Gashaw Hailemichael

Gashawh1@umbc.edu

Credit Risk Analysis and Prediction of Loan Defaults

Spring 2020

1. **Introduction**

Credit risk is the possibility of a loss resulting from a borrower's failure to repay a loan or meet contractual obligations. In order for lenders help to minimize the possible losses and increase the volume of credits, Credit Risk assessment is a crucial to evaluate if a loan applicant can be a defaulter at a later stage so that lenders can go ahead and grant the loan or not. This project focuses to analyze the credit risk involved in peer-to-peer (P2P) lending system of Lending Club Company based in the United States, in which investors provide funds for potential borrowers and investors earn a profit depending on the risk they take. Machine learning algorithms and preprocessing techniques will be used to explore, analyze and determine the factors, which play crucial role in predicting the credit risk involved in Lending Club. The dataset used existed in the open source Kaggle website through the URL link,<https://www.kaggle.com/wendykan/lending-club-loan-data>that ranges from 2007 to 2015. It has 890 thousand records and 75 features.

1. **Statement of the problem**

In the financial industry, lending loans has been playing a significant role and it is quite beneficial for both the lenders and the borrowers. However, evaluating the creditworthiness of loan applicants is a common challenge typically where loans are unsecured. In decision-making process, it is a tough problem for investors to decide whether or not to fund a particular loans. Therefore, it is important for the lending industry to apply risk management by providing investors with comprehensive risk assessment. This study proposes solutions with the goal at helping potential investors in their decision making process.

1. **Related Work**

There have been many studies on classification models predicting LendingClub loan default. According to the data, examining borrowers’ characteristics and their influence on borrowers’ default is important. The study shows in both paper are in consensus that the Credit Grade assigned by Lending Club is the best predictor for borrowers’ default. Moreover, Revolving Credit Line Utilization is another variable influencing the default rate mentioned in all. Findings of other default determinants vary. The first factor is the selection of variables potentially having an impact on borrowers’ default. For instance, both references [1 & 2] found out that the FICO score has an influence on default. In the research paper, the dataset used also another factor creating differences between the findings for instance, the timeframe. Reference [1] used only 36-month loans. Reference. [2] Used both, 36- and 60-month loans also classification of loan status or type of loan length might be the cause. Borrower characteristics can be used to predict a loan’s default chances. However, since default determinants depend on the loan’s risk class, caution is warranted. What seems to be a good predictor of loan default based on overall data may not be reliable in the highest loan risk class.

1. **Methodology**

The focus on this study would be understanding which features are most relevant or most predictive for separating those loans likely to be paid off from those loans likely to be defaulted. The level of default risk will be best predicted with predictive modeling using machine-learning algorithm techniques. Data Processing, Exploratory Data Analysis (EDA) and data Visualization would be performed using Python.

1. **Purpose of Study**

Using Lenders Club dataset, the goal is to provide a data product that enables investors to avoid loans likely to default. This will be achieved by conducting EDA and by developing a model using various machine-learning techniques that predict the probability of default for a potential loan to avoid loans that are predicted to default.

1. **Data Cleaning**

In order to achieve our goal, it is important to identifying the relevant features from a set of data and removing the irrelevant or less important features, that does not contribute much to our target variable. Such as dropped the columns, which had more than 50% of missing data and imputed the remaining attributes of missing data using median for continuous variables and mode for categorical variables. Also removed the columns that has no direct relation to our analysis such as ID’s and other unique identifiers. Removing redundant attributes that carried same or similar information values for employment length – such as 10+ years. We labeled our data based on the features provided in data. “Default", "Charged Off", "Late (31-120 days)", "Late (16-30 days)", "Does not meet the credit policy. Status:Charged Off“ were considered as defaulted loans. Our dataset is heavily skewed for the loan to be defaulted while there are only 61,176 Default, there are 209,711 Fully Paid. This is imbalanced data and we will correct this situation during the data preprocessing stage. If we leave the balance heavily skewed towards Fully Paid, our model will rarely ever predict a loan as Default. Since there are only two options for a loan (Default/Fully Paid), we want the distribution in our dataset to be balanced.

We dropped 22 features which were irrelevant for our analysis, some of those where ID attributes which had a different value in each cell or some attributes had just one unique value.

Also 3 features has been dropped because they had a strong correlation to another feature. The correlations were identified using Pearson Correlation matrix.

We would only keep features which are known at the time of loan application using which we can find predictive solutions for helping investors to determine whether they should invest in a given loan based on how likely it is that the loan will be defaulted and whether the loan’s borrower is likely to be a good borrow. Therefore, we removed 20 additional attributes which have meaning only after a loan has been funded by investors.

1. **Exploratory Data Analysis**

It is difficult to understand the data set and make conclusions without looking through the entire data set. Understanding and interpreting data from large data sets can be very challenging. Spending more time exploring the data to get a better representation of the data set is important. The very first step in exploratory data analysis is to identify the type of variables in the dataset. Variables are of two types Numerical and Categorical. Once the type of variables is identified, the next step is to identify the features (input) and Target (output) variables.

In this project, Exploratory Data Analysis is majorly performed using the following methods [1]

* Univariate visualization : to provides summary statistics for each field in the raw data set
* Bivariate visualization : to find the relationship between each variable in the dataset and the target variable of interest
* Dimensionality reduction: to understand the fields in the data that account for the most variance between observations and allow for the processing of a reduced volume of data.

The target variable is loan status, and we would like to know whether the loan would be defaulted or not using the selected features from the dataset. Our goal is to become confident that the data set is ready to be used in a machine-learning algorithm. [2]

1. **Data transformation**

In this step, we have performed one-hot encoding on the categorical variables and then scaled the numerical columns; there are many ways to convert categorical values into numerical values. Each approach has its own trade-offs and impact on the feature set. We use one-hot transformation for the categorical variables as well as scale the numerical columns using the StandardScaler module from the sklearn preprocessing module.

1. **Model**

This project focused on building different machine learning algorithms and evaluating the performance of each models. The resulting dataset was randomly split into two data frames, with one consisting of 80% of the instances for training set and the other consisting of 20% of the the instances for test set. Since our data is an imbalanced class, we have used oversampling the minority class by Synthtic Minority Over-Sampling Technique (SMOTE) in the training set. The classification accuracy is not a good measure for imbalanced classification which is very common situation in investment and banking sectors, and could lead decision makers to make wrong decisions. For these kind of models, the better measure is the precision/ recall output and ROC-curve, which shows true positive rates against false positive rates.

For our dataset, we use the following Machine learning algorithms,

* Naïve Bayes
* Logistic Regression
* Random Forest

**Naive Bayes**

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes’ theorem with the “naive” assumption of conditional independence between every pair of features given the value of the class variable. Despite its simplicity, the Naive Bayesian classifier often does widely used because it often outperforms more sophisticated classification methods.

**Logistic Regression**

[Logistic regression](http://www.statisticssolutions.com/academic-solutions/membership-resources/member-profile/data-analysis-plan-templates/data-analysis-plan-logistic-regression/) is the appropriate regression analysis to conduct when the dependent variable is binary. It is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

**Random Forest**

Random forests is an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the mode of the classes. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction.

1. **Results**

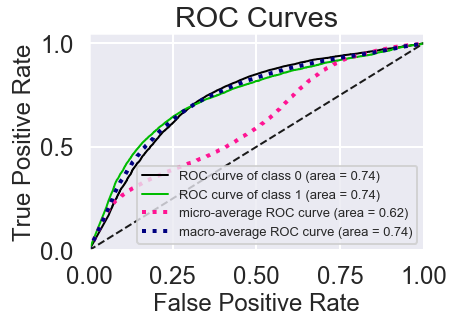
In this project, a comparative analysis of four different classification algorithms has been done on Lending Club Loan dataset. Our goal is to select an algorithm, which the most accurately predicts if a loan will be defaulted or fully paid. The algorithms that we compared are; Naïve Bayes, Logistic Regression and Random Forest. Random Forest is an ensemble method that combine several learning algorithms to make better predictions. In the final model result, it was found that Random Forest algorithm showed the highest accuracy of 97.4%, while Logistic Regression algorithm showed the second highest accuracy of 94.6% and Naïve Bayes showed the lowest accuracy of 54.1%. The results of implementing the models are summarized below from the classification report as well as ROC (Receiver Operating Characteristic) curve output of AUC (Area under ROC curve.

**Naïve Bayes**

When the model run on the 20% of test dataset, a ‘Default’ recall of 84% and a ‘Fully Paid’ precision of 91% were achieved. The recall score of Naïve Bayes tells us that we can use this model to classify 84% of the bad loans that could end up to be defaulted. [4]

Performance of Naïve Bayes Model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Naïve Bayes | Precision | Recall | F1 Score | Support |
| Fully Paid | 0.91 | 0.46 | 0.61 | 41867 |
| Default | 0.31 | 0.84 | 0.46 | 12311 |



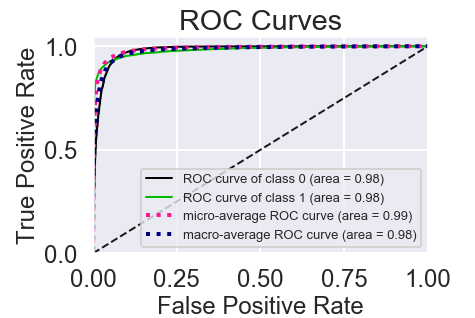
The ROC curve for Naïve Bayes shows that AUC of 0.74 for loan to be defaulted.

**Logistic Regression**

When the model was run on the 20% test dataset, we get a ‘Default’ recall score of 92% and a ‘Fully Paid’ precision of 98% were achieved. The recall score of logistic regression tells us that we can use this model to classify 92% of the bad loans that could end up to be defaulted. [4]

Performance of Logistic Regression Model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Logistic Regression | Precision | Recall | F1 Score | Support |
| Fully Paid | 0.98 | 0.95 | 0.96 | 41867 |
| Default | 0.85 | 0.92 | 0.89 | 12311 |



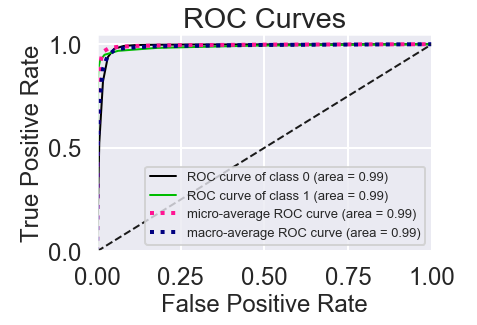
The ROC curve for Logistic Regression shows that AUC of 0.98, which is better than the Naïve Bayes model.

**Random Forest**

When the model was run on the 20% test dataset, a ‘Default’ recall of 88% and a ‘Fully Paid’ precision of 97% were achieved. The recall score indicates that the Random Forest can classify 88% of the applicants that would be default will be appropriately identified.

Performance of Random Forest Model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Random Forest | Precision | Recall | F1 Score | Support |
| Fully Paid | 0.97 | 1.00 | 0.98 | 41867 |
| Default | 0.99 | 0.88 | 0.93 | 12311 |



The ROC curve for Random Forest shows that 0.99, which is the highest of all the three models.

1. **Conclusion**

In this project, we have developed a model using Naïve Bayes, Logistic Regression and Random Forest to predict if a loan could be defaulted. The predictive models created based on a historical data provided by Lending Club and extracted from Kaggle website. Due to imbalanced classification, prediction of Loan Defaults provides excellent performance in predicting true negative instances; however, it provides very poor performance in predicting true positive instances. Therefore, we improved our model performance by applying Synthetic Minority Over-Sampling Technique (SMOTE) to balance the data on the training set.

Comparing our models from other related works, ours have shown better results in precisionas well as recall scores. Three models Logistic Regression, Random Forest and Naïve Bayes classifier successfully created and compared to help investors avoid loans that would end up being defaulted while effectively identifying loans, which are most likely to be “Fully Paid”. The Logistic Regression model provided “Default” recall of 92% and “Fully Paid” precision of 98%. The Random Forest model provided a “Default” recall of 88% and a “Fully Paid” precision of 97%. The “Default” precision has been compromised by the given high performance metrics. The other model tested - Naïve Bayes did not perform well. From the result of ROC cure, AUC for Logistic Regression is 0.98 a little less than the AUC of Random Forest, which is 0.99. In general, ROC curves should be used when there are roughly equal numbers of observations for each class. However, precision/ recall scores should be used when there exist class imbalance. Therefore, we recommend investors to use Logistic Regression algorithm to identify whether a loan to be “Defaulted” or “Fully Paid”.

1. **Potential Future Work**

* Find other features that might help to improve overall model performance.
* Develop grid search to optimize Random Forest model performance.
* Find out other approaches to work around highly imbalanced data.
* Try out different classification algorithms such as XGBoost, neural networks or other classification algorithms that could provide better performance.

**Reference:**

1. Credit Risk by Olivia Labarre, May 2019 retrieved from <https://www.investopedia.com/terms/c/creditrisk.asp>
2. Default rates at Lending Club & Prosper: When loans go bad by Simon Cunningham on October 17, 2014 in [P2P Lending Basics](https://www.lendingmemo.com/category/p2p-lending-basics/) <https://www.lendingmemo.com/lending-club-prosper-default-rates/>
3. <https://www.roselladb.com/credit-risk-analysis.htm>
4. Emekter, R.; Tu, Y.; Jirasakuldech, B.; Lu, M. Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. Appl. Econ. 2015, 47, 54–70.
5. Carmichael, D. Modeling Default for Peer-to-Peer Loans. 2014. Available online: http://ssrn.com/abstract= 2529240 (accessed on 31 August 2018).
6. Determinants of Borrowers’ Default in P2P Lending under Consideration of the Loan Risk Class by Michal Polena and Toblas Regner (2018) <https://www.mdpi.com/2073-4336/9/4/82>

# Exploratory Data Analysis by [Swetha Lakshmanan](https://medium.com/@swethalakshmanan14?source=post_page-----ebdf643a33f6----------------------) 2019 retrieved from <https://medium.com/@swethalakshmanan14/exploratory-data-analysis-in-python-ebdf643a33f6>

# InData Labs, 2017, Exploratory Data Analysis: the Best way to Start a Data Science Project retrieved from

# <https://medium.com/@InDataLabs/why-start-a-data-science-project-with-exploratory-data-analysis-f90c0efcbe49>

# scikit-learn developers, 2007 - 2019, Naïve Bays retrieved from

# <https://scikit-learn.org/stable/modules/naive_bayes.html>

# Statistics Solutions 2020, [Logistic regression](http://www.statisticssolutions.com/academic-solutions/membership-resources/member-profile/data-analysis-plan-templates/data-analysis-plan-logistic-regression/) retrieved from <https://www.statisticssolutions.com/what-is-logistic-regression/>

# [Tony Yiu](https://towardsdatascience.com/@tonester524?source=post_page-----58381e0602d2----------------------), 2019,Understanding Random Forest retrieved from <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>

1. Kirasich, Smith, Sadler, 2019 Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets from

<https://scholar.smu.edu/cgi/viewcontent.cgi?article=1041&context=datasciencereview>

1. Jason Brownlee, August 2018, How to Use ROC Curves and Precision-Recall Curves for Classification in Python

<https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/>