**LOAN APPROVAL PREDICTION**

Mohamed Jafir Ashraf M

September 25, 2020

1. **Introduction** 
   1. **Background**

Loan eligibility is primarily dependent on the income and repayment capacity of the individual(s). There are other factors that determine the eligibility of loan such as age, financial position, credit history, credit score, and other financial duty.

* 1. **Problem Description**

Loan Approval Prediction is automate loan eligibility process from customer details. These details are Gender, Marital status, Education, Number of dependents, Self-employed, Monthly income, Loan Application amount and Credit history. In this process identify the customer segmentation and those are eligible for the loan amount.

* 1. **Interesting**

Financial Company or banks needs to automate the loan eligibility process based on customer detail provided while ﬁlling online application form. In this project used to automate loan eligibility process from historical data for customer details.

1. **Data Information**
   1. **Data Attributes**

|  |  |
| --- | --- |
| Loan\_ID | Unique Loan ID |
| Gender | Male/ Female |
| Married | Applicant married (Yes/No) |
| Dependents | Number of dependents |
| Education | Applicant Education (Graduate/ Not Graduate) |
| Self\_Employed | Self employed (Yes/N0) |
| ApplicantIncome | Applicant income |
| CoapplicantIncome | Co applicant income |
| LoanAmount | Loan amount in thousands |
| Loan\_Amount\_Term | Term of loan in months |
| Credit\_History | Credit history meets guidelines |
| Property\_Area | Urban/ Semi Urban/ Rural |
| Loan\_Status | Loan approved (Y/N) |

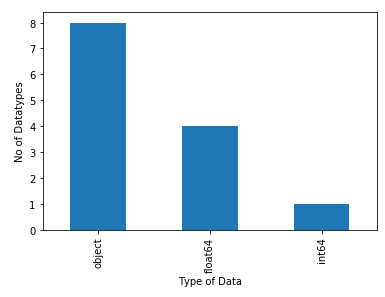
* 1. **Data Sets**

Range Index: 614 entries (0 to 613)

Data columns: 13 columns

* 1. **Data types**

Dataset have three type of data types of the values. These are Object, Float and Integer. The number of categorical datatype in shown in figure 1.1.



**Figure 1.1**

* 1. **Data Missing Values**

Data contains some missing values. The number of missing values in each columns is given below.

Loan\_ID - 0

Gender - 13

Married - 3

Dependents - 15

Education - 0

Self\_Employed - 32

ApplicantIncome - 0

CoapplicantIncome - 0

LoanAmount - 22

Loan\_Amount\_Term - 14

Credit\_History - 50

Property\_Area - 0

Loan\_Status - 0

1. **Data Acquisition/Cleaning** 
   1. **Data Cleaning**

First downloaded the data from source and find the data information, description and shape of the data in after analysis. There were missing values from dataset, because of lack of record keeping. And also lot of data is object type. It is trouble to feature prediction. So we change the datatype from object to numeric values.

Data set has to several problems, so start the cleaning of data. First, there were lot of missing values. There are **Gender, Married, Dependents, Self\_Employed, LoanAmount, Loan\_Amount\_Term** and **Credit\_History**. These missing values replace from zero to mean and some kind of categorical values are contain missing values replace to ‘Undefined or other’ categorical value. For example, we consists of columns are **Gender, Married, Dependents, Self\_Employed.** Theseare categorical values, so that missing value of these columns replace to another categorical values like, undefined or others.

Second, data has lot of data is object data type, so encode the data. Encoding is used to change the values from object to numeric. The columns are **Gender, Married, Dependents, Self\_Employed** and **Property\_Area.** Theseare categorical values, so encode that data. For example, we consists of columns are **Gender**. It has two values are Male and Female. These values are change in male (0) and female (1). It is purpose easy to implement predictive modelling.

After fixing these problems, I checked for outliers in the data. I found there were some extreme outliers, mostly caused by some types of small sample size problem.

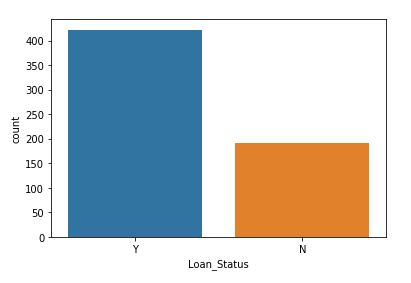
* 1. **Feature Selection**

Final step of the Data acquisition, Feature selection is important to the predictive modelling. After data cleaning, there were 194673 samples and 49 features in the data. Upon examining the meaning of each feature, it was clear that there was some redundancy in the features. But these data set contain all features are important to future predicting.

**4. Exploratory Data Analysis**

**4.1 Calculation of target variable**

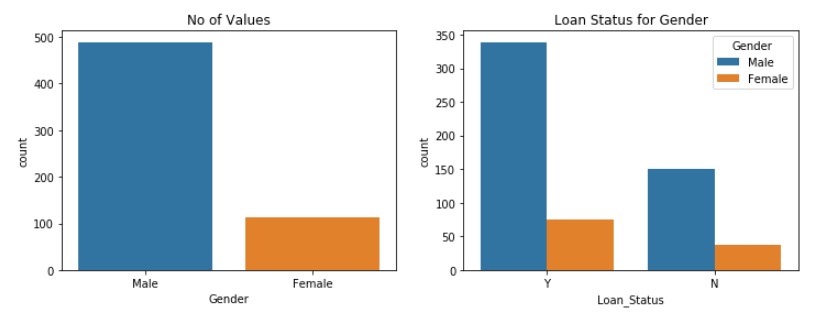
The problem is Loan Approval prediction. So the target value is **Loan Status.** The Loan \_Status column contain loan approved or not. It is categorical (Yes / No) values in shown figure 1.2.

**Figure 1.2**

**4.2 Calculate the features**

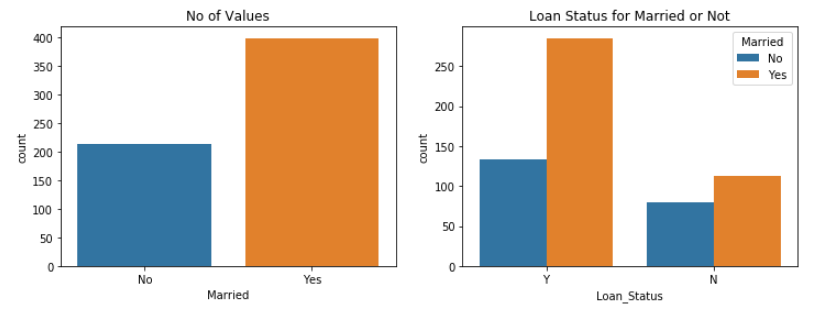
**4.2.1 Gender**

Gender is common feature of identical data. It is the range of characteristics pertaining to, and differentiating between, masculinity and femininity. It has two different types of feature in this columns. There are male and female. The total number of male is 489 and total number of female is 112. And then given plot is define how many male or female approval for loan in shown figure 1.3.



**4.2.2 Married**

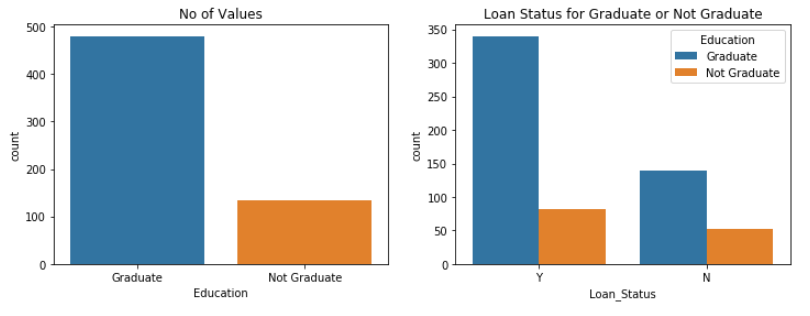
Marital status is one of the important features. It distinct options that describe a person’ relationship with a significant other. Here are some of the important ways a change in marital status can affect the target variable. The value of ‘Married’ are ‘Yes’ and ‘No’. The total number of ‘Yes’ is 398 and total number of ‘No’ is 213. And then given plot is define marital status for loan approval in shown figure 1.4.



**Figure 1.4**

**4.2.3 Education**

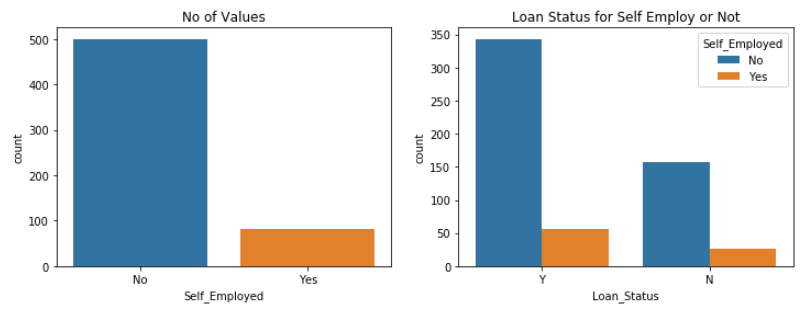
Education is major role on loan approval prediction. Loan eligibility criteria is pursing graduate or post graduate degree. It has to contain two values, there are ‘Graduate’ and ‘Not Graduate’. The total number of ‘Graduate’ is 480 and total number of ‘Not Graduate’ is 134. And then given plot is define Education for loan approval in shown figure 1.5.



**Figure 1.5**

**4.2.4 Self\_Employed**

Self-Employment is another major role on loan approval prediction. Here the maximum applicants are not self-employed for this dataset. The value of ‘Self-Employed’ are ‘Yes’ and ‘No’. The total number of ‘Yes’ is 82 and total number of ‘No’ is 500. And then given plot is define Self-Employed for loan approval in shown figure 1.6.



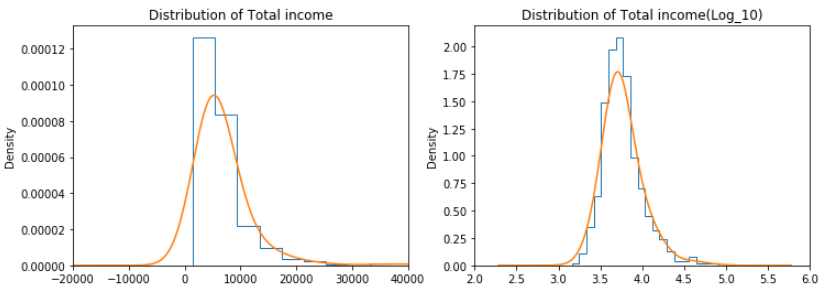
**Figure 1.6**

**4.3 Analysis the feature**

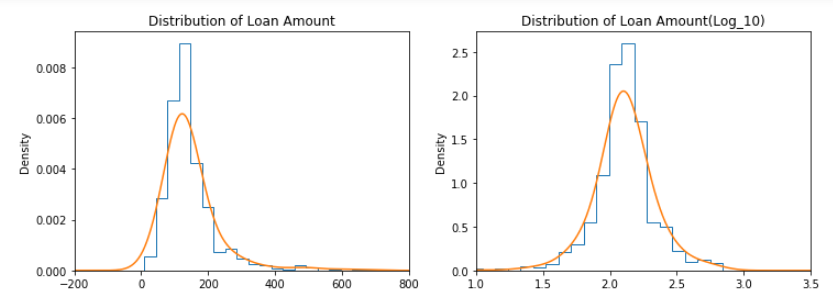
**4.3.1 Distribution of major features**

The distribution is represent for log transformation. The log transformation can be used to make highly skewed distributions less skewed. This can be valuable both for making patterns in the data more interpretable and for helping to meet the assumptions of inferential statistics.

The plots are different between normal distribution and log normal distribution for total income in shown figure 1.7.



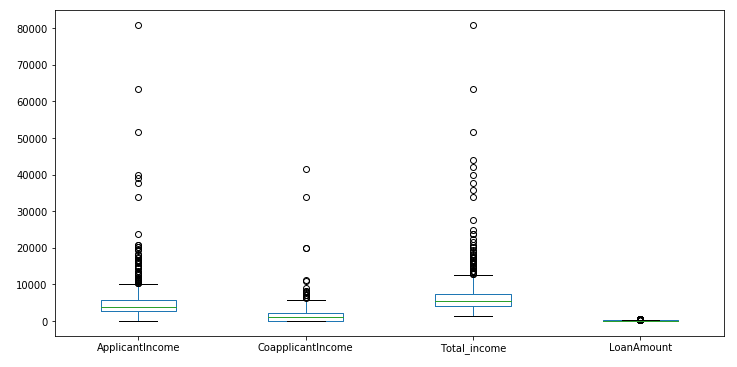
This plot are different between normal distribution and log normal distribution for loan amount in shown figure 1.8.

****

**Figure 1.8**

**4.3.2 Box plot of major features**

The box plot is represent shape of distribution, its central value, and its variability. Its helps to understanding of data in shown figure 1.9.

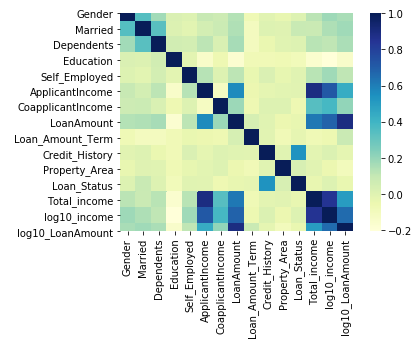


**Figure 1.9**

**4.3.3 Correlation between each data**

Correlation of data is define relationship between each columns in the data. It has to understand how to handle the data and columns. In this plot of diagram is represent with different colours of range (-0.2 to 1.0). The range of colour is negative value, it low rate of correlation between the data. For example, Education and Loan Amount is low correlation of each data. Its correlation value is -0.2.

The range of colour is equal to zero, it normal of correlation between the data. In dataset more amount of columns are correlation between each data should be normal. For example, Gender and Self\_Employed, ApplicantIncome and Loan\_Amount\_Term.

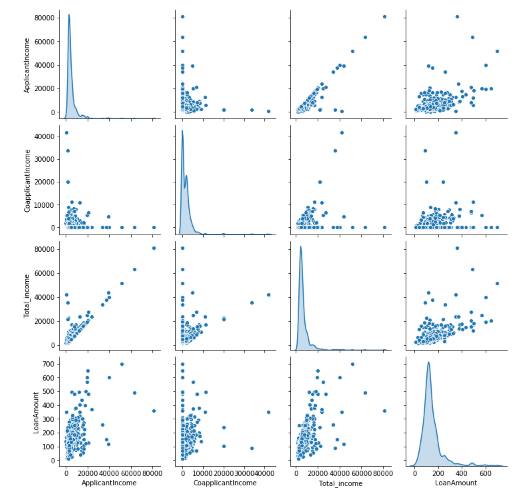


**Figure 1.10**

The range of colour is positive value, it low rate of correlation between the data. Each columns are highly correlated in x-axis and y-axis. So both are same values. And otherwise some columns are highly correlated. For example, Total income and Application Income is high correlation of each data. Its correlation value is 1.0. The correlation of each data is visualize in shown figure 1.10.

**4.3.4 Relationship between major features**

The matrix plot is represent to the relationship between major features. In this plot visualize the data based on scatter plot and histogram. The histogram is present in diagonal. The scatter plot is represent to the relationship between major columns without diagonal. The matrix plot in shown figure 1.11.



**Figure 1.11**

**5. Predictive Modelling**

Predictive modelling uses statistics to predict outcomes. It is the general concept of building a model that capable of making predictions. Typically, such a model includes a machine learning algorithm that learns certain properties from a training dataset in order to make those predictions.

**5.1 Models**

Models can use one or more classiﬁers in trying to determine the probability of a set of data belonging to another set. There are two types of models, Regression and Classification.

Regression is Supervised Learning task where output is having continuous value. The goal here is to predict a value as much closer to actual output value as our model can and then evaluation is done by calculating error value. The smaller the error the greater the accuracy of our regression model.

Classification is a Supervised Learning task where output is having deﬁned labels (discrete value). The goal here is to predict discrete values belonging to a particular class and evaluate on the basis of accuracy. It can be either binary or multi class classiﬁcation. In binary classiﬁcation, model predicts either 0 or 1; yes or no but in case of multi class classiﬁcation, model predicts more than one class.

In this project target value is categorical type (discrete value). So I choose classification model.

**5.2 Applying standard Classification algorithms**

Classiﬁcation in machine learning and statistics is a supervised learning approach in which the computer program learns from the data given to it and make new observations or classiﬁcations. A classiﬁcation model attempts to draw some conclusion from observed values. Given one or more inputs a classiﬁcation model will try to predict the value of one or more outcomes. Outcomes are labels that can be applied to a dataset.

There are a number of classiﬁcation models. Classiﬁcation models include K nearest neighbour and Naive Bayes, Logistic regression, Decision tree, and Random forest.

**5.3 Performance of Models**

Model evaluation metrics are required to quantify model performance. I choose the model evaluation metrics depends on our machine learning task such as classification algorithms. In precision – recall are useful for multiple tasks.

The Classification metrics include classification accuracy, confusion matrix, log loss and f – measure.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Models | F1-score | True Positive | True Negative | False Positive | False Negative |
| KNN | 0.801 | 2 | 81 | 37 | 3 |
| Decision Tree | 0.872 | 17 | 82 | 22 | 2 |
| Logistic Regression | 0.805 | 0 | 83 | 39 | 1 |
| Naive Bayes | 0.870 | 18 | 81 | 21 | 3 |
| Random Forest | 0.858 | 18 | 79 | 21 | 5 |

I applied some classification metrics for model evaluation. There are Classification accuracy and Confusion matrix. Classification accuracy is the number of correct prediction made as a ratio of all predictions made. Confusion matrix provide a more detailed breakdown of correct and incorrect classification for each class. And also fine Actual and predicted values (True and False).

It estimated performance of a model tells as how well it preform on unseen data. And also I find best classifier on this problem based on the table, it is **decision tree**. It has to high accuracy and better result on confusion matrix. In table explain performance of different models (Table 1.1).

**6. Conclusion:**

Finally, I predicted the severity of the collision based on further analysis. I achieved above 80% accuracy in classification algorithms. That is helps to identify the eligibility of loan. And also analysis of major features, it used to get better result of this problem.

Purpose of this project was to predict the **Loan Approval**. Company wants to automate the loan eligibility process based on customer details. These details are Gender, Marital status, Education, Number of dependents, Self-employed, Monthly income, Loan Application amount and Credit history. In this process identify the customer segmentation and those are eligible for the loan amount.