combination didn't make great changes in the algorithms. We also tried the same data with smaller parameters, and it turned out that although the results as a whole were much lower than those with larger parameters, our combination affected them in a more obvious way.

3) The selection of the data (the range of starting and ending years) affects the final result greatly.

The results shown above were after our preprocessing which cut down the first half of the dataset. At first we used all the data but we soon found that some of the data were too old to have reference value, and it indeed turned out that without those data, our algorithms worked much better than before.

References

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

Emekter, R., Tu, Y., Jirasakuldech, B., & Lud, M. (2015). Evaluating Credit Risk and Loan Performance in Online Peer-to-Peer (p2p) Lending. *Applied Economics, 47,* 54-70. https://doi.org/10.1080/00036846.2014.962222

Klafft, M. (2008). Online Peer-to-Peer Lending: A Lender's Perspective. In H. R. Arabnia & A. Bahrami (eds.), *Proceedings of the International Conference on E-Learning, E-Business, Enterprise Information Systems, and E-Government* (pp. 371-375). Las Vegas: CSREA Press. https://doi.org/10.2139/ssrn.1352352

LendingClub.com, 2015. Accessed January 27th. http://www.lendingclub.com/public/about-us.action

Lopez, S. H. (2009). Social Interactions in p2p Lending. In *Proceedings of the 3rd Workshop on Social Network Mining and Analysis* (pp. 1-8). ACM.

Malekipirbazari, M., & Aksakalli, V. (2015). Risk Assessment in Social Lending via Random Forests. *Expert Systems with Applications, 42,* 4621-4631. https://doi.org/10.1016/j.eswa.2015.02.001

Zang, D. G., Qi, M. Y., & Fu, Y. M. (2015). The Credit Risk Assessment of P2P Lending Based on BP Neural Network. In *Industrial Engineering and Management Science: Proceedings of the 2014 International Conference on Industrial Engineering and Management Science (IEMS 2014)* (Vol. 2, p. 91). 8-9 August 2014. Hong Kong: CRC Press.