Exploratory and Descriptive Analysis on Loan Dataset

by

Jyoti Singh

Pratiksha Diddi

Honey Soni

Sahil Bhange

**INDEX**

|  |  |
| --- | --- |
| **Sr. No.** | **Topics** |
| 1 | Introduction |
| 2 | Objectives |
| 3 | Data Preparation |
| 4 | Data Visulization |
| 5 | Statistical tools |
| 6 | Classification Models |
| 7 | Performance metrics |
| 8 | Scope and Limitations |
| 9 | Conclusion |

# 

# Introduction

In the past a few years, P2P lending (peer-to-peer lending) has arisen as a major innovation in fintech industry. Without traditional financial intermediaries, through online platforms, people are now able to lend and borrow money much more conveniently and efficiently than before. But how to find profitable and safe loans to invest? The answer is to avoid lending money to someone who is likely to default.

One of the main concerns people have doing investment is its average default rate, which means, the rate at which people fail to pay back the loans. This is a really valid issue considering the investor takes the full brunt of the loss when a borrower defaults. Risks like a platform bankruptcy or new regulatory burdens are important to keep in mind, but nothing has a more direct and negative impact upon our investment’s overall return than the rate at which our borrowers default on their loans.

With this project, we are conducting a set of exploratory analysis and performing various machine learning techniques to predict borrower’s default rate with that we have tried various data visualisation technique to visualise the data how it is distributed.

In general, loan default prediction is a credit scoring problem. Some previous analysis are focused on assigning a score to loan applicants, while many others are focused on predict whether the loan is good or bad. For this project, our objective is to apply binary classification on loans. We defined ”Paid” loans as “1” that are already paid off, and ”Default” loans on which the borrower default as “0”.

We applied Logistic Regression, Decision tree to the problem also using data from Kaggle.com. Their results Logistic Regression are among the best performers when it comes to loan default prediction. We have pre-processed the data in order to gain a strong predictive power. In this report, we employed exploratory analysis techniques to classify loans to two classes: “Default” and “Paid”. The models will be trained and tested on data from data set. We have also interpreted the results from an investor’s point of view and make recommendations on how to utilize the models and discrimination threshold to help with investment decisions.

**Objective**

Now a day’s the prediction of defaulting the borrower in future is a challenging task for credit card companies. Therefore the main objective of this project is to develop prediction models for defaulting the borrower in the future by taking advantage of available technological advancement.

One of the main concerns people have doing investment is its average default rate, which means, **the rate at which people fail to pay back the loans**. This is a really valid issue considering the investor takes the full brunt of the loss when a borrower defaults. Risks like a platform bankruptcy or new regulatory burdens are important to keep in mind, but nothing has a more direct and negative impact upon our investment’s overall return than the rate at which our borrowers default on their loans.

With this project, we are conducting a set of exploratory analysis and performing various machine learning techniques to predict borrower’s default rate with that we have tried various data visualisation technique to visualise the data how it is distributed.

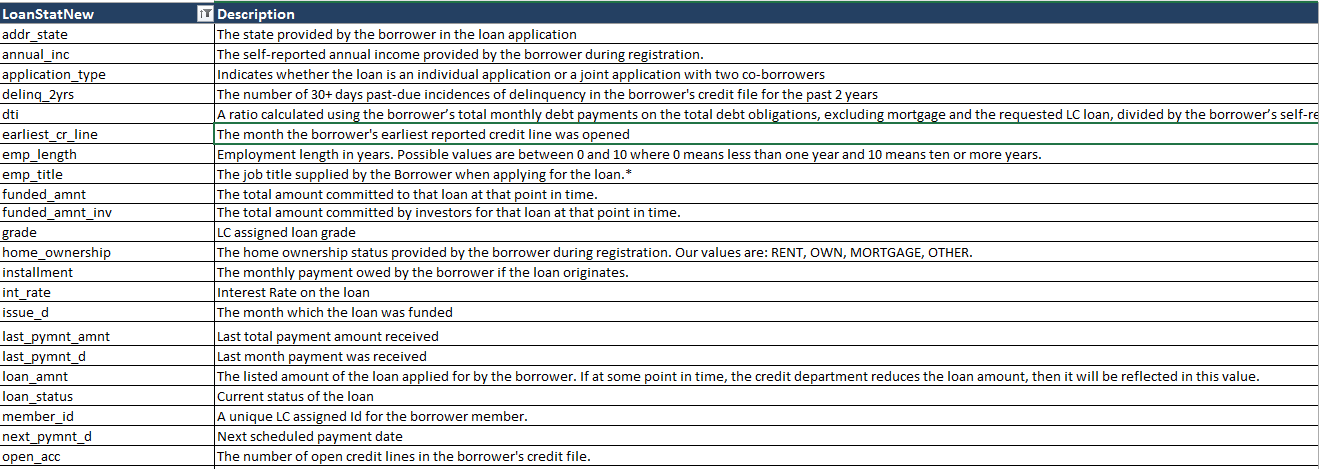
# Data Preparation

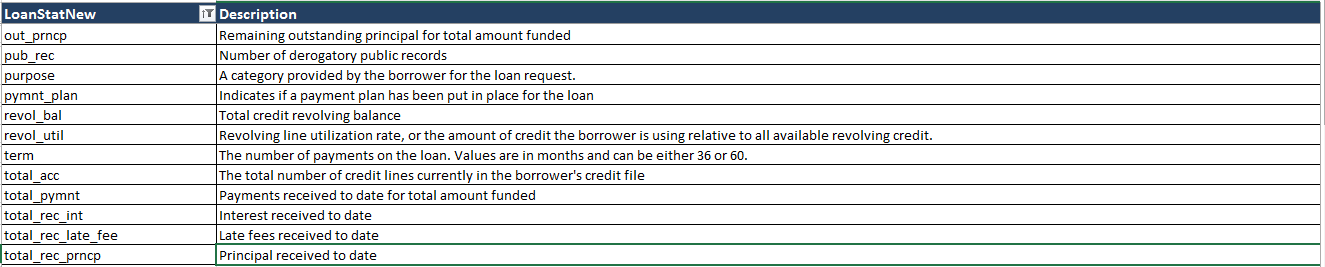
## Obtain the data

We downloaded loan data from <https://www.kaggle.com/c/loan-default-prediction>.

The file contains various parameters such as Monthly Income, Number of Dependents, age, number of open credit lines and loans etc. It is stored in csv format.

**Data description and data set**





* 1. **Terminology:-**

We describe some term related with this project like credit card, delinquency, credit score, utilization etc.

**i) Credit Card:-**

Credit card is a type of financial account. By using credit cards, customers can offers a bank’s money instead of their own to pay for a product or service today, and over time, they repay the bank. For the benefit of using someone else’s money, customers will often need to pay interest, as expected with other types of loans.

**ii) Default:-**

When payments are not made in time and according to the agreement signed by the card holder, the account is said to be in default.

**iii) Delinquency:-**

The term delinquent commonly refers to a situation where a borrower is late or overdue on a credit payment. If borrower is in default, the bank that has issued card is entitled to charge borrower a higher interest rate for that particular billing period as a penalty.

**iv) Credit Score:-**

Credit score is numerical expression based on a level of analysis of a person’scredit files. Simply it is representation of creditworthiness of a borrower.

**v) Utilization:-**

Utilization is a percentage of consumer’s available credit that borrower has used. The credit utilization ratio is key component of borrower credit card.

## Preprocess the data

First we deleted those observations that contain most missing observation. After deleting missing observations, reduced the irrelevant dimensions with business semantics and field co-relation.

The data provided in this project is relatively clean and formatted, however. Based on the preprocessed data, we conducted a series of exploratory data analysis to get a taste of how the data looks like or how it is distributed.

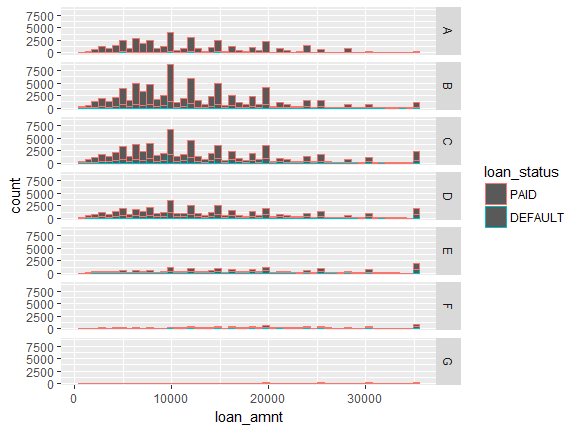
First, we deleted the fields that almost all instances have the same value. For example, for “policy code” field, every instance is 1, and for “pymnt plan”, every instance is “n”. Second, we deleted the instances that have at least one empty field. Third, since there are some fields, such as “is\_inc\_verified”, that earlier loans do not have, we deleted the earlier instances that do not have the up-to-date data fields. There are also some redundant fields which we removed from the data asset.

The original data set has 3 different values for the label field “loan status”: “Charged Off”, “Current”, and “Paid”. Lending Club defines that “Charged Off” occurs when a loan is 150 days past due (i.e. 30 days after the Default status is reached). “Current” means the borrower is paying interest as planned and no default has occurred. “Paid” means the borrower has paid off both interest and principal.

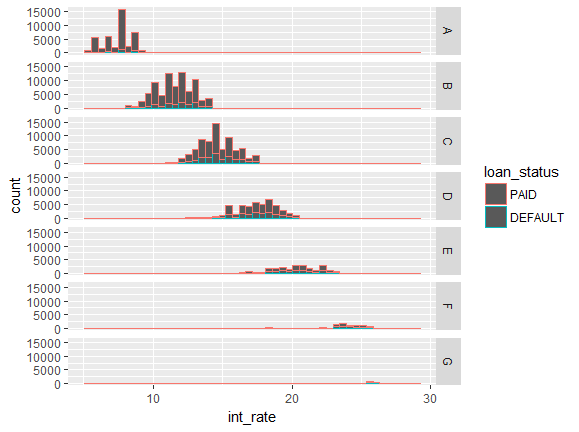
Then we used normalization to bring the data into 2 categories. Finally, I split the data to training data (80%) and testing data (20%).

**2. Data Visualization**

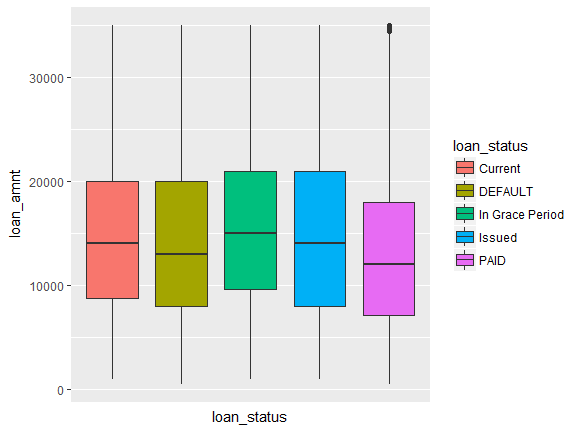
1. **1 Relation between Loan Amount vs. Count**



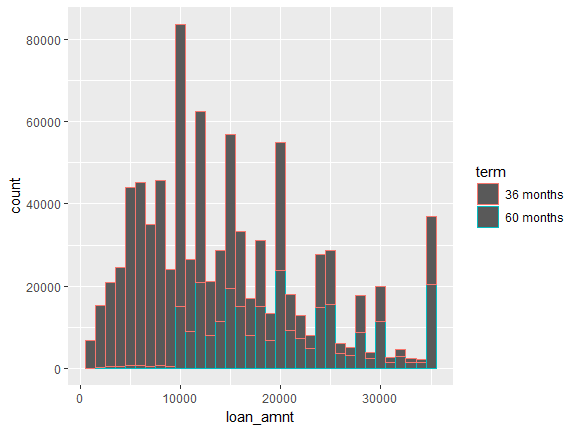
**2.2 Relation between Count vs. Interest Rate**



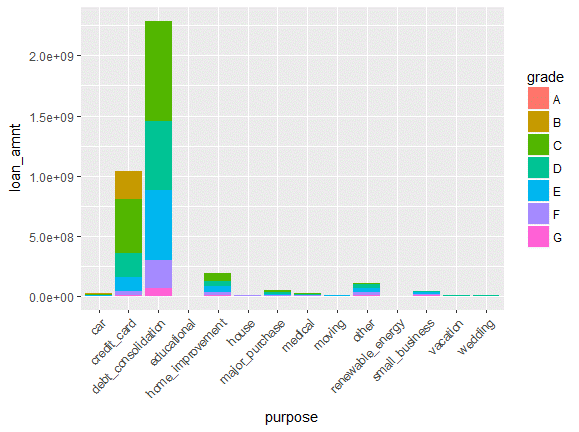
**2.3 BOXPLOT for Loan Status vs. Loan Amount**



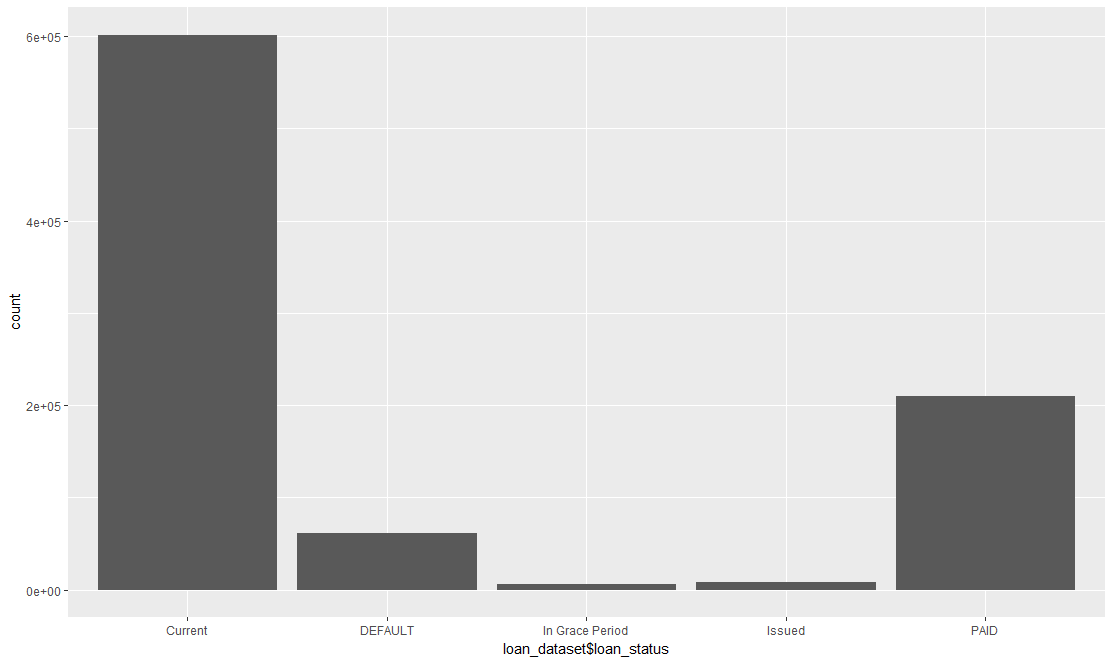
**2.4 Relation between Loan Amount and Count**



**2.5 Relation between Purpose and Loan Amount**



**2.6 Relation Loan dataset vs. Count**

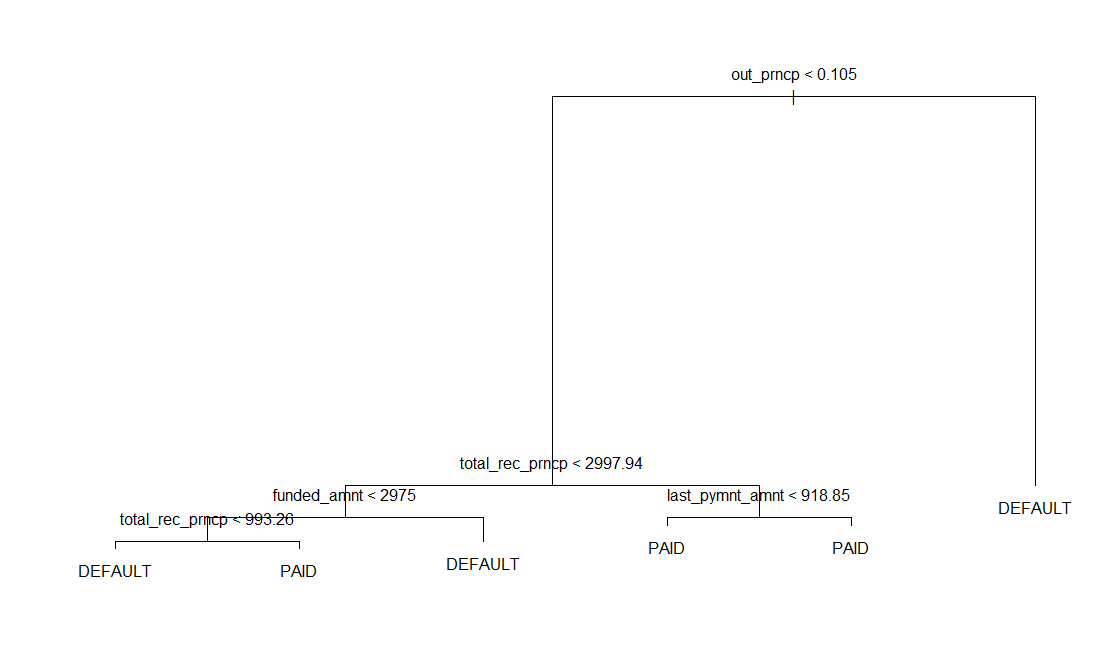


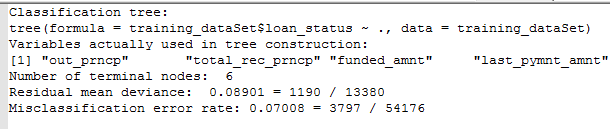
# Classification Models

## 3.1 Decision Tree

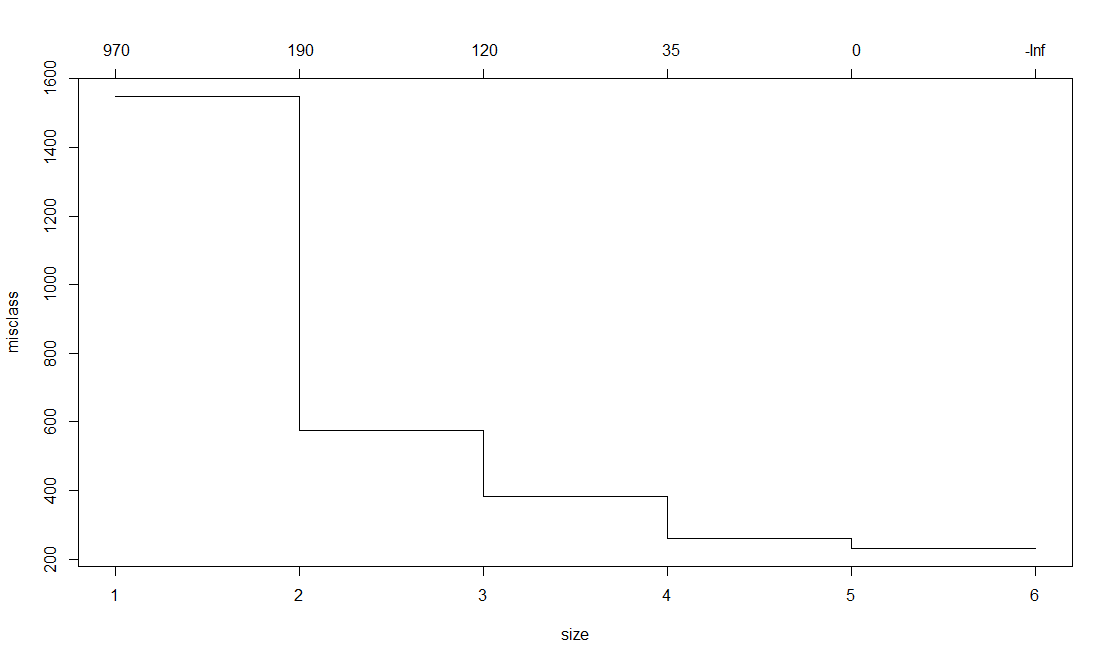
Decision tree learning is a widely used classification model. Decision tree are powerful classification algorithm that are becoming increasing more popular with the growth of data mining in the field of information systems.

As the name implies, this technique recursively separates observations in branches to construct a tree for the purpose of improving the prediction accuracy. In doing so, they use mathematical algorithms like information gain, Gini index to identify a variable and corresponding threshold for the variable that splits the input observation into two or more subgroups. This step is repeated at each leaf node until the complete tree is conducted. The objective of the splitting pair that maximizes the homogeneity of the resulting two or more subgroup samples.



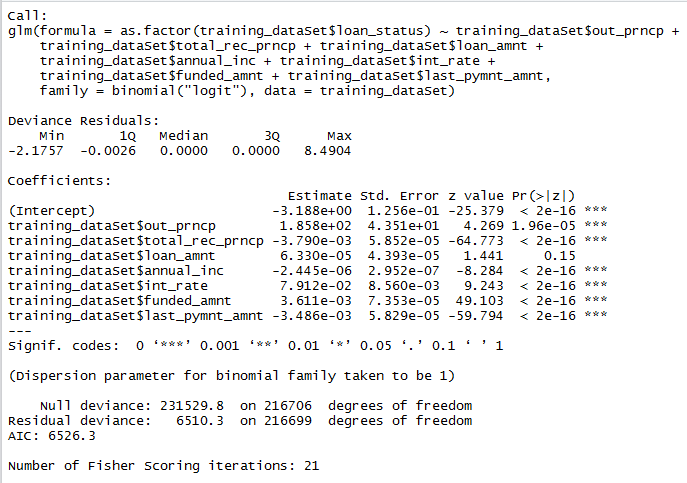


## Optimum Number of Nodes for decision tree obtained through cross validation



## Logistic Regression

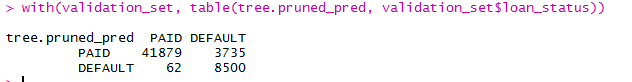
* It is the most commonly used statistical method to develop the prediction models using a large dataset (200000 instances)
* Logistic Regression is used to determine which features of the data set contribute towards someone paying off their loan or defaulting on their loan in United States.
* Modelling with logistic regression helped us to learn more about the data features and get the basic performance of our prediction.
* Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more independent variables.



# Performance Metrics

## 4.1Confusion Matrix

A classifier is typically evaluated by a confusion matrix:



R code for Decision tree and logistic regression model.



**Scope and Limitations**

**Scope:**

As studied, analysis of this project is concerned with credit risk or bank loan defaults. Also our analysis gives better results. Similarly this analysis is useful for other problems related with medical studies like predicting cancer survivability. Also this project is useful for dealing with imbalance dataset.

**Limitation**:

For this project we have data on various variables listed as above. Some variable like occupation, education, living status etc of borrowers are not consider in these study. We develop model only with available variables but if add other important variables then we expect that our models gives better results.

**Conclusion:**

In conclusion, we can show that it is possible to use a superior analytics algorithm through the use of different sampling method to correctly classify customers according to their probability of default. We believe that decision tree give better results when we make data balance by using oversampling method.