PROPENSITY TO DEFAULT MODEL FOR ONLINE CREDIT MARKETPLACE

Pavanraj Talawar
PGDM in Business Analytics,
REVA University, Bengaluru
Pavanrajt.BA01@reva.edu.in

Lakshmi D
PGDM in Business Analytics,
REVA University, Bengaluru
Lakshmid.BA01@reva.edu.in

ABSTRACT

Default risk is the chance that companies or individuals will be unable to make the required payments on their debt obligations. Lenders and investors are exposed to default risk in virtually all forms of credit extensions. In the event of a default, lenders may lose out on periodic interest payments and in many cases, the entire principal amount. Our objective of this study is to build a 'Propensity to Default' model which can predict the probability of default. We have also tried to understand the important variables that explain the default behaviour. For the purpose of our study, we have used the loan approval data of Lending Club, an online credit marketplace. Lending Club is the world's largest online credit marketplace, facilitating personal loans, business loans, and financing for elective medical procedures. The company operates fully online with no branch infrastructure, and use technology to lower the cost. The predictive models are built on the actual data collected on the customers of this lending platform over a period of 4 years (2007 - 2011). We have used various Python libraries to perform detailed Exploratory Data Analysis (EDA) to display the vital statistics of different features. We have also built multiple models (Logistic Regression, Random Forest, Deep Learning etc.) to compare and select the most appropriate and accurate model. These predictive models can be used by the online credit marketplaces to optimize their loan repayment success by targeting the right borrowers and modifying their loan structure if required.

Keywords: *Default risk, Propensity to Default, Lending Club*

---

1. INTRODUCTION

Peer to peer lending, called, P2P lending, has grown rapidly in recent years where markets allow individuals to borrow and lend money without financial institutions acting as intermediaries (Zhang et.al. 2017). In 2014, the online lending volume in the US marketplace

was around $5.5 billion and the accounting firm PWC expects the same volume to grow $150 billion by 2025. Another study by Juniper titled Fintech Future: Market Disruption, Leading Innovators & Emerging opportunities 2016-2021, the online lending revenues will be doubled and will exceed $10 billion globally by 2020. However, due to the high volume of online applicants, as well as due to the difficulty of verifying each of the online applications, online lenders are facing the issue of handling fraudulent applications (Identity Mind Global report, 2015).

While there is a steady increase in volume in the P2P lending segment, there is a probable estimation that an average default rate is 8.38% in 2012, 8.87% in 2013 and 8.62% in 2014. Due to which P2P online lending market have an estimated unpaid loans of $74 million in 2012, $257 million in 2013 and $569 million in 2014.

It is estimated that in P2P lending the default and fraud rates – the percentage of loan losses caused by fraud increases to approximately 17.4% (Identity Mind Global report, 2015).

2. CHALLENGES

According to the research from the credit-reporting firm the TransUnion, Loan stacking or taking multiple loans from different lenders has doubled between 2013 and 2015 which also include creating online fraudulent activities. Borrowers might receive a part of the loan which is originally approved by one lender and to cover the same, the borrower procures the remainder of the requested amount form the other lenders and where they end up paying higher interest rates (Pat, Phelan, 2016).

In a report by the Wall Street Journal (Telis, Demos, 2016), "an average day about 4.5% of people who take an unsecured personal loan seek for other lenders later on the same day and turn as a defaulter". Due to this, many lenders are discouraging loan-stacking as it increases the probability that a borrower will default. A borrower who applies for the second loan within 15 days of the first loan is identified as four times more likely to be a defaulter with no intent to repay the loan and ten times more likely to be fraudulent if the same borrower applies for a third loan application. Another major issue with online lending industry is synthetic fraud, a fraudster can also use stolen identities, documentation or hacked profiles to apply for a fraudulent loan application.

Even though a high level of fraud and risks are involved in the rapidly growing P2P lending industry, however by taking a few necessary preventative and precautionary steps listed below, lenders can reduce the frauds.

1. Lenders can reach the third party agencies to ensure a borrower's data and credit history so that they can take a more informed decision.
2. Instead of just relying on the underwriting and technology, lenders can take experts advice to get a more holistic decision which can help the lenders to get a better accuracy rate.

3. By using modern technology, lenders can predict the bad loan behaviours and patterns hence by implementing a proactive data-driven anti-stacking strategy in the online loan management software lenders can reduce their loss.
4. To prevent fraud, online lenders require a complete compliance framework including Sanctions Screening, pre-populated Suspicious Activity Report (SAR) filing, Geo-Fencing, Anti-Money Laundering (AML) monitoring and advanced case management, to comply with existing regulations.

3. LITERATURE REVIEW

P2P loans do not have any kind of guarantee fund, so it's a risky activity for individual lenders which would transfer as credit risk. A credit score is a number which represents an assessment of the creditworthiness of a person or which defines a person's willingness and ability to repay. Klafft (2008) did his analysis in Prosper platform where it shows that credit rating has the most significant impact on interest rate while the debt-to-income ratio is less significant. Also his analysis suggested that verified borrower's bank account or verified home ownership are not at all impacting on interest rates. However, an existence of a borrower's bank account plays a significant role in determining to fund the loan.

A few of the researchers (Berger et al, 2009) had focused on stakeholders and other determining factors and they found that early lenders do not understand the market risk but over time they become effective in reducing the risk. There are few researchers who gave importance to P2P websites which act as an intermediary between borrowers and lenders and bring these group together. Sometimes these groups form small communities and concentrate their interests as being in small groups help the financial markets to understand different regulatory restrictions in different countries (Herrero-Lopez, 2009).

A research by (Galloway, 2009) suggested there is a necessity of bank involvement but only be restricted to facilitate the process of lending. However, data confirmation is required by a credit bureau or any other external agencies which vary from country to country, as per respective country regulations.

Colliers et.al., in their paper gave importance in distinguishing between financial and demographic characteristics of the borrower, also social relationships with friends and other group intermediation. This research gave insights on how the respective factors affect the borrower's likelihood of successful funding of the interest rate. Few studies show that the loan interest rate decrease only for the active bidding of the group leader (Collier et. al., 2009). However, Freedman & Jin (2008) did not agree with these findings instead, a combination of group leaders bidding, actually increase the average interest rate.

Pope et.al., (2008) did a research on gender and age-based – male or female and statistical discrimination. Discrimination occurs when lenders offer higher interest to the old age people as this age group people statistically proved higher default rate than other borrowers (Phelps, 1972).

As per research by Barasinska (2009), female lenders are less risk averse than male lenders. This research also shows that the female lenders are funding a loan to the borrower with lower interest rate and lower credit ratings. In another study by Petersen (2004) shows the difference between the soft and the hard information. In his study he describes soft information as "difficult to completely summarize in a numeric score" in contrast to hard factors like financial data of a borrower. Again, in another research Lin (2009) describes the "soft credit information" as the detail about the borrowers' risk generated by his or her social network in the P2P lending in the community.

4. DATA DESCRIPTION

The dataset used for analysis is the customer data from the world's largest P2P lending platform, Lending Club. The company enables borrowers to create unsecured personal loan listings on its website for any amount from $1,000 to $40,000. Based on different factors like the borrower's credit score, credit history, loan amount etc, the Lending Club assigns credit grades which determine the fees amount and interest rate. The repayment duration is either three years or in some cases five years with additional fees and higher interest rates.

The dataset contains details of ~40,000 approved loans during the period of 4.5 years (June 2007 to December 2011). The loan amount, term/duration of the loan, interest rate, grade, annual income, issue date, the purpose of the loan, total payment made and loan status are few of the important variables available in this public dataset. It has 39,786 observations with 137 features.

5. DATA PREPARATION

Out of the 137 features, 85 have more 60% of null values. 9 of them do not have any variation and have only one unique value across all observations which does not contribute to the proposed predictive model. Individual features are then updated and formatted to reflect the correct type of data. The feature 'term' defines the duration of the loan in months. The word 'months' is removed from the observations within this feature and the feature is converted to float from a string. A similar exercise is performed on the 'int_rate' and 'emp_length' variables to convert them to the correct data type, float. After cleaning the data, the final dataset contains 37,900 observations with 31 features.

The dependent/target variable is the 'loan_status' feature. This provides the details of whether the customer has fully paid the loan, or if the loan has been charged off. The overall split between 'Fully Paid' and 'Charged Off' is 86% and 14%. The dependent variable is encoded to 0 and 1 to be used in building the model.

The 4 important categorical variables (grade, home_ownership, ver_status, purpose) are encoded using dummy variables to be used for building the model. The final dataset available for modelling had 37,900 observations and 51 features. The following heat map shows the correlation between the different features. The target variable, loan_status is highly correlated with the features, 'recoveries' and 'collection_recovery_fee'.

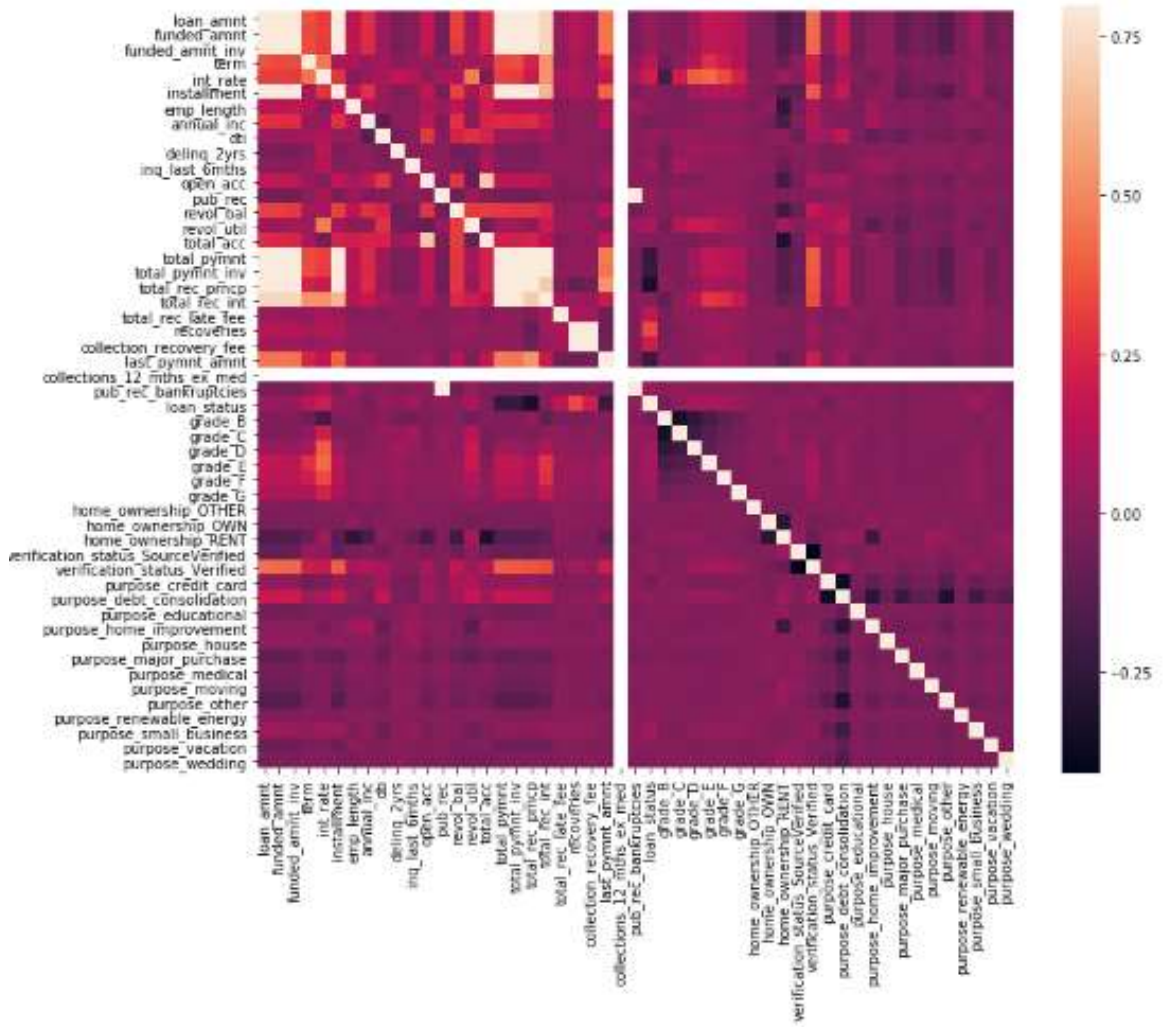

Figure 1 Correlation Matrix

6. METHODOLOGY

The objective of this research paper is to build different models to predict the propensity of default if the loan is approved to a borrower and to identify the best predictive among the various models. To build various models, the first step is to identify the most important features that contribute towards the model. Feature selection is important as it reduces the risk of overfitting while improving the accuracy of the model and reducing the processing time. Out of the many different techniques used for feature selection, the method used for this research is by identifying the Gini importance of features using Random Forest predictive model.

As quoted by Leo Breiman and Adele Cutler (2012) in their article on Random Forests states, “Every time a split of a node is made on variable m the Gini impurity criterion for the two descendent nodes is less than the parent node. Adding up the Gini decreases for each individual variable over all trees in the forest gives a fast variable importance that is often very consistent with the permutation importance measure”. The following table provides the list of the top 15 features according to their Gini importance value. As already observed in the correlation matrix, ‘recoveries’ and ‘collection_recovery_fee’ top the list with the highest Gini importance values.

The Python script with the models and datasets are uploaded to GitHub. The link can be accessed here.

	Gini_Imp	Features
21	0.462551	recoveries
22	0.246749	collection_recovery_fee
18	0.088241	total_rec_prncp
16	0.039087	total_pymnt
0	0.025616	loan_amnt
2	0.024526	funded_amnt_inv
17	0.022163	total_pymnt_inv
23	0.019128	last_pymnt_amnt
1	0.018672	funded_amnt
5	0.015381	installment
20	0.007852	total_rec_late_fee
19	0.006476	total_rec_int
4	0.006240	int_rate
3	0.004719	term
14	0.001751	revol_util
13	0.001610	revol_bal

Table 1 Important features

The top 15 features contribute towards 99% of the feature importance. These 15 features are now used to build different predictive models. However, before proceeding with building the models, Multicollinearity has to be identified between features and remediated. This can be addressed using finding the Variance Inflation Factor (VIF) for all the features. The feature with the highest VIF is removed to reduce the multicollinearity. This step is repeated multiple times until the VIF for all the features is less than 5. The final 9 independent variables that can be used to build the model are as follows.

	VIF Factor	features
0	31.215247	Intercept
5	3.396997	total_rec_int
1	3.031724	recoveries
4	2.995397	loan_amnt
2	2.883735	collection_recovery_fee
8	1.947485	int_rate
6	1.584668	term
3	1.466769	last_pymnt_amnt
9	1.328504	revol_util
7	1.028356	total_rec_late_fee

Table 2 Variance Inflation Factor (VIF)

The target and predictor variables are first separated into two different datasets, X and y. The predictor variables are normalized so that all the variables are on the scale of 0 to 1. The dataset is then split into training dataset (70%) and test datasets (30%). The following 8 different predictive models are built and validated against the test datasets.

- Logistic Regression
- Random Forest
- Deep Learning
- Support Vector Machines
- KNN
- Naive Bayes
- Linear SVC
- Decision Tree

The comparison of the ROCR and AUC is provided below along with the confusion matrices.

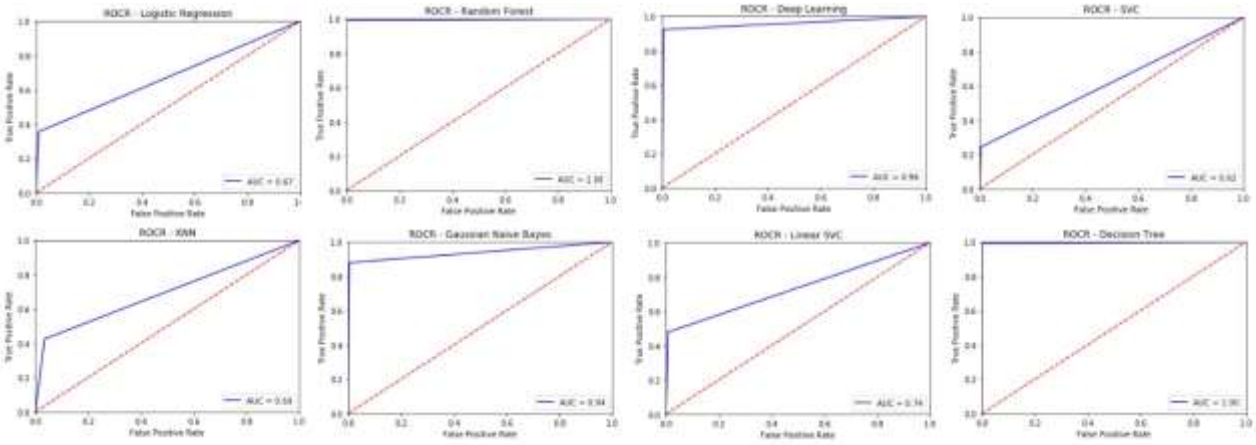

Figure 2 Comparison of ROC curve and AUC

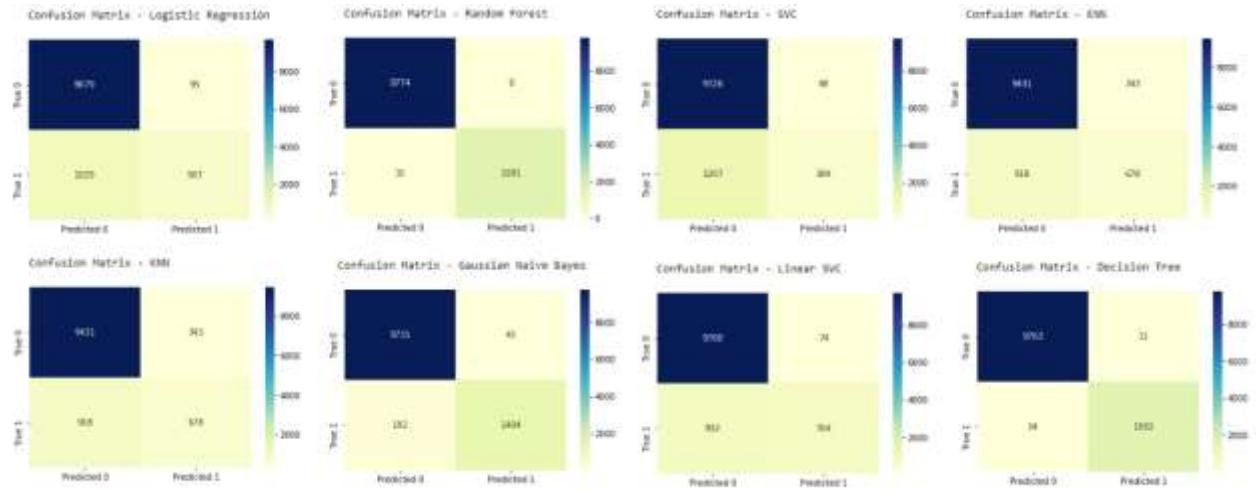

Figure 3 Comparison of Confusion Matrices

7. FINDINGS/DISCUSSION

While comparing classification models, precision and recall scores are a better indicator than the accuracy score. Accuracy is defined as the correct number of predictions out of the

overall predictions. As the number of defaulters (loan_status = 1) is very less compared to the non-defaulters (loan_status =0), the overall accuracy is largely decided by the percentage of non-defaulters. However, to build a good propensity to default, predictive model, the number of correct predictions of defaulters is more important that the correct prediction of non-defaulters.

Precision is the number of correct predictions of defaulters out of the total predictions as defaulters. i.e., True Positive / (True Positive + False Positive). The recall is the number of correct predictions of defaulters out of the total number of defaulters. i.e., True Positive / (True Positive + False Negative). The following table provides a comparison of these scores for all the 8 predictive models.

	Model	Precision_Score	Recall_Score	Score_Accuracy	Value_AUC
1	Random Forest	1.000	0.991	0.999	0.995
7	Decision Tree	0.992	0.991	0.998	0.995
2	Deep Learning	0.970	0.927	0.986	0.961
5	Naive Bayes	0.970	0.880	0.979	0.938
6	Linear SVC	0.912	0.479	0.920	0.736
3	Support Vector Machines	0.890	0.244	0.890	0.619
0	Logistic Regression	0.856	0.355	0.901	0.673
4	KNN	0.664	0.425	0.889	0.695

Table 3 Precision, Recall, Accuracy and AUC comparison

8. Conclusion/Implications

Based on the precision and recall scores of the various models, it is evident that the tree-based predictive models perform much better in comparison to the Logistic Regression or KNN models. While the deep learning model's performance is also very good, the time and resource utilisation of this model is very high. The online credit lending marketplaces can utilise either the Random Forest or Decision Tree predictive models to identify the potential defaulters and minimise their risk.

BIBLIOGRAPHY

Bachmann, A., Becker, A., Buerckner, D., Hilker, M., Kock, F., Lehmann, M., ... & Funk, B. (2011). Online peer-to-peer lending-a literature review. *Journal of Internet Banking and Commerce*, *16*(2), 1.

Barasinska, N. (2009). The role of gender in lending business: evidence from an online market for peer-to-peer lending. *The New York Times, 217266*, 1-25.

Collier, B. C., & Hampshire, R. (2010, February). Sending mixed signals: Multilevel reputation effects in peer-to-peer lending markets. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work* (pp. 197-206). ACM.

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. In Ensemble machine learning (pp. 157-175). Springer US.

Freedman, S. M., & Jin, G. Z. (2011). *Learning by Doing with Asymmetric Information: evidence from Prosper. com* (No. w16855). National Bureau of Economic Research.

Galloway, I. (2009). Peer-to-peer lending and community development finance. *Community development investment centre working paper*, (39), 19-23.

Herrero-Lopez, S. (2009, June). Social interactions in P2P lending. In *Proceedings of the 3rd Workshop on Social Network Mining and Analysis* (p. 3). ACM.

Klafft, M. (2008). Peer to peer lending: auctioning microcredits over the internet. Proceedings of the 2008 International Conference on Information Systems, Technology and Management (ICISTM 08), March, Dubai, United Arab Emirates

Kumar, S. (2007). Bank of one: Empirical analysis of peer-to-peer financial marketplaces. *AMCIS 2007 Proceedings*, 305.

Phelps, E. S. (1997). 'The Statistical Theory of Racism and Sexism', American Economic Review, 62, 659-61. *INTERNATIONAL LIBRARY OF CRITICAL WRITINGS IN ECONOMICS*, *81*, 551-553.

Pope, D. G., & Sydnor, J. R. (2011). What's in a Picture? Evidence of Discrimination from Prosper. com. *Journal of Human Resources*, *46*(1), 53-92.

Zhang, K., & Chen, X. (2017). Herding in a P2P lending market: Rational inference OR irrational trust?. *Electronic Commerce Research and Applications*, *23*, 45-53.

WEBLIOGRAPHY

https://blog.identitymindglobal.com/online-lenders-an-emerging-target-for-fraud
Last accessed on 05/11/2017

http://www.businessinsider.com/online-lending-is-making-fraud-easier-2016-10?IR=T
Last accessed on 05/11/2017

https://www.pwc.com/us/en/consumer-finance/publications/assets/peer-to-peer-lending.pdf
Last accessed on 05/11/2017

https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#giniimp
Last accessed on 06/11/2017

Pat, Phelan (2016, October 26th), Fraud in the Digital Age: Loan Stacking and Synthetic Fraud https://www.transunion.com/blog/fraud-in-the-digital-age-loan-stacking-and-synthetic-fraud (Last accessed on 24/10/2017)

Telis, Demos. (2016, October 27th), Borrower or Fraudster? Online Lenders Scramble to Tell the Difference. https://www.wsj.com/articles/borrower-or-fraudster-online-lenders-scramble-to-tell-the-difference-1477580637 (Last accessed on 24/10/2017)