Credit Risk Analysis in Peer-to-Peer Lending System

Conference Paper · September 2016

DOI: 10.1109/ICKEA.2016.7803017

5 authors, including:

Natarajan Subramanyam
PES Institute of Technology
19 PUBLICATIONS **90** CITATIONS

SEE PROFILE

Credit Risk Analysis in Peer-to-Peer Lending System

Vinod Kumar L
Software Engineer
CISCO Systems
Bangalore, India
e-mail:
vinodkumarlogan@gmail.com

Natarajan S
Department of ISE
PESIT (main campus)
Bangalore, India
e-mail: natarajan@pes.edu

Keerthana S, Chinmayi K M, Lakshmi N
Department of ISE
PESIT (main campus)
Bangalore, India
e-mail: keerthana10295@gmail.com,
chinmayi.1994@gmail.com,
lak.blue@gmail.com

***Abstract*—This research paper aims to analyze the credit risk involved in peer-to-peer (P2P) lending system of "LendingClub" Company. The P2P system allows investors to get significantly higher return on investment as compared to bank deposit, but it comes with a risk of the loan and interest not being repaid. Ensemble machine learning algorithms and preprocessing techniques are used to explore, analyze and determine the factors which play crucial role in predicting the credit risk involved in "LendingClub" publicly available 2013-2015 loan applications dataset. A loan is considered "good" if it's repaid with interest and on time. The algorithms are optimized to favor the potential good loans whilst identifying defaults or risky credits.**

Keywords-credit risk analysis; machine learning applications; ensemble classification; data mining for banking

I. INTRODUCTION

Peer-to-peer lending is a new decentralized model for investors lend capital and potential borrowers to avail credit. Multiple investors and potential borrowers collaborate on common online platform that doesn't involve any financial institutions or middlemen in the end-to-end process. The borrower requests for loan amount, which range is categorized on basis of interest rates. The investor has the liberty to choose the amount of capital, the interest cap and even the borrower.

These loans are not completely secure as they involve substantial risk of default [1] and hence require added effort to identify and determine a borrower from a pool on unknown users. Identifying potential credit defaults is mandatory in peer-to-peer lending platforms, but determining the good credits and financing them has higher priority, as the P2P revenue model is dependent on the number of loans, volume of credit etc., but incurring losses on account of defaults will make the business unviable.

The processing of borrower's loan application involves collection of user data like annual income, credit history, bank balance, other loans, etc. The foremost priority is to identify the subset of these attributes that is capable of classifying the loan application as one with potential default risk or not.

Since the number of attributes collected for every credit record is much smaller than the number of credit records, tree based classifiers are used, as the run-time complexity of building the tree is linearly proportional to the number of features. They are also capable of handling the skewness of the data, which is not possible in the case of "Logistic Regression", where oversampling is needed.

Tree based classification is a data mining technique which recursively builds a set of rules based on multiple input variables that are either numerical or categorical and outputs a class. Tree based classifiers like Decision Tree [2], Random Forest [3], Bagging [4] and Extra Trees [5] are used to train prediction models for the peer-to-peer lenders data.

The return on investment for an investor on any loan is high if the loan has a higher interest rate, but the risk of defaulting is also proportionally high. Investor can also choose to play safe by investing in low interest loans which has higher chances of being repaid. Hence the tree based models are optimized to get a better precision ahead of accuracy. The details and split up of the loan's interest rates, is explained in the upcoming sections.

II. LITERATURE SURVEY

Recent peer-to-peer lending trend in the last decade has led to a boom in online lending platforms such as "LendingClub", "Peerform", "Upstart", "Funding Circle", "Lendbox", etc. Borrowers with low credit scores in traditional financial institutions are more likely to apply for the same loan in these new platforms, increasing the potential credit risk of default, to the investor. Hence the significant rise in research in credit risks analysis.

The "Peer Lending Risk Predictor" [6] research was carried out on LendingClub's 2007-2013 dataset having 91,520 credit records, to build machine learning models capable of predicting loan defaults. The classifiers used in model are "Logistic Regression", "Support Vector Machines", "Naive Bayes" and "Random Forest". The model for each of these algorithms was optimized to prioritize precision ahead of accuracy, to identify good credits. The initial dataset had to be oversampled to accommodate the skewness of 18.26% default to 81.74% paid-off. The best precision was obtained by "Logistic Regression" 95.9% with 23.9% accuracy. In this model, tree

978-1-5090-3471-0/16/$31.00 ©2016 IEEE

based ensemble classifiers are used to better the statistical inference on 2013-2015.

Ensemble learning classifiers for credit scoring [7], [8] involves creating multiple models for subsets of data and combine them to improve the overall performance of the base algorithm. "Random Forest", "Extra Trees" and "Bagging" are ensemble classifiers using "Decision Tree" as a base model. The randomness and sampling produces more accurate models.

"Extra Tree Classifier" is an ensemble classifier which randomizes the splits on the generated trees instead of optimizing it like "Random Forest", can also be used for feature selection. The best subset of attributes is selected with respect to splits, from the generated forest.

III. METHODOLOGY

A. Dataset Description

The dataset consists of 656,724 loan records which were issued by "LendingClub", between the years 2013 and 2015. There are a total of 115 attributes describing the loan application. The "Loan Status" attribute which describes the current state of the loan, has the following values, "Issued", "Current", "Fully paid", "Default", "Charged off", "Late (16-30 days)", "Late (31-120 days)" and "In grace period". These statuses are used to reduce them to a binary classification problem, i.e., the loan applications with "Charged off", "Default", "Late (31-120 days)" and "Late (16-30 days)" are considered as "bad" or "defaulted" loans while "Current", "Fully Paid" and "In grace period" are classified as "good" loans and remaining are ignored. The "loan status" attribute is replaced with "is bad" which takes the values of 0 for a good credit and 1 for a bad credit or default.

The loan amount ranges from $1000 to $35,000 and each loan has a "grade" (ranging from A-G) associated with them. The grade specifies the range of interest rates in ascending order, starting from 5.32% to 29%. As expected, the loans with higher interest rate have a higher risk of default. 31% of Loans in grade G are bad while it's only 3% in grade A.

B. Preprocessing and Cleaning

The publicly available "LendingClub" dataset for the years 2013-2015 obtained from the official website was presented in a "Comma-Separated-Value" file. The dataset was split as two sets, 2013-2014 and 2015, having 235,629 and 421,095 records, 111 and 74 attributes respectively. Only the attributes which are common to both the datasets are considered, since "LendingClub" has removed attributes which are not significant for the analysis and added a few new ones.

The records which had "loan status" as "Issued" are ignored for analysis. Since this research was carried out in early 2016, the 2015 dataset had only 60,000 records that could be classified into either "good" or "bad" credit, whilst the remaining were in "Issued" state. 9.3% of 2013-14 dataset was also in the same bracket and hence ignored. The preprocessed data contains of 10.9% credit defaults.

Attributes having non-numerical values like "Home ownership", "Application type", "Employee title" are labeled with numbers starting from 1. The combined dataset has 279,169 with 70 attributes.

C. Feature Selection

1) Removing Redundant Attributes

The attributes that carried same or similar information are removed. Attributes like "Funded amount", "Funded amount by investors" were found to be the same as loan approved amount in most of the records. Similarly, a total of 16 attributes are eliminated.

2) Missing Values

Attributes with 80% (and above) missing values are ignored, as most of them were found to be optional. The remaining attributes with 10-15% of missing data, are handled by using median of the available data, as the placeholder. There are 25 such attributes ignored in this process.

3) Removing Excessive Attributes

Attributes such as "Last payment date", "Last credit pull date", and "Outstanding principal amount" are populated only when the borrower has started repaying it. Training the model with these attributes might lead to undesired results, like extremely high accuracy and precision (close to 100%), as it allows the algorithm to learn from the future. Hence 14 such attributes are removed from the dataset.

4) Using Domain Knowledge

An intuitive approach was taken, on the remaining attributes, with an objective to eliminate the least important ones. "Initial list status", "Address state", "Public records" and "Loan description" are ignored for classification. A total of 6 least important attributes are ignored.

5) Extra Tree Classifier

Extra Tree or Extremely Randomized Forest classifier is a tree based ensemble classifier that uses Decision Tree as the base learner. It selects samples from entire dataset during attribute split at the node, at random, for every combination in the feature set and choosing the optimal fit amongst them. This cheaper to train as the node splits are completely randomized as opposed Random Forest algorithm, where the split creation is optimized.

D. Classification

1) Decision Tree

Decision Tree classifier is a supervised learning algorithm which builds a binary tree where each node level has an attribute value split. A Decision Tree is built top-down from a root node and involves partitioning the data into subsets that contain similar values. Entropy is a measure that is used to calculate the homogeneity. A completely homogeneous sample set will have entropy value of 0 and an evenly distributed sample will have the value 1 as its entropy. Information gain at each level or attribute is calculated by using the entropy of the parent and the weighted sum of entropy of its children. This information gain value acts as the split at each node. This tree represents the set of rules used for predicting.

The decision tree obtained for the "LendingClub" dataset has the attribute "Interest rate" as the primary split as it has the highest information gain followed by "Debt to income ratio", "Installment", "Annual income" and "Loan amount". The resulting model gives us 97.1% precision and 81.3% accuracy.

TABLE I. CONFUSION MATRIX FOR DECISION TREE

	Predicted Positive	Predicted Negative
Condition Positive	202586	46154
Condition Negative	6051	24378

2) Random Forest

Random forest classifier is another tree based ensemble supervised learning algorithm that generates multiple Decision Trees (forest) for random subsets of the data, and predicts the class with highest frequency after running the sample on all the Decision Trees generated.

Random forest helps in overcoming the over-fitting problem experienced in Decision Tree classification. It provides a significant increase in the accuracy of the model. The trees generated had "Interest rate" as the primary split just like Decision Tree but the remaining attributes varied due to randomness in the subsets. This model has a precision of 96.2% and 88.5% accuracy.

TABLE II. CONFUSION MATRIX FOR RANDOM FOREST

	Predicted Positive	Predicted Negative
Condition Positive	225545	23195
Condition Negative	8909	21520

3) Bagging (Bootstrap Aggregating)

Bootstrapping is a sampling technique to obtain an approximate statistic estimate of the sampling distribution by choosing a random subset with replacement, multiple times and learning from it.

Bagging involves bootstrapping the original dataset and using the subsets to estimate the performance another classifying algorithm, by voting. The randomness introduced in bootstrapping reduces the variance or the over-fitting problem.

TABLE III. CONFUSION MATRIX FOR BAGGING

	Predicted Positive	Predicted Negative
Condition Positive	224856	23884
Condition Negative	8639	21790

Decision Tree was used as the base classifying algorithm or estimator for bagging. A precision of 96.3% and an accuracy of 88.35% was achieved for bagging. The random sampling induced results in better performance in identifying defaults as the randomized samples might contain a healthy number of true negatives, that enables the base Decision Tree algorithm to extrapolate the defaulting patterns.

IV. ALGORITHM PERFORMANCE ANALYSIS

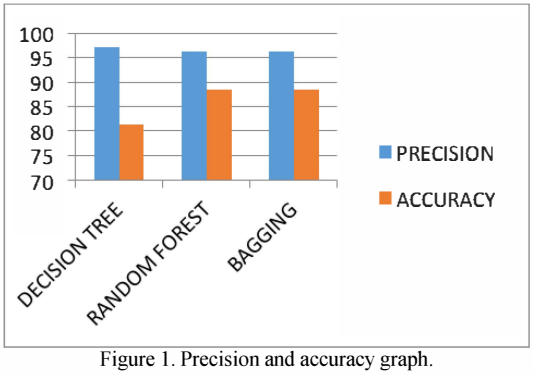

Figure 1. Precision and accuracy graph.

Improving predicting power (precision) using Decision Tree, Random Forest and Bagging algorithms was the highlight of this paper. Decision Tree was the best as shown in the graph with the precision of 97.1% as compared to Random Forest and Bagging having precision of 96%. The precision of the latter is lower due to the randomness and the skewness involved in the subsets used for generating the trees. Few of these subsets have favored the identification of defaults ahead of the good credits. Hence much higher accuracy of 88.5% and 88.35% was obtained in the ensemble classifiers.

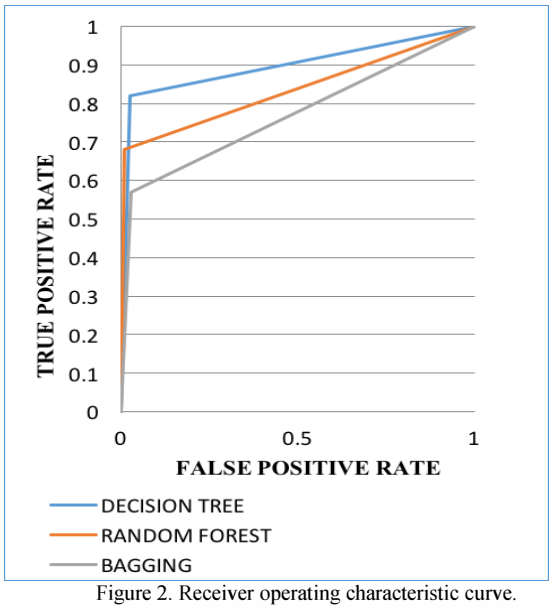

Figure 2. Receiver operating characteristic curve.

Receiver operating characteristic (ROC) [9] curve is a 2D plot representing the performance of a binary classifier; the curve is plotted with true positive rate (TPR) versus the false positive rate (FPR) at multiple threshold values.

$$TPR = \frac{\sum True\ Positive}{\sum True\ Positive + \sum False\ Negative} \quad (1)$$

$$FPR = \frac{\sum False\ Positive}{\sum False\ Positive + \sum True\ Negative} \quad (2)$$

The following is the algorithm's performance for a given range of area under curve (AUC):

TABLE IV. AUC PERFORMANCE COMPARISON

AUC ranges	Performance
0.90 to 1.0	Excellent
0.80 to 0.90	Good
0.70 to 0.80	Fair
0.60 to 0.70	Poor
0.50 to 0.60	Failure

The AUC for Decision Tree is 0.91 which is larger compared to Random Forest's 0.83 and Bagging's 0.77. This confirms that Decision Tree model will perform well while classifying a good credit.

V. CONCLUSION

The credit risk involved in peer-to-peer lending system of "LendingClub" has been minimized, by identifying the correct set of features during data preprocessing and using fine-tuned tree based ensemble machine learning classifiers like Decision Tree, Random Forest, Extra Trees and Bagging, to achieve high precision and accuracy. Although Decision Tree has the highest precision, the accuracy of Random forest is 88.5%, i.e. 7.2% higher than Decision Tree. This makes the Random Forest predictor model better in identifying the defaults, while Decision Tree is more powerful in finding the good credits. This concludes that the overall return on investment will be high as the model identifies most of the good credits whilst identifying potential defaults.

VI. FUTURE WORK

A. Exploring Artificial Neural Network techniques like Feedforward, Backpropagation on "LendingClub" dataset. Artificial Neural Networks [10] are capable of finding hidden layers and patterns that can improve the predicting power and help us determine defaulting loans.

B. Performance analysis on latest datasets and on other peer-to-peer lending systems with the same set of features and model.

C. Explore probability based classifiers like Naive Bayes and Anomaly detection algorithms with features following Gaussian distribution.

REFERENCES

[1] T. Van Gestel, B. Baesens, Credit Risk Management: Basic Concepts: Financial Risk Components, Rating Analysis, Models, Economic and Regulatory Capital: Oxford University Press, 2008.

[2] J.R. Quinlan, "Induction of Decision Tree", Machine Learning Journal, Volume 1, Issue 1, Pages 81-106, 1986.

[3] L. Breiman, "Random Forest", Machine Learning Journal, Volume 45, Issue 1, Pages 5-32, 2001.

[4] L. Breiman, "Bagging Predictors", Machine Learning Journal, Volume 24, Issue 2, Pages 123-140, 1996.

[5] P. Geurts, D. Ernst., and L. Wehenkel, "Extremely randomized trees", Machine Learning Journal, Volume 63, Issue 1, Pages 3-42, 2006.

[6] K. Tsai, S.Ramiah, S. Singh, "Peer Lending Risk Predictor", Stanford University, 2014.

[7] F.N Koutanaei, H. Sajedi, M. Khanbabaei, "A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring", Journal of Retailing and Consumer Services 27, Pages 11-23, 2015.

[8] G. Wang, J. Mac, A hybrid ensemble approach for enterprise credit risk assessment based on Support Vector Machine, Expert Systems with Applications 39 (2012) 5325–5331, 2012.

[9] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters 27, 2006.

[10] E. Angelini, G. di Tollo, A. Roli, A Neural Network Aapproach for Credit Risk Evaluation, The Quarterly Review of Economics and Finance 48(4): Pages: 733-755, 2008