

# **Customer Loan Prediction Analysis**

Submitted in partial fulfillment of the requirements for  
the award of  
Bachelor of Engineering degree in Computer Science and Engineering

by

Gopinath Mahankali (Reg. No. 37110419)

Maheep Kunagu (Reg. No. 37110392)



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING  
SCHOOL OF COMPUTING**

## **SATHYABAMA**

**INSTITUTE OF SCIENCE AND TECHNOLOGY  
(DEEMED TO BE UNIVERSITY)**

**Accredited with Grade "A" by NAAC  
JEPPIAAR NAGAR, RAJIV GANDHI SALAI,  
CHENNAI – 600 119**

**MARCH - 2021**



# **SATHYABAMA**

INSTITUTE OF SCIENCE AND TECHNOLOGY  
(DEEMED TO BE UNIVERSITY)

Accredited with "A" grade by NAAC  
Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai – 600 119

[www.sathyabama.ac.in](http://www.sathyabama.ac.in)



## **DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

### **BONAFIDE CERTIFICATE**

This is to certify that this project report is the bonafide work of **GOPINATH MAHANKALI (Reg. No. 37110419)** and **KUNAGU SRINIVAS SHANKAR MAHEEP (Reg. No.37110392)** who carried out the project entitled "**CUSTOMER LOAN PREDICTION ANALYSIS**" under my supervision from **August 2020** to **March 2021**.

#### **Internal Guide**

Mr. R. Sethuraman B.E., M.E., (Ph.D.,)

#### **Head of the Department**

Dr. S. VIGNESHWARI, M.E, Ph.D.

---

Submitted for Viva voce Examination held on \_\_\_\_\_

Internal Examiner

External Examiner

## **DECLARATION**

I **GOPINATH MAHANKALI** hereby declare that the Project Report entitled “**CUSTOMER LOAN PREDICTION ANALYSIS**” is done by me under the guidance of **Mrs. R. Sethuraman, B.E., M.E., (Ph.D.,)** Department of Computer Science and Engineering at Sathyabama Institute of Science and Technology is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in Computer Science and Engineering.

**DATE:**

**PLACE: CHENNAI**

**SIGNATURE OF THE CANDIDATE**

## ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala, M.E., Ph.D., Dean**, School of Computing, **Dr. S. Vigneswari, M.E., Ph.D.,** and **Dr. L. Lakshmanan, M.E., Ph.D., Heads of the Department** of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Mrs. R. Sethuraman, B.E., M.E., (Ph.D.) Assistant Professor**, for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

## **ABSTRACT**

In our banking system, banks have many products to sell but main source of income of any banks is on its credit line. So they can earn from interest of those loans which they credits. Loan approval is a very important process for banking organizations. Banking Industry always needs a more accurate predictive modeling system for many issues. Predicting credit defaulters is a difficult task for the banking industry.

A bank's profit or a loss depends to a large extent on loans i.e. whether the customers are paying back the loan or defaulting. By predicting the loan defaulters, the bank can reduce its Non- Performing Assets. This makes the study of this phenomenon very important. Previous research in this era has shown that there are so many methods to study the problem of controlling loan default. But as the right predictions are very important for the maximization of profits, it is essential to study the nature of the different methods and their comparison. A very important approach in predictive analytics is used to study the problem of predicting loan defaulters: The Logistic regression model.

Logistic Regression models have been performed and the different measures of performances are computed. The models are compared on the basis of the performance measures such as sensitivity and specificity. The final results have shown that the model produce different results. Model is marginally better because it includes variables (personal attributes of customer like age, purpose, credit history, credit amount, credit duration, etc.) other than checking account information (which shows wealth of a customer) that should be taken into account to calculate the probability of default on loan correctly. Therefore, by using a logistic regression approach, the right customers to be targeted for granting loan can be easily detected by evaluating their likelihood of default on loan. The model concludes that a bank should not only target the rich customers for granting loan but it should assess the other attributes of a customer as well which play a very important part in credit granting decisions and predicting the loan defaulters.

## TABLE OF CONTENTS

### ABSTRACT

v

### LIST OF FIGURES

viii

| CHAPTER No. | TITLE                                  | PAGE No.  |
|-------------|----------------------------------------|-----------|
| <b>1.</b>   | <b>INTRODUCTION</b>                    | <b>1</b>  |
|             | 1.1 INTRODUCTION TO LOANS              | 1         |
|             | 1.2 LOAN APPROVAL AUTOMATION PROCESS   | 2         |
|             | 1.3 RESEARCH AND SIGNIFICANCE          | 3         |
| <b>2.</b>   | <b>AIM AND SCOPE</b>                   | <b>5</b>  |
|             | 2.1 AIM OF PROJECT                     | 5         |
|             | 2.2 OBJECTIVES                         | 5         |
|             | 2.3 SCOPE                              | 6         |
|             | 2.2.1 ADVANTAGES                       | 7         |
| <b>3.</b>   | <b>SYSTEM DESIGN &amp; METHODOLOGY</b> | <b>8</b>  |
|             | 3.1 EXISTING SYSTEM                    | 8         |
|             | 3.2 PROPOSED SYSTEM                    | 8         |
|             | 3.3 SYSTEM CONFIGURATION               | 9         |
|             | 3.4 MODULES                            | 9         |
| <b>4.</b>   | <b>MODEL DEVELOPMENT METHODOLOGY</b>   | <b>13</b> |
|             | 4.1 DESCRIPTION OF DIAGRAM             | 13        |
|             | 4.2 ARCHITECTURAL DESIGN               | 15        |
|             | 4.3 IDENTIFYING THE ACTORS             | 15        |
|             | 4.4 SYSTEM IMPLEMENTATION              | 15        |
|             | 4.5 MODULAR DESIGN                     | 18        |

|           |                                   |           |
|-----------|-----------------------------------|-----------|
| <b>6.</b> | <b>RESULTS AND DISCUSSION</b>     | <b>20</b> |
| <b>7.</b> | <b>CONCLUSION AND FUTURE WORK</b> | <b>24</b> |
|           | <b>REFERENCES</b>                 | <b>26</b> |
|           | <b>APPENDIX</b>                   |           |
|           | A. PLAGIARISM REPORT              | 27        |
|           | B. JOURNAL PAPER                  | 28        |
|           | C. SOURCE CODE                    | 34        |

## LIST OF FIGURES

| FIGURE No. | FIGURE NAME               | PAGE No. |
|------------|---------------------------|----------|
| 4.1        | BLOCK DIAGRAM             | 13       |
| 4.2        | ARCHITECTURE DIAGRAM      | 15       |
| 4.3        | LOGISTIC REGRESSION GRAPH | 18       |
| 5.1        | BOX PLOT 1                | 20       |
| 5.2        | BOX PLOT 2                | 21       |
| 5.3        | BOX PLOT 3                | 21       |
| 5.4        | BOX PLOT 4                | 22       |
| 5.5        | MODEL BUILDING            | 22       |
| 5.6        | ACCURACY RATE RESULT      | 23       |
| 5.7        | OUTPUT FILE               | 23       |

# **CHAPTER 1**

## **INTRODUCTION**

### **1.1 INTRODUCTION TO LOANS**

In most developing countries, the banking sector continues to play a leading role in promoting economic growth, and this in spite of efforts to boost their financial markets, which remained underdeveloped. Credit risk is among the most significant risks to which banks are exposed. This allowed the proliferation of risk management methods.

Distribution of the loans is the main business part of almost every bank. The main portion of the bank's asset is directly from the profit earned from the loans distributed by the banks. When credit is granted, and regardless of the expected gain, it can be exposed to uncertainty of default. The credit institution is not always safe to recover its funds and is exposed to counter party risk.

Banking domain is to invest their assets in safe hands. Lending money to unsuitable loan applicants results in the credit risk. Today many banks approve loans after a long procedure of verification, yet there is no guarantee whether the picked candidate is the right candidate or not. Estimating the risk, which is involved in a loan application, is one of the most significant concerns of the banks in order to survive in the highly competitive market. Through our proposed model we can predict whether that specific customer is safe or not.

Today many banks/financial companies approve loan after a regress process of verification and validation but still there is no surety whether the chosen applicant is the deserving right applicant out of all applicants. The objective of this paper is to present models for predicting the probability of default of the counterparty based on the score method, which pits discriminant analysis against logistic regression. Credit scoring is a method that helps the bank to rationalize its process for credit granting decision. Its principle is to synthesize a set of financial ratios as one indicator able to distinguish between good and bad customers.

Loan Prediction is very helpful for employee of banks as well as for the applicant also. The aim of this Paper is to provide quick, immediate and easy way

to choose the deserving applicants.

This Process is exclusively for the managing authority of Bank/finance company, whole process of prediction is done privately no stakeholders would be able to alter the processing. Result against particular Loan Id can be send to various department of banks so that they can take appropriate action on application..

For banks, the big problem has become not only to decide whether to grant a loan or not but also to predict the probability of default of a borrower in case of a credit agreement. It is to anticipate, for a determinate period, the quality of the borrower (good or bad borrower) and its ability to repay its debt. When credit is granted, and regardless of the expected gain, it can be exposed to uncertainty of default. The credit institution is not always safe to recover its funds and is exposed to counter party risk.

## **1.2 LOAN APPROVAL AUTOMATION PROCESS**

Credit scoring is a widely used technique that helps financial institutions evaluates the likelihood for a credit applicant to default on the financial obligation and decide whether to grant credit or not. the lowest credit score referred to as 'zero' and the highest credit score referred to as 'one'. Here, t he highest correlated factor on the loan status of a customer is Credit score. This fact is inferred by the feature extraction and correlation matrix of a customer's data.

Credit scoring is a statistical method for estimating the probability of default of the borrower using historical data and statistical data to reach a single indicator that can distinguish good borrowers from bad borrowers. The score function is based on a method of financial analysis on financial ratios presented as a single indicator can distinguish between healthy and failing companies.We can find it by implementing Logistic Regression Model.

The increased number of potential applicants impelled the development of sophisticated techniques that automate the credit approval procedure and supervise the financial health of the borrower. The large volume of loan portfolios also imply that modest improvements in scoring accuracy may result in significant savings for financial institutions (West, 2000). The goal of a credit scoring model is to classify credit applicants into two classes: the "good credit" class that is liable to reimburse

the financial obligation and the “bad credit” class that should be denied credit due to the high probability of defaulting on the financial obligation. The classification is contingent on sociodemographic characteristics of the borrower (such as age, education level, occupation and income).

In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

### **1.3 RESEARCH AND SIGNIFICANCE**

The primary aim of this proposed system i.e. “Customer Loan Prediction Analysis” is to find whether the Customer applied for Loan is eligible to take by evaluating all his details related to his credit score, Gender, Annual Income and Nature of Place living.

This paper proposes a credit scoring model of consumer loans based on various analytical models. The rest of this paper is organized as follows. In the next section, exploratory analysis and data transformation is presented. This is followed by a description of the data sets and a comparison of the predictive accuracy of the models.

In our proposed model we had used Logistic regression which is one of the popular Machine Learning algorithms that comes under the Supervised Learning technique. It is applicable for categorical dependent variable using a given set of independent variables. Thus, the outcome must be a categorical or discrete value. The output can be either Yes or No, 0 or 1, true or false, etc. but instead of giving the exact value as 0 or 1, it gives some probabilistic values which lies between 0 and 1. Logistic regression is much similar to linear regression except that how they are used. It is used for solving regression problems, whereas Logistic regression is used for solving the classification problem.

In Logistic regression, rather than fitting a regression line, we fit an "S" shaped logistic function, which predicts two greatest values (0 or 1). The curve from the logistic function demonstrates the probability of something, for example, regardless of whether the cells are destructive or not, a mouse is corpulent or not founded on its weight, and so on. It is a significant algorithm because it can provide

probabilities and classify the use of different types of data and easily determines the most effective variables that are used for classification. Regression is the task of forecasting the value of a continuously changeable variable (e.g. a price, a temperature) given some input variables. As well as the lowest credit score tends to offer the highest interest rates on the loan. Actually, the credit score saves the lenders to trust the unreliable borrowers. The Credit Score plays a major role in the customer's information. when the FICO Score was created according to the money lending process was fairly subjective, and potential borrowers were often judged by how trustworthy their character seemed.

The past machine learning models have the highest accuracy but with the least precision score. The precision score is the number of true positives per the total number of true positives and false positives. If the precision score is low means the large number of False Positives generated by the model. Generating higher number of False Positives than the False Negatives decreases the model exactness and also increases the risk factor.

## **CHAPTER 2**

### **AIM AND SCOPE**

#### **2.1 AIM OF PROJECT**

The primary aim of this proposed system i.e. "Customer Loan Prediction Analysis as we know that now-a-days there is a rapid growth in banking sector, resulting lots of people are applying for bank loans. Finding out the applicant to whom the loan will be approved is a difficult process. In this paper, we proposed a model which predicts loan approval/rejection of an applicant using machine learning techniques. This can be done by training the model with the data of the previous records of the people applied for loan.

For this, First we will analyze the data from the Dataset download from Kaggle and then Preprocess the data using Exploratory Data Analysis and after that build the model using Train Dataset and use that model for Test Data to Predict the output. To do this, we need Logistic Regression for finding the Loan Approval status of the applicants.

#### **2.2 OBJECTIVES**

The proposed model "Online Crime Reporting System (OCRS)" is based on the new technology Logistic Regression where rather than fitting a regression line, we fit an "S" shaped logistic function, which predicts two greatest values (0 or 1). It is a significant algorithm because it can provide probabilities and classify the use of different types of data and easily determines the most effective variables that are used for classification.

After that we will Clean and process the Dataset and understanding the Data fields and coming to and understanding by finding the criteria's and variables giving positive impact for Loan approval. Various steps are involved to do this process, after doing all the necessary steps like building model and training it to the Test data set.

Then after implementation we will get an output.csv Dataset file containing all the predicted approval status based on the model we build which can be used by

Banks to automate the Loan approval Process in a easier and faster way in no time.

## 2.3 SCOPE

This model focusses more on the both accuracy and cross validation score of a model. The higher Precision Score decreases the lender's risk factor. By implementing this Banks can save more time in processing the Customers or Applicants Loan application by automating the process so that it saves more time and human work intervention.

Also to improve the accuracy of the estimation of logistic regression coefficients when we have a binary response variable and a large number of highly correlated categorical explanatory variables. For this purpose, we first employ a categorical principal component analysis to deal with multi collinearity problem among binary explanatory variables, and then use uncorrelated principal components instead of original correlated.

To implement this, the model fed by the data which impacts more on the loan status and also to decrease the more ambiguity the credit score factor consists only two unique values  $\{0,1\}$  where '0' represents lower credit score and '1' represents higher credit score. By doing this we can predict the Approval status for the customers who applied for the loan in the respective Banks.

The main goal of our model is to allocate credit borrowers to two groups: "good credit" group that is likely to repay the financial obligation or "bad credit" group with a high risk of default on its financial obligations. Our sample consists of 195 SMEs using credit files distributed over five sectors of activity namely "agriculture, industry, services, tourism and trade."

The objective of this paper is to present models for predicting the probability of default of the counterparty based on the score method, which pits discriminant analysis against logistic regression. Credit scoring is a method that helps the bank to rationalize its process for credit granting decision. Its principle is to synthesize a set of financial ratios as one indicator able to distinguish between good and bad customers.

The past machine learning models have the highest accuracy but with the least precision score. The precision score is the number of true positives per the

total number of true positives and false positives. If the precision score is low means the large number of False Positives generated by the model. Generating higher number of False Positives than the False Negatives decreases the model exactness and also increases the risk factor. The credit risk analysis is based on various information about the borrower, which may be summarized as qualitative and quantitative data such as financial ratios, since access on qualitative data seems be difficult, our study will be only based on quantitative variables; hence it becomes necessary to make a wise choice of our financial ratio.

### **2.3.1 ADVANTAGES**

- Time saving as it is operated online.
- Automate the Loan approval process.
- This system lets the model to predict accurately.
- This model decreases the risk and efforts of a lender.
- This process is based on digitalization.
- Useful to minimize this risk without much effort on the lender.
- Reduces the Human Involvement in this process for banks
- Having more Accuracy Rate when compared to other models.

## **CHAPTER 3**

### **SYSTEM DESIGN & METHODOLOGY**

#### **3.1 EXISTING SYSTEM**

Machine Learning implementation is a very complex part in terms of Data analytics. Working on the data which deals with prediction and making the code to predict the future of outcomes from the customer is a challenging part. All the other methods are simply making the process way more complexed and not giving accurate Cross Validation score in predicting results.

##### **Disadvantages of Existing System:**

- Complexity in analyzing the data.
- Prediction is a challenging task working in the model.
- Coding is complex maintaining multiple methods.
- Libraries support was not that much familiar.

#### **3.2 PROPOSED SYSTEM**

Python has a good area for data analytical which helps us in analyzing the data with better models in data science. The libraries in python makes the prediction for loan data and results with multiple terms considering all properties of the customer in terms of predicting. Logistic Regression is used to build the model and used to get the output by predicting results accurately.

The objective of this paper is to present models for predicting the probability of default of the counterparty based on the score method, which pits discriminant analysis against logistic regression. Credit scoring is a method that helps the bank to rationalize its process for credit granting decision. Its principle is to synthesize a set of financial ratios as one indicator able to distinguish between good and bad customers.

##### **Advantages of Proposed System:**

- Libraries help to analyze the data.
- Statistical and prediction is very easy comparing to existing technologies.
- Results will be accurate compared to other methodologies.

### **3.3 SYSTEM CONFIGURATION**

#### **H/W System Configuration**

Processor : Pentium 3, Pentium 4 and higher RAM : 2GB/4GB RAM.  
Hard disk : 40GB and higher.

#### **S/W System Configuration**

Operating System : Windows 7 , Windows 8, (or higher versions)  
Language : Python 3.5 Mozilla Firefox(or any browser)  
IDLE : Jupyter Notebook.

### **3.4 MODULES**

The Customer Loan Prediction Analysis Model consists of mainly Modules. They are Reading and Cleaning Dataset, Model Building and finally Testing the Dataset. Here the primary task is to maintain the database required for analyzing the data in the form of .csv files. It needs a Train dataset, Test Dataset. The user will have access to both the datasets to train and build the testing model.

Firstly, we need to implement Exploratory Data Analysis for Spotting mistakes in the provided or collected Dataset and mapping them to underlying structure of the data. Finding the Most important variables and fields in the dataset given by EDA by visualizing graphs and boxplots using python libraries NumPy and Pandas. Listing Anomalies and Outliers comes to the next stage scenario. Let us look into each module in detail.

#### **1. Reading and Cleaning Dataset**

Now we need to import Pandas, NumPy and sklearn libraries and use them to process the data. Reading both training and testing dataset in a data frame using Pandas. Using head () function, we will be able to see first 10 rows of the dataset so that we will have a clear picture on what fields does dataset contains in it. After then we will store the length of rows and columns in the dataset. Here comes the crucial part of this module nothing but processing the data as required for the analysis. For that we need to understand the various features and columns of the dataset. We will use describe () function to get the summary of all the numerical and categorical value fields in the dataset which contains Applicant income, Co-applicant income, Credit history and Loan amount.

For the non-numerical values (e.g., Property Area, Credit History etc.), we can look at frequency distribution to understand whether they make sense or not. Getting the unique values and their frequency of variable Property Area and also Understanding Distribution of Numerical Variables like Applicant Income and Loan Amount can be done by using Boxplot to understand and finding the outliers of the Dataset fields. We can see that there is no substantial different between the mean income of graduate and non-graduates. But there are a higher number of graduates with very high incomes, which are appearing to be the outliers. So after this we will be able to find fields which are having positive impact on Loan approval status and coming to a conclusion by analysis.

We can see that there is no substantial different between the mean income of graduate and non-graduates. But there are a higher number of graduates with very high incomes, which are appearing to be the outliers. Loan Amount has missing as well as extreme values, while Applicant Income has a few extreme values. Now it is time to Understanding Distribution of Categorical Variables.

By making Cross table with both Credit History and Loan Amount Fields we will see that Loans Approved in the Train Data set are more with applicants having credit history equals to 1. Then we will write a function to find out the percentage of applicants whose loans are approved with credit history equals to 1 and it shows more than 79% people have got loans with Credit history of 1. Now we will move forward and understand outliers in a better way in next module to build the model.

## **Module 2 – Model Building**

A value that deviated significantly from the rest of the values in a feature is referred to as an Outlier. They can be caused either by measurement or execution error. Analyzing the outlier data from the data point is referred to as outlier analysis. Removing or decreasing the outliers is important when that particular feature effecting more on the dependent variable means if the feature is important in getting the output.

In this project, we approached Box Plot methodology on the numerical features. Exploratory Data Analysis is graphically representing how the features are in the dataset to discover the patterns, anomalies and to derive the insights and assumptions used to get full knowledge on the given data to build a model. EDA is all about making sense of data in hand, before getting them dirty with it. What we did

earlier in this project also comes under EDA.

In this phase, we were discussing a bit more depth than before for better clarification on the insights we got. The insights we got from the feature analysis were the highest point values were male, married, graduation, not self-employed, and semi-urban areas. But there, we applied the statistics one by one and we were not sure about how these features affecting the target variable Loan Status when we grouped those together.

Also, we were not sure that we should clear the outliers in numerical features or not. sklearn requires all inputs to be numeric, we should convert all our categorical variables into numeric by encoding the categories. Before that we will fill all the missing values in the dataset. In order to prepare and apply a model to this dataset, we'll first have to break it into two subsets.

The first will be the training set on which we will develop the model. The second will be the test dataset which we will use to test the accuracy of our model. We will allocate 75% of the items to Training and 25% items to the Test set. Once our dataset has been split, we can establish a baseline model for predicting whether a credit application will be approved.

This baseline model will be used as a benchmark to determine how effective the models are. First, we determine the percentage of credit card applications that were approved in the training set. We can see from the summary output that the Debt variable has missing values that we'll have to fill in. We could simply use the mean of all the existing values to do so.

Another method would be to check the relationship among the numeric values and use a linear regression to fill them in. Regression models are useful for predicting continuous (numeric) variables. However, the target value in Approved is binary and can only be values of 1 or 0. The applicant can either be issued a credit card or denied- they cannot receive a partial credit card.

We could use linear regression to predict the approval decision using threshold and anything below assigned to 0 and anything above is assigned to 1. Unfortunately, the predicted values could be well outside of the 0 to 1 expected range. Therefore, linear or multivariate regression will not be effective for predicting the values. Instead, logistic regression will be more useful because it will produce probability that the target value is 1. Probabilities are always between 0 and 1 so the output will more closely match the target value range than linear regression.

Convert all non-numeric values to number and creating Generic Classification Function for accessing performance.

Logistic Regression is a classification algorithm for supervised data and also similar to linear regression, the difference between them is logistic regression uses Logistic function/ sigmoid function which is an S-shaped curve[5]. The sigmoid function can take any real-valued number/integer and maps that value into a value between 0 and 1, but won't be 0 or 1.

The chances of getting loan will be higher for Applicants having credit history equals to 1, Applicants with higher applicant income and Co-applicant income, with Higher education and having properties in Urban areas. So we make our model with 'Credit History', 'Education' & 'Gender'.

### **Module 3 – Testing Data set with Model**

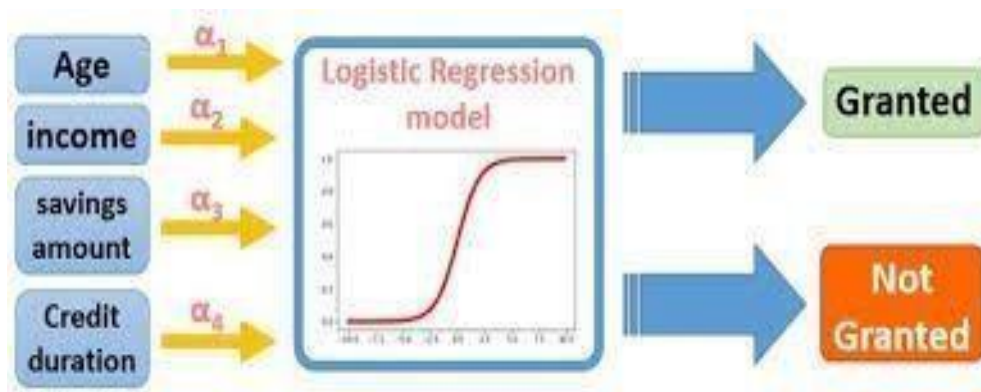
Model building will be completed by combining both training and test data set and converting into log transformation to nullify outliers and model is built. Now let's use it to test with the test data set to predict the output for the applicants. Both Test and train dataset are combined and do reverse encoding to get outcome. Final Output will be stored in the local storage with Output.csv and also we can see Accuracy Rate and cross validation score seems to be higher than any other models with more than 80%. Results can be viewed from the Output.csv file which can be used by banks for Loan approvals and Processing them to the applicants.

So finally, we have builded a model and used it to predict the output for this Loan Approval Process.

## CHAPTER 4

### MODEL DEVELOPMENT METHODOLOGY

#### 4.1 Description of Diagram



**Fig 4.1 Block Diagram**

An effective model was proposed for predicting the right customers who have In our proposed model we had used Logistic regression which is one of the popular Machine Learning algorithms that comes under the Supervised Learning technique. It is applicable for categorical dependent variable using a given set of independent variables. Thus, the outcome must be a categorical or discrete value. The output can be either Yes or No, 0 or 1, true or false, etc. but instead of giving the exact value as 0 or 1, it gives some probabilistic values which lies between 0 and 1. Logistic regression is much similar to linear regression except that how they are used. It is used for solving regression problems, whereas Logistic regression is used for solving the classification problems. applied for loan. The input dataset is the bank dataset of customers who applied for the loan. The dataset is a CSV file. The dataset can read into the python environment by using the `read_csv()` method in pandas. For that, should import pandas into the present python environment. The features of the

Customers dataset: Loan\_ID, Gender, Married, Dependents, Education, Self Employed, ApplicantIncome, CoapplicantIncome, LoanAmount, LoanAmount\_Term, Credit\_History, Property Area and Loan\_Status. Missing Values considers as noisy data in the dataset. If no information is provided while collecting the data that particular entries in the database will refer to as Missing values. The missing data is either represented as Nan or None in pandas. None is referred to as missing data in code and also as singleton object of python. Mean imputation replacing the mean value of a column in place of null values. Median imputation replacing the median value of a column in place of null values. A value that deviated significantly from the rest of the values in a feature is referred to as an Outlier. They can be caused either by measurement or execution error. Analyzing the outlier data from the data point is referred to as outlier analysis. Removing or decreasing the outliers is important when that particular feature effecting more on the dependent variable means if the feature is important in getting the output. Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with structured (tabular, multidimensional, potentially heterogeneous) and time series data both easy and intuitive.

### **Steps in Logistic Regression:**

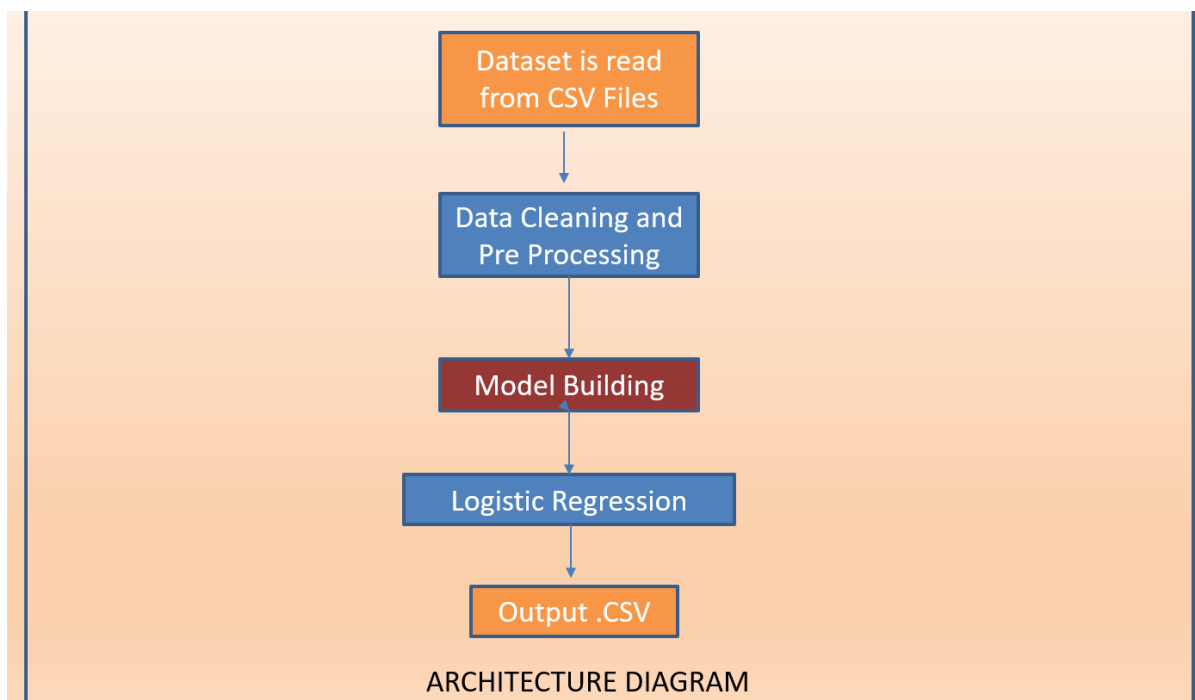
For implementing the Logistic Regression using Python, we had done the following steps:

1. Data Pre-processing.
2. Fitting Logistic Regression to the Training set.
3. Predicting the test result.
4. Test accuracy of the result.
5. Visualizing the test set result.

Based on the data given by the loan applicant, we can predict whether the loan of particular applicant is approved or not using a User Interface. User interface contains input variables with their corresponding fields and a field to display the output. Input variables are Gender, Marital status, Dependents, Education, Applicant income, Loan Amount, Loan amount term, Credit History, Property Area. The applicant need to give these values and based on these, the model will predict whether the loan will be approved or not.

As per our Analysis, Credit History, Income, Education, Gender and Property Areas of the Applicant makes more impact so we created a model using above steps with these fields.

## 4.2 ARCHITECTURE DIAGRAM



**Fig 4.2 ARCHITECTURE DIAGRAM**

## 4.3 IDENTIFYING THE ACTORS

The actors in the system are Loan Approval Authority , Applicant and Builed System consisting of maintain and export data and also preprocessing the data by building the model. The main goal of the Loan Approval Authority is to find out the predictable status of Granting loan or not based on the details given by the applicant by using the system.

## 4.4 SYSTEM IMPLEMENTATION

We will get input dataset is the bank dataset of customers who applied for the loan. The dataset is a CSV file. The dataset can read into the python environment by using the `read_csv()` method in pandas. For that, should import pandas into the present python environment. Lets us look into the variable description of each field in the Dataset.

Based on the data given by the loan applicant, we can predict whether the loan of particular applicant is approved or not using a User Interface. User interface contains input variables with their corresponding fields and a field to display the output. Input variables are Gender, Marital status, Dependents, Education, Applicant income, Loan Amount, Loan amount term, Credit History, Property Area. The applicant need to give these values and based on these, the model will predict whether the loan will be approved or not.

| Variable          | Description                     |
|-------------------|---------------------------------|
| Loan_ID           | Unique Loan ID                  |
| Gender            | Male/ Female                    |
| Married           | Applicant married (Y/N)         |
| Dependents        | Number of Dependents            |
| Education         | Graduate/ Non-Graduate          |
| Self_Employed     | Self_Employed(Y/N)              |
| ApplicantIncome   | Applicant Income                |
| CoapplicantIncome | Coapplicant 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)             |

Various steps are included in making this process.

Packages Used are NumPy, Matplotlib, Boxplot and sklearn.

**Pandas:** Pandas is mainly used for data analysis. Pandas allows importing data from various file formats such as comma-separated values, JSON, SQL, Microsoft Excel. Pandas allows various data manipulation operations such as merging, reshaping, selecting, as well as data cleaning, and data wrangling features.

**NumPy:** NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level

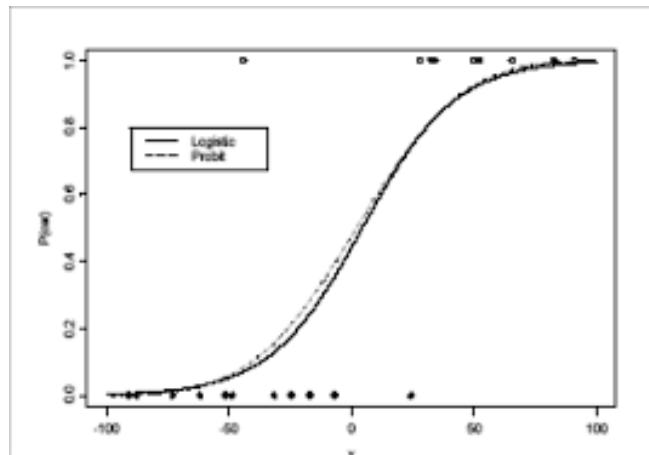
mathematical functions to operate on these arrays.

**Matplotlib:** Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits and has advanced Mathematical features.

**Scikit-learn:** Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines.

**Logistic Regression:** Regression models are useful for predicting continuous (numeric) variables. However, the target value in Approved is binary and can only be values of 1 or 0. The applicant can either be issued a credit card or denied- they cannot receive a partial credit card. We could use linear regression to predict the approval decision using threshold and anything below assigned to 0 and anything above is assigned to 1. Unfortunately, the predicted values could be well outside of the 0 to 1 expected range. Therefore, linear or multivariate regression will not be effective for predicting the values. Instead, logistic regression will be more useful because it will produce probability that the target value is 1. Probabilities are always between 0 and 1 so the output will more closely match the target value range than linear regression.

The sigmoid function is a numerical function used to outline predicted values to probabilities. It maps any real value to another must be in between 0 and 1 which means it can exceed the limit, then it forms a curve like the “S” form. The S-structure curve is also known as the sigmoid function or the logistic function. logistic regression, we utilize the concept of threshold value probability of either 0 or 1 and the value below the threshold values tends to 0.



**FIG 4.4 Logistic Regression Model Graph**

## **4.5 MODULAR DESIGN**

To automate this Loan approval process, we need to follow steps below mentioned.

### **4.5.1 Collecting Data**

Step 1: Download Test and Train Data set from Kaggle Website.

Step 2: Understanding the fields of the Data set by using Data description.

Step 3: Importing NumPy and Pandas libraries.

Step 4: Reading the both Data sets .

Step 5: using head () func to see the data set is actually loaded or not.

### **4.5.2 Understanding the various features (columns) of the dataset.**

Step 1: Using describe() func to get summary of the training Data set.

Step 2: Get the unique values and their frequency of variable Property Area.

Step 3: Understanding Distribution of Numerical Variables like Applicant Income and Loan Amount by drawing Box Plot .

Step 4: Boxplot is drawn to understand distribution and find Outliers..

Step 5: We will find outliers from Boxplot in Education and Gender.

### **4.5.3 Understanding Distribution of Categorical Variables**

Step 1: Finding Cross table for Credit History and Loan status in train data set.

Step 2: Write Function to output percentage row wise in a cross table.

Step 3: We will see Loan approval rate for customers having Credit History (1) is 79%.

Step 4: Nullify Outliers of Loan Amount and Applicant Income .

Step 5: This can be done by using Log Transformation by combining both Applicant and Co applicant Income.

#### **4.5.4 Data Preparation For Model Building and Generic Classification**

Step 1: sklearn requires all inputs to be numeric, we should convert all our categorical variables into numeric.

Step 2: We will fill all the missing values in the dataset.

Step 3: Import models from scikit learn module.

Step 4: Generic function for making a classification model and accessing performance.

Step 5: Perform k-fold cross-validation with 5 folds.

Step 6: Training the algorithm using the predictors and target.

Step 7: Fit the model again so that it can be referred outside the function.

#### **4.5.5 Model Building and Testing**

Step 1: We have chosen Logistic Regression for Model Building.

Step 2: The chances of getting a loan will be higher for

- Applicants having a credit history (we observed this in exploration.)
- Applicants with higher applicant and co-applicant incomes.
- Applicants with higher education level.
- Properties in urban areas with high growth perspectives.

So let's make our model with 'Credit\_History', 'Education' & 'Gender'.

Step 3: Create logistic regression object.

Step 4: Train the model using the training sets.

Step 5: Reverse encoding for predicted outcome.

Step 6: Store Output and view Accuracy Rate and Cross Validation Score.

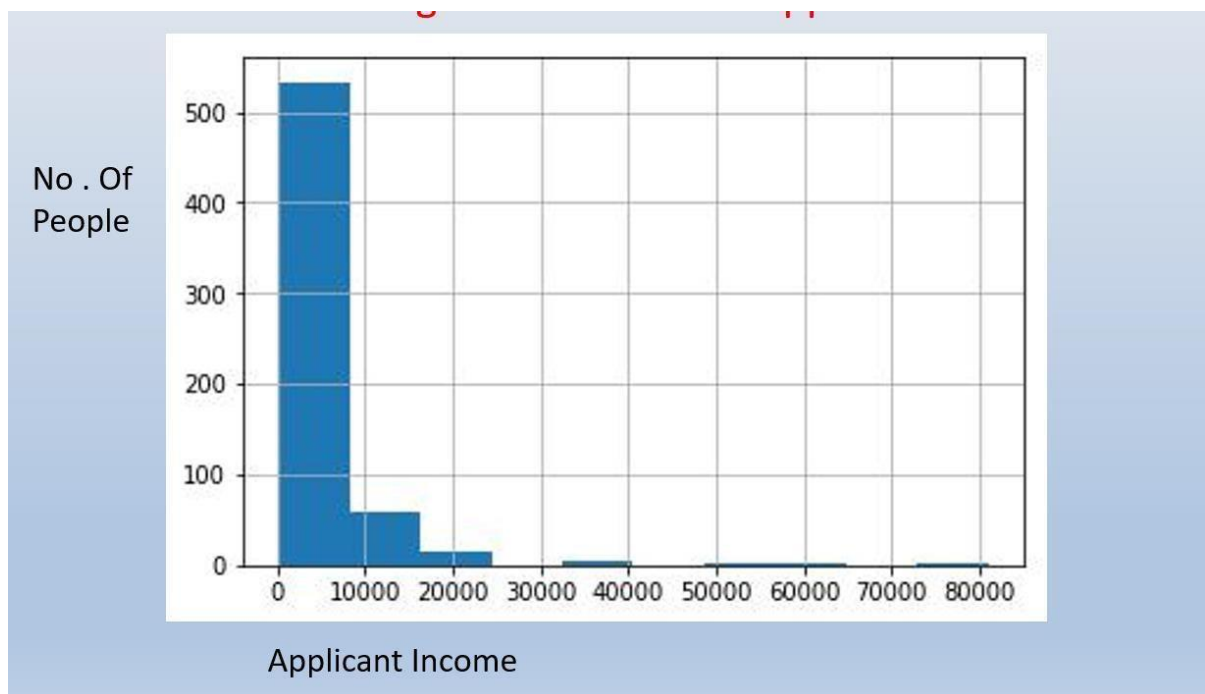
Accuracy Rate : It is the percentage of correct predictions for a given dataset

Cross Validation Score: Statistical method used to estimate the skill of Models

## CHAPTER 5

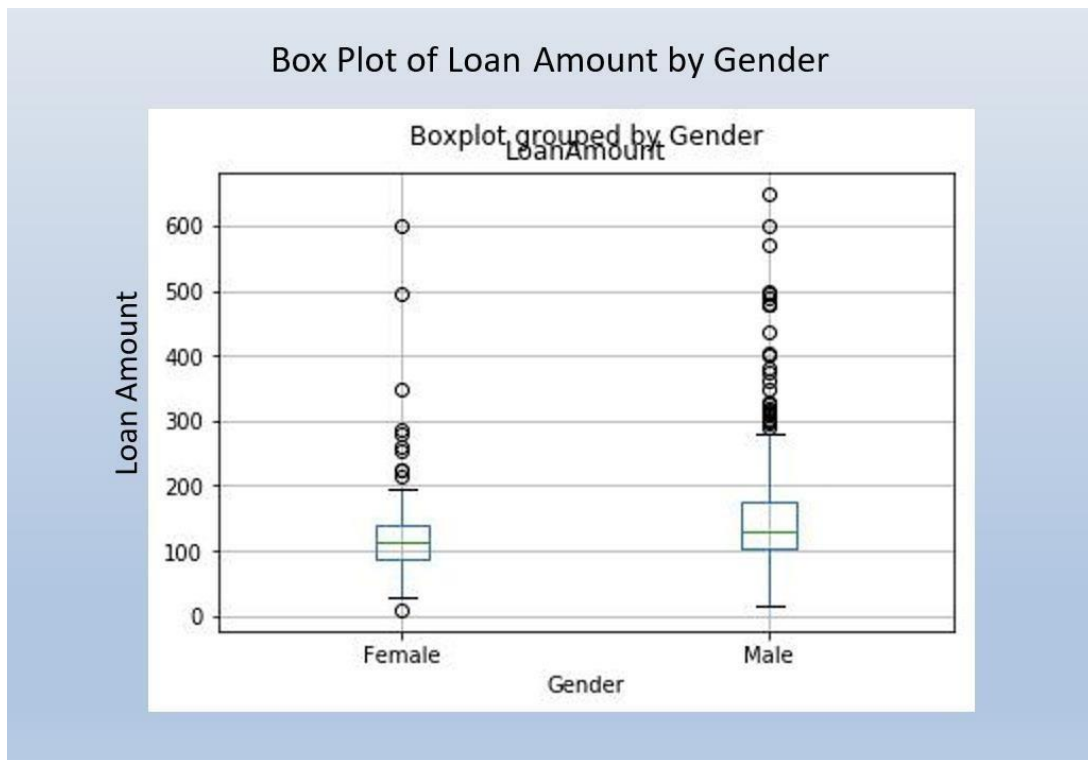
### RESULTS AND DISCUSSION

Finally, in our model by using logistic regression model we predict whether the loan is approved or not. In order to implement this various input variables were used to get the output. Whenever program takes the input data it gives the output in the form of binary i.e., either 0 or 1. If the output is 1 then '1' will be displayed and it indicates that loan is approved. If the output is 0 then '0' will be displayed and it indicates that loan is not approved. Here, we had implemented loan credibility prediction system that helps the organizations in making the right decision to approve or reject the loan request of the customers. This will definitely help the banking industry to open up efficient delivery channels. In this model, Logistic Regression algorithm is used for the prediction.

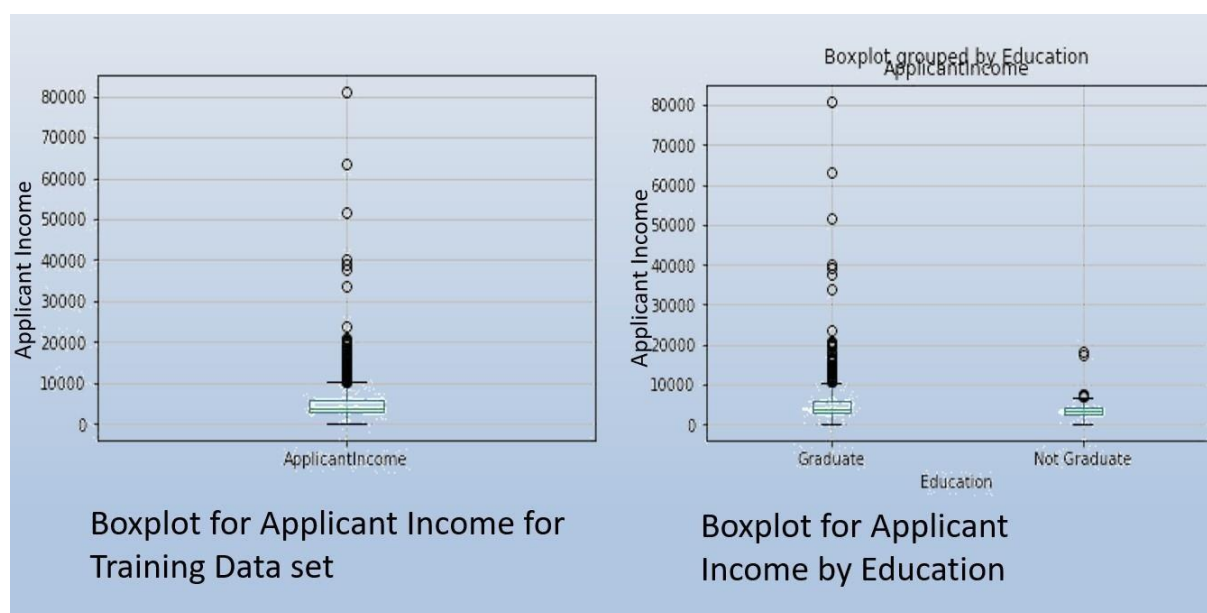


**Fig 5.1 Boxplot for Variable Applicant Income of Training Data set**

The above Box Plot confirms the presence of a lot of outliers/extreme values. This can be attributed to the income disparity in the society.

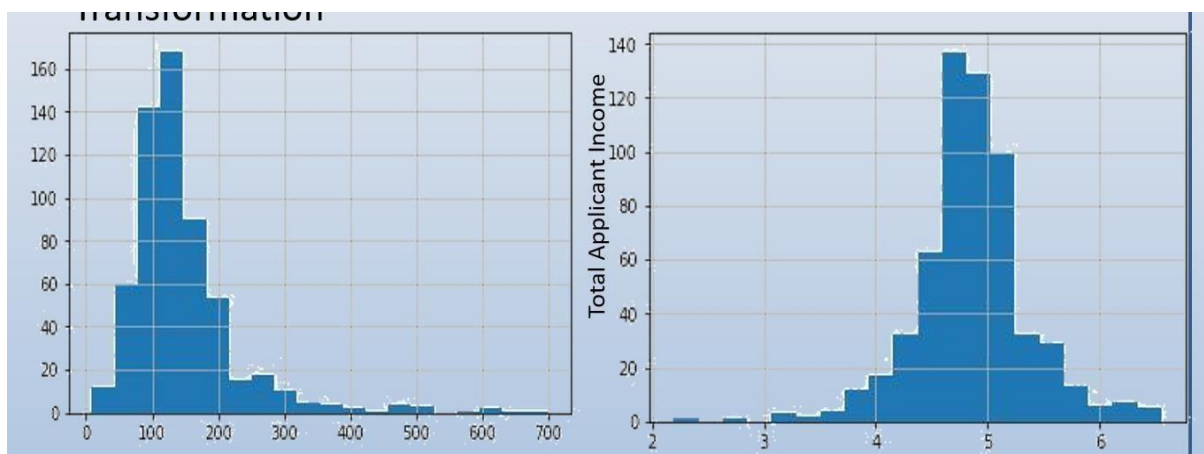


**Fig 5.2 Box Plot for understanding the distributions and to observe the outliers**  
 Above Fig 5.2 shows that Outliers are present in Male Gender than compared to female



**Fig 5.3 Box plot for both Applicant Income by Education**

In Fig 5.3, We can see that there is no substantial different between the mean income of graduate and non-graduates. But there are a higher number of graduates with very high incomes, which are appearing to be the outliers



**Fig 5.4 Box plot for Applicant Income before and after Log Transformation**

The extreme values are practically possible, i.e. some people might apply for high value loans due to specific needs. So instead of treating them as outliers, let's try a log transformation to nullify their effect. Also we combine both Applicant and Co Applicant Income for better results

```

In [21]: # Impute missing values for Gender
df['Gender'].fillna(df['Gender'].mode()[0],inplace=True)

# Impute missing values for Married
df['Married'].fillna(df['Married'].mode()[0],inplace=True)

# Impute missing values for Dependents
df['Dependents'].fillna(df['Dependents'].mode()[0],inplace=True)

# Impute missing values for Credit_History
df['Credit_History'].fillna(df['Credit_History'].mode()[0],inplace=True)

# Convert all non-numeric values to number
cat=['Gender','Married','Dependents','Education','Self_Employed','Credit_History','Property_Area']

for var in cat:
    le = preprocessing.LabelEncoder()
    df[var]=le.fit_transform(df[var].astype('str'))

df.dtypes

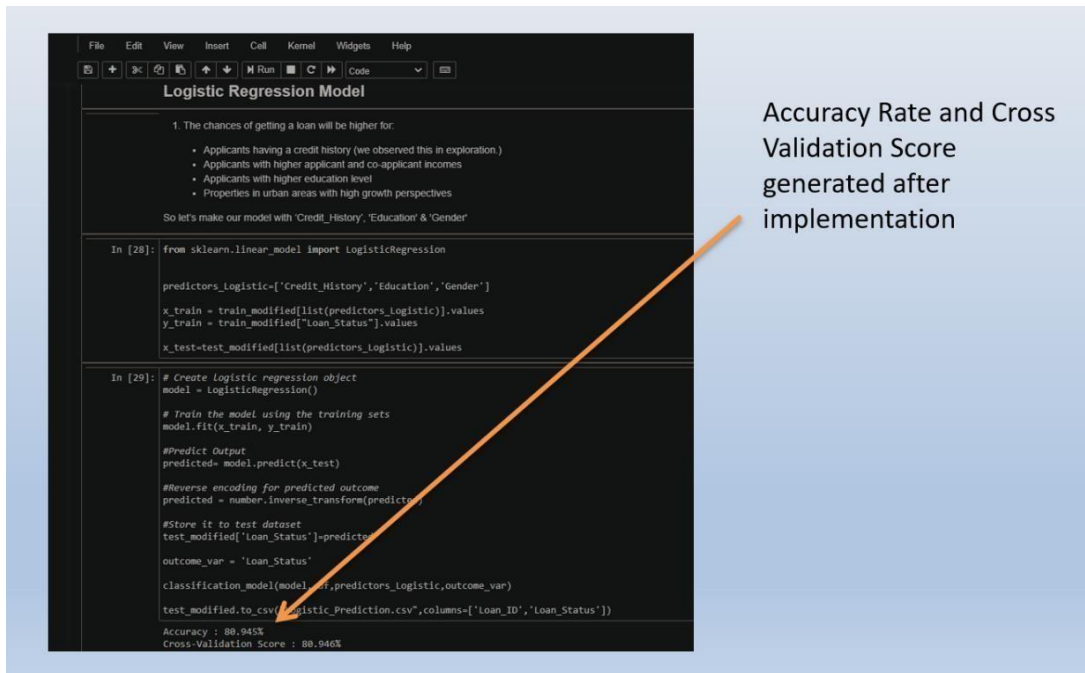
```

|                   |         |
|-------------------|---------|
| Loan_ID           | object  |
| Gender            | int64   |
| Married           | int64   |
| Dependents        | int64   |
| Education         | int64   |
| Self_Employed     | int64   |
| ApplicantIncome   | int64   |
| CoapplicantIncome | float64 |
| LoanAmount        | float64 |
| Loan_Amount_Term  | float64 |
| Credit_History    | int64   |
| Property_Area     | int64   |
| Loan_Status       | object  |
| TotalIncome       | float64 |
| LoanAmount_log    | float64 |
| dtype:            | object  |

Dataset after filling Missing Values And Converting Non-Categorical Values into Numeric Values.

**Fig 5.5 Data Preparation for Model Building**

In Fig 5.5 sklearn requires all inputs to be numeric, we should convert all our categorical variables into numeric by encoding the categories. Before that we will fill all the missing values in the dataset.



**Fig 5.6 Accuracy Rate and Cross Validation Score**

In Fig 5.6 we can see the Accuracy Rate is 80.945% with Cross Validation Score of 80.946% which is best value when compared to recent models based on other Machine Learning Models. This better values makes us understand how accurate this model.

|    | A | B       | C           | D | E | F | G | H | I |
|----|---|---------|-------------|---|---|---|---|---|---|
| 1  |   | Loan_ID | Loan_Status |   |   |   |   |   |   |
| 2  |   | 0       | LP001015    | Y |   |   |   |   |   |
| 3  |   | 1       | LP001022    | Y |   |   |   |   |   |
| 4  |   | 2       | LP001031    | Y |   |   |   |   |   |
| 5  |   | 3       | LP001035    | Y |   |   |   |   |   |
| 6  |   | 4       | LP001051    | Y |   |   |   |   |   |
| 7  |   | 5       | LP001054    | Y |   |   |   |   |   |
| 8  |   | 6       | LP001055    | Y |   |   |   |   |   |
| 9  |   | 7       | LP001056    | Y |   |   |   |   |   |
| 10 |   | 8       | LP001059    | Y |   |   |   |   |   |
| 11 |   | 9       | LP001067    | Y |   |   |   |   |   |
| 12 |   | 10      | LP001078    | Y |   |   |   |   |   |
| 13 |   | 11      | LP001082    | Y |   |   |   |   |   |
| 14 |   | 12      | LP001083    | Y |   |   |   |   |   |

**Fig 5.7 Output.csv File**

Output Generated and stored in Local Disk as shown in Fig 5.7 with Loan Status—

## **CHAPTER 6**

### **CONCLUSION AND FUTURE WORK**

#### **CONCLUSION**

From a proper analysis of positive points and constraints on the component, it can be safely concluded that the product is a highly efficient component. This application is working properly and meeting to all Banker requirements. This component can be easily plugged in many other systems. There have been numbers cases of computer glitches, errors in content and most important weight of features is fixed in automated prediction system, So in the near future the so – called software could be made more secure, reliable and dynamic weight adjustment

Finally, in our model by using logistic regression model we predict whether the loan is approved or not. In order to implement this various input variables were used to get the output. Whenever program takes the input data it gives the output in the form of binary i.e., either 0 or 1. If the output is 1 then '1' will be displayed and it indicates that loan is approved. If the output is 0 then '0' will be displayed and it indicates that loan is not approved.

In this paper, data preprocessing and transformation techniques are applied and results are generated by implementing analytical models. The performance is analyzed using the confusion matrix table. We can also use this model to make detail testing selections. Any credit application that does not have the same outcome as predicted by the model is potential audit exception. The inherent risk is that a credit card was issued to someone that should have been denied. This account is more likely to default than a properly approved account which, in turn, exposes the company to loss.

#### **FUTURE WORK**

Here, we had implemented loan credibility prediction system that helps the organizations in making the right decision to approve or reject the loan request of the customers. This will definitely help the banking industry to open up efficient delivery channels. In this model, Logistic Regression algorithm is used for the prediction. Incorporation of other techniques that outperform the performance of popular data mining models has to be implemented and tested for the domain.

The inherent risk is that a credit card was issued to someone that should have been

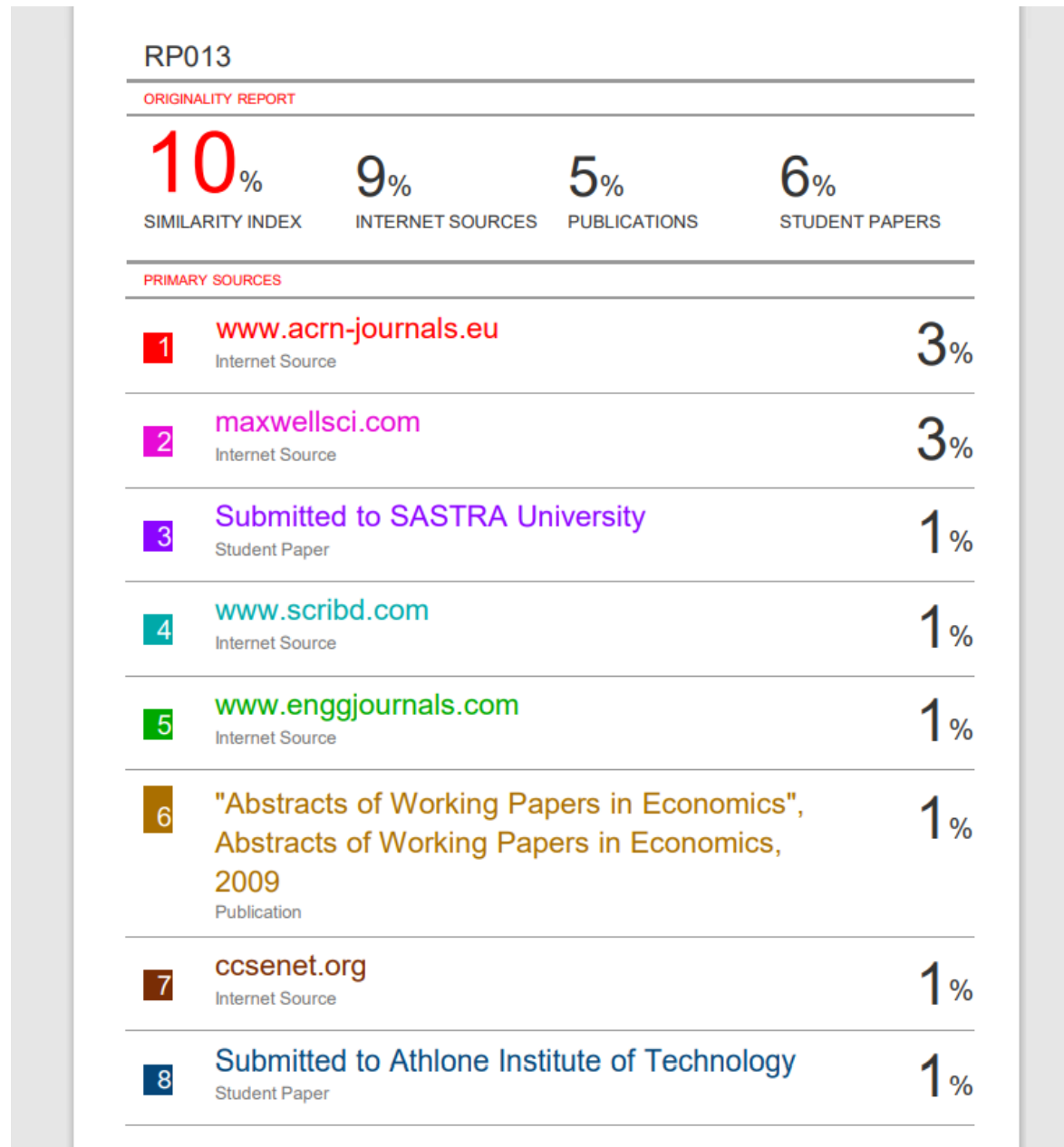
denied. This account is more likely to default than a properly approved account which, in turn, exposes the company to loss. The different machine learning models can be implemented and the performance can be compared. We can also do this automation by using Neural Networks, Naive Bayes and some other Machine Learning Algorithms which are in progress.

## REFERENCES

- [1]. Sudhamathy G and Jothi Venkateswaran “Analytics Using R for Predicting Credit Defaulters”, IEEE international conference on advances in computer applications (ICACA), 978-1-5090-3770-4, 2016.
- [2]. M. Sudhakar, and C.V.K. Reddy, “Two Step Credit Risk Assessment Model For Retail Bank Loan Applications Using Decision Tree Data Mining Technique”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 5, no.3, pp. 705-718, 2016.
- [3]. J.H. Aboobyda, and M.A. Tarig, “Developing Prediction Model Of Loan Risk In Banks Using Data Mining”, Machine Learning and Applications: An International Journal (MLAIJ), vol. 3, no.1, pp. 1–9, 2016.
- [4]. Z. Somayyeh, and M. Abdolkarim, “Natural Customer Ranking of Banks in Terms of Credit Risk by Using Data Mining A Case Study: Branches of Mellat Bank of Iran”, Jurnal UMP Social Sciences and Technology Management, vol. 3, no. 2, pp. 307–316, 2015.
- [5]. A.B. Hussain, and F.K.E. Shorouq, “Credit risk assessment model for Jordanian commercial banks: Neural scoring approach”, Review of Development Finance, Elsevier, vol. 4, pp. 20–28, 2014.
- [6]. T. Harris, “Quantitative credit risk assessment using support vector machines: Broad versus Narrow default definitions”, Expert Systems with Applications, vol. 40, pp. 4404– 4413, 2013.
- [7]. Dileep B. Desai, Dr. R.V.Kulkarni “A Review: Application of Data Mining Tools in CRM for Selected Banks”, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (2), 2013, 199 – 201.
- [8].Gang Wang, Jian Ma, “Study of corporate credit risk prediction based on integrating boosting and random subspace”, 2011.
- [9] Hussain Ali Bekhet , Shorouq Fathi Kamel Eletter , “Credit risk assessment model for Jordanian commercial banks: Neural scoring approach” ,Apr2014.
- [10] M. Yaghini , T. Zhiyan , and M. Fallahi, “A Prediction Model for Recognition of Bad Credit Customers in Saman Bank Using Neural Networks”, 2011

## APPENDIX

### A. PLAGIARISM REPORT



# Customer Loan Approval Prediction using Logistic Regression

MAHANKALI GOPINATH <sup>[1]</sup>, K. SRINIVAS SHANKAR MAHEEP <sup>[2]</sup>, R. SETHURAMAN <sup>[3]</sup>

[1],[2] UG Student, Dept of CSE, Sathyabama Institute of Science and Technology, Chennai, India

[3] Associate Professor, Dept of CSE, Sathyabama Institute of Science and Technology, Chennai, India

## ABSTRACT

In Banking Sector, a loan is a process of lending or borrowing a sum of money by one or more individuals, organizations, etc. from Banks. The Person who lends that money from respective financier incurs a debt, and he is responsible to pay back the money with the Interest decided by Bank within a certain period. Generally what Banks look into before applying for a loan is Credit History, Credit loss and Income of Applicant. So basically, loans play a major role regarding Income for Bank. Due to rapid urban development people who are applying for loans got increased rapidly. As a result, finding the applicant to whom loan can be approved become a complexed process. In this paper, we want to automate the loan eligibility process (real time) based on customer details. Fields that required are Marital Status, Income, Education, Gender, Number of Dependents, Loan Amount, Credit History and others. To predict the status, we will use Logistic Regression to spot the customers, those are eligible for loan amount so that bank can reach out them for granting loans to those people who can payback in a given time.

**Keywords:** Loan, Exploratory Data Analysis, Prediction, Logistic Regression

## Introduction

The banking industry plays a significant role in present mostly in developing countries where money is usually required for all of them, so they will increase their market to capital value by gaining profits. Banks allows their customers to save lots of money in individual accounts. So, then Banks allows to lend money to business people or others who can utilize it for his or her capital growth and meet their Business requirements, and payback to the bank within a specific period of your time including Interest amount. So, interest is the profit gained by the banks by giving loans to folks that are in need. But Banks are worried about whether the person whose loan got granted will be ready to payback loan amount or not. so as to predict it, they basically inspect things associated with applicant like Credit score, Applicant Income. Credit Score. Here the credit score plays the key role in information given by the customer. In most scenarios Credit score is required for Loan sanction. If the applicant didn't payback his loan amount, then eventually his credit score will automatically get decreased. Giving Loans to people is one among the most business strategies for pretty much every bank. Banks will get most of the profits from the loan within the types of Interest. The most goal of Bank authorities is to

grant loans to trustworthy people, so they'll pay back on the deadline comes. In recent times, banks are approving loans at their customers after a step-by-step procedure, but there's still no guarantee that the applicant's loan was granted or not. To approve loan, banks will undergo to estimate risk involved within the application, which is important for them as they cannot lend money to those that cannot afford to pay back in time which affects their economic status during this huge competitive market. So, we've got collected Dataset from the Kaggle contains loan applicant's details contains various fields like Gender, Applicant Income, Credit history etc.

After doing an Exploratory Data analysis on these data sets, we have discovered that probabilities of getting Loan granted are higher for applicants who have credit history equals to '1' with greater applicant income, with Education level mostly graduated and eventually who lives urban areas with properties. So, we'd like a model supported Fields Credit history, Education and Gender of Applicant. This model developed using Logistic Regression. we'll use this model to check with another data set and results obtained are stored in other file with Predicted Approval status as 'Yes' or 'No'. We have chosen Logistic Regression because

it gives an Accuracy Rate of 80.945% approximately.

## 2. Literature Review

In **Somayyeh. Z et al** [1] a model was proposed for identifying and predict the right customers who have applied for loan. Decision Tree is applied to estimate the traits where accuracy rate is not much appreciable.

In **Sudhamathy G et al** [2], Banks hold huge volumes of customer information from which they are unable to judge. Data Mining is a promising area of an analysis which aims to get useful knowledge from an amount of data sets. This work aims to develop a model. It was a decision tree-based classification model which uses the functions available in the R Package. Before building the model, the dataset is processed, made it ready to provide effective predictions. The final model is used for predict using the test dataset and the experimental results shows the efficiency of the model.

In **Dileep. B et al** [3], Data analysis was done by figuring out techniques like Bayes classification, Decision Tree, Boosting, Bagging algorithm, Random Forest etc. Techniques like Support Vector Machine, Logistic Regression, k-means algorithm Neural Network, Perception model are combined in this model. The accuracy rate for each of these techniques is studied. The results show the overall performance is very good based on accuracy.

In **Allen et al** [4], The authors studied the economic effects of small business credit scoring (SBCS) with both high average prices and more risk levels for small business credits below \$100,000. These holdings are stable with a net increase in lending to risky “marginal borrowers” who would not receive credit, but who would pay relatively high prices when they are funded. The authors also find that a) bank-specific and industry-wide learning curves are very important b) SBCS effects differ for banks that follow “rules” and c) SBCS effects differ to people with slightly larger credits.

In **Altman, E. I.** [5], This paper provides some empirical results of a study considering financial ratios as predictors of Japanese corporate failure. A few empirical studies related to corporate bankruptcy in Japan have been taken into analysis. However, the results of come after studies are not generalizable, due to the limited data is provided. In contrast, the model proposed in this is independent of industry zone and size. This study shows that the

model can predict bankruptcy with approximately 86.14% accuracy of industry and size.

In **J.H. Aboobyda et al** [6], In this paper, a model for classifying loan risk in the banking sector is implemented by using data mining algorithms. The model has been prepared using data from the banking sector to predict the loan status. Three algorithms have been used to build the proposed model: They are j48, Naive Bayes and Bayes Net. Using the Weka application, the model has been implemented and tested. The results have been discussed and a comparison between these three algorithms was conducted. J48 was selected as best among the three based on accuracy.

In **A.B. Hussain et al** [7], Two data mining models were created calculating the credit scoring that helps in decision making of giving loans for the banks in Jordan citizen. This paper will help to find the accurate way to identify the right customers By the accuracy rate, the regression model is found to perform better than the radial function model.

In **T. Harris** [8], This work tries to see the probability of default as a tool to live credit risk in an exceedingly Tunisian bank. A scoring model was built per the normal technique of logistic regression (LR), and computer science techniques i.e. (ANN) and SVM. Then a comparison was made between these models using performance metrics like the confusion matrix and therefore the area under the ROC curve (AUC) to spot the foremost efficient model. Our results show that the Radial Basis Function kernel SVM was the foremost performing method in terms of accuracy, sensitivity and specificity with the smallest amount error rates. Thus, within the Tunisia context, this model is worth implementing in banking institutions so as to boost their credit risk management measures to watch and control credit.

In **Charles Kwofie et al** [9], This study shows the performance of logistic regression in predicting probability of default using data from a microfinance company. A logistic multivariate analysis was conducted to predict default status of loan beneficiaries using 90 sampled beneficiaries for model building and 40 out of sample beneficiaries for prediction. Age, legal status, gender number of years of education, number of years in business and base capital were used as predictors. The predictors that were significant within the model were legal status, number of years in business and base capital. The variability within the response variable within the logistic regression was very weak.

### 3. Proposed Work

Python has could be a good area for data analytical which helps us in analysing the data with better models in data science. The libraries in python make the prediction for loan data and results with multiple terms considering all properties of the customer in terms of prediction. Logistic Regression is deployed to create the model and used to get the output by predicting results accurately. Credit history could be a method that helps the bank to rationalize its process for credit granting decisions. Its principle is to synthesize a group of financial ratios together able to distinguish between good and bad customers. For predicting results, we have collected Data sets in Kaggle. In order to build this model, we will import some python packages which will be used to analyse the data sets. They are NumPy, Pandas, Matplotlib and sklearn Libraries. All these libraries are available in Python.

#### 3.1 Dataset Description

There are two Datasets which are required to build this model. 1) Train Dataset and 2) Test Dataset. The Test dataset contains list of customers applied for loan. By using Train Dataset, we can train the model and use it to predict loan status for Test dataset. The dataset is in CSV format. In python, Pandas is used to read the dataset. Refer the below table for field variables of the dataset.

| Variable           | Description                     |
|--------------------|---------------------------------|
| Loan_ID            | Unique Loan ID Number           |
| Gender             | Male/ Female                    |
| Married            | Applicant Married(Y/N)          |
| Dependents         | Number of dependents            |
| Education          | Graduate/Under Graduate         |
| Self-Employed      | Self Employed (Y/N)             |
| Applicant_Income   | Applicant Income                |
| Coapplicant_Income | Co Applicant Income             |
| Loan_Amount        | Loan Amount in thousands        |
| Loan_Amount_Term   | Term in Months                  |
| Credit_History     | Credit History meets guidelines |
| Property_Area      | Urban/ Rural                    |
| Loan_Status        | Loan Approved (Y/N)             |

### 3.2 System Requirements

#### Software Requirements

- Windows XP, Windows 7, Windows 10
- Python 3.5 Mozilla Firefox (or any browser)
- Jupyter Notebook IDLE.

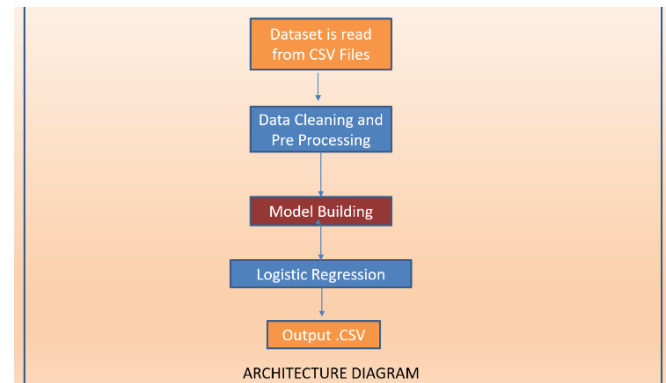
#### Hardware Requirements

- Minimum HDD: 20 GB
- Random Access Memory: 512 MB
- I3 processor-based computer or higher.

### 4. Software Methodology

After reading the dataset, we will implement Exploratory Data Analysis to understand and find out the outliers in the dataset.

#### 4.1 Block Diagram



In the above diagram, It shows what are the steps involved in building this model.

#### 4.2 Modules

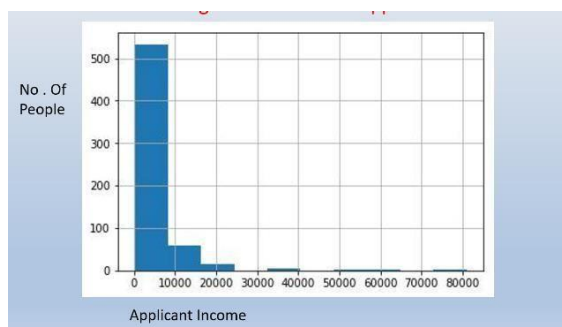
The Customer Loan Approval Prediction Model has mainly 3 modules. They are

- 1) Reading and Cleaning Dataset.
- 2) Model Building.
- 3) Testing Dataset with Model.

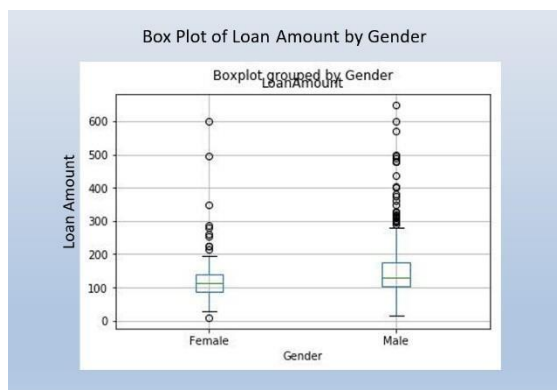
Let us look into each Module

**a) Reading and Cleaning Dataset:** We need to import Pandas, NumPy, and sklearn libraries and use them to process the information. Reading both training and testing dataset using Pandas. Now Exploratory Data Analysis (EDA) is used to analyse the Dataset and identify specific Outliers. By Using head () function, we are going to be ready to see the first 10 rows of the dataset so that we'll have a clear picture of what fields does dataset contains in it. After then we will store the length of rows and columns within the dataset. we'd like to grasp the

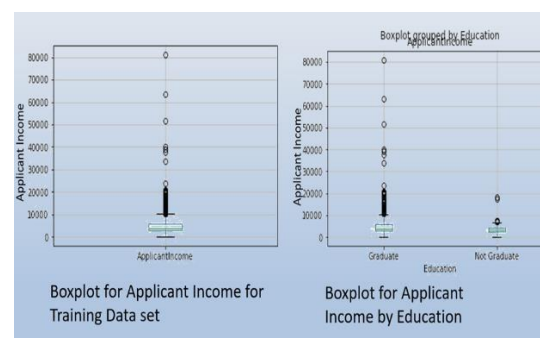
varied features and columns of the dataset. We will use describe () function to urge the summary of all the numerical and categorical value fields within the dataset which contains Applicant income, Co-applicant income, Credit history, and Loan amount. For the non-numerical values (e.g., Property Area, Credit History, etc.), we can take a look at distribution to grasp whether or not they add up or not. Understanding Distribution of Numerical Variables like Applicant Income and Loan The amount may be done by using Boxplot to know and finding the outliers of the Dataset fields. But there are a better number of graduates with very high incomes, which are appearing to be the outliers. Now it's time to understanding distribution of Categorical Variables. By making a Cross table with both Credit History and Loan Amount Fields we will see that Loans Approved within the Train Data set are more with applicants having credit history equals 1. Then we are going to write a function to search out the proportion of applicants whose loans are approved with credit history adequate to 1 and it shows more than 79% of individuals have gotten loans with a Credit history of 1. Now we are going to move forward and understand outliers in an exceedingly better way within the next module to create the model. Outliers are slightly having extreme values when compared to normal ones.



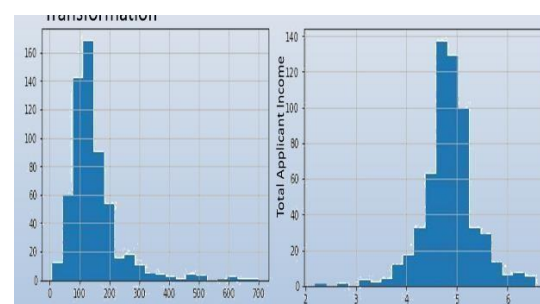
Boxplot for Applicant Income in Train Dataset. The above Box Plot confirms the presence of a lot of outliers/extreme values. This can be attributed to the income disparity in the society.



This Boxplot of Loan Amount by Gender. Outliers are present in Male Gender than compared to female



Box plot for Applicant Income by Education shows that there are a higher number of graduates with very high incomes, which are appearing to be the outliers.



The extreme values are practically possible, i.e., some people might apply for high value loans due to specific needs. So instead of treating them as outliers, let's try a log transformation to nullify their effect. Also, we combine both Applicant and Co Applicant Income for better results.

Also, we need to fill missing values for Data preparation for Model Building in next modules. We generally fill the missing values using pandas by Null values.

## b) Model Building

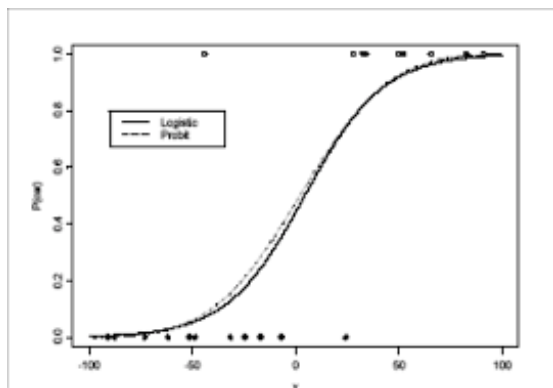
After converting all our categorical variables into numeric values as sklearn all inputs to be numeric. Here we can see all the variables in the dataset are in numeric Data type. Now we will build Generic Classification Function by importing scikit learn module and accessing performance. Fit the model, make predictions on dataset Perform k-fold cross-validation with 5 folds. Training the algorithm using target and predictors. Fit the model again so that it can be referred outside the function by adding Accuracy rate and cross validation score.

Now Combining both train and test dataset and create a flag for both datasets. Look at the summary of missing values of both datasets and fill them with

Null values for both categorical and numeric variables

```
Out[62]: Loan_ID      object
Gender      int64
Married     int64
Dependents  int64
Education   int64
Self_Employed int64
ApplicantIncome int64
CoapplicantIncome float64
LoanAmount  float64
Loan_Amount_Term float64
Credit_History int64
Property_Area int64
Loan_Status object
dtype: object
```

Now all the variables are converted to numeric for processing in model. Now we create a new column named as Total Income by adding both Applicant Income and Co Applicant Income and create label encoders for them. We have used Logistic Regression in building the model.



Logistic Regression is one of the popular algorithms that belongs to Supervised Learning Techniques. In this algorithm, the result must lie between 0 and 1 which means Yes or No (True/False). So, it was mainly used for solving classification-based problems. It predicts only two values, so it makes our model job very accurate and give good results as we expected.

- Applicants having a credit history (we observed this in exploration.)
- Applicants with higher applicant and co-applicant incomes.
- Applicants with higher education level.
- Properties in urban areas with high growth possibilities.

So, we made our model with 'Credit\_History', 'Education' & 'Gender'.

### c) Testing Dataset with Model

We can see the Accuracy Rate is 80.945% with Cross Validation Score of 80.946% which is best value when compared to recent models based on other Machine Learning Models.

**Accuracy Rate:** It is the percentage of correct predictions for a given dataset.

**Cross Validation Score:** Statistical method used to estimate the skill of Models

```
x_test=x_test_modified[list(predictors_logistic)].values

In [29]: # Create logistic regression object
model = LogisticRegression()

# Train the model using the training sets
model.fit(x_train, y_train)

# Predict Output
predicted= model.predict(x_test)

# Reverse encoding for predicted outcome
predicted = number.inverse_transform(predicted)

# Store it to test dataset
test_modified['Loan_Status']=predicted
outcome_var = 'Loan_Status'

classification_model(model, df,predictors_logistic,outcome_var)

test_modified.to_csv("Logistic_Prediction.csv",columns=['Loan_ID','Loan_Status'])

Accuracy : 80.945%
Cross-Validation Score : 80.946%
```

Output CSV file is stored in local storage with fields Unique Loan ID with predicted Loan Approval Status as Yes or No. By using this results Banks can go ahead and complete the loan approval process for right applicants.

| Loan_ID     | Loan_Status | Loan_Amount_Term |
|-------------|-------------|------------------|
| 0 LP001015  | Y           |                  |
| 1 LP001022  | Y           |                  |
| 2 LP001031  | Y           |                  |
| 3 LP001035  | Y           |                  |
| 4 LP001051  | Y           |                  |
| 5 LP001054  | Y           |                  |
| 6 LP001055  | Y           |                  |
| 7 LP001056  | Y           |                  |
| 8 LP001059  | Y           |                  |
| 9 LP001067  | Y           |                  |
| 10 LP001078 | Y           |                  |
| 11 LP001082 | Y           |                  |
| 12 LP001083 | Y           |                  |

### Experimental Results

#### 4. Conclusion and Future Scope

Finally, by using logistic regression model we can predict whether the loan can be approved or not. so as to implement this various input variables were accustomed to get the output. Whenever a program takes the computer file it gives the output within the type of binary i.e., either 0 or 1. If the output is 1 then '1' is displayed and it indicates that the loan is approved. If the output is 0 then '0' is displayed and it indicates that the loan isn't approved. This will definitely help the banking system to open up efficient delivery channels by taking right decision in approving or rejecting Loan applicants

## 5. References

1. Z. Somayyeh, and M. Abdolkarim “Natural Customer Ranking of Banks in Terms of Credit Risk by Using Data Mining A Case Study: Branches of Mellat Bank of Iran”, Jurnal UMP Social Sciences and Technology Management, vol. 3, no. 2, pp. 307–316, 2015.
2. Sudhamathy G and Jothi Venkateswara” Analytics Using R for Predicting Credit Defaulters”, IEEE international conference on advances in computer applications (ICACA), 978-1-5090-3770-4, 2016.
3. Dileep B. Desai, Dr. R.V.Kulkarni “A Review: Application of Data Mining Tools in CRM for Selected Banks”, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (2), 2013, 199 – 201.
4. Allen, N., Berger, W., Scott, F., & Nathan, H. M. (2002). Credit Scoring and the Availability, Price, and Risk of Small Business Credit. FRB of Atlanta Working Paper No. 2002-6, FEDS Working Paper No. 2002-26.
5. Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. The Journal of Finance, 23(4), 589-609.
6. J.H. Aboobyda, and M.A. Tarig, “Developing Prediction Model Of Loan Risk In Banks Using Data Mining”, Machine Learning and Applications: An International Journal (MLAIJ), vol. 3, no.1, pp. 1–9, 2016.
7. A.B. Hussain, and F.K.E. Shorouq, “Credit risk assessment model for Jordanian commercial banks: Neural scoring approach”, Review of Development Finance, Elsevier, vol. 4, pp. 20–28, 2014.
8. T. Harris, “Quantitative credit risk assessment using support vector machines: Broad versus Narrow default definitions”, Expert Systems with Applications, vol. 40, pp. 4404– 4413, 2013.
9. Kwofie, Charles & Owusu-Ansah, Caleb & Boadi, Caleb. (2015). Predicting the Probability of Loan-Default: An Application of Binary Logistic Regression. Research Journal of Mathematics and Statistics. 7. 46-52. 10.19026/rjms.7.2206.

### C. SOURCE CODE

```
#Import models

from sklearn module:

from sklearn import metrics

from sklearn.cross_validation import KFold

#Generic function for making a classification model and accessing performance:
def classification_model(model, data, predictors, outcome):

#Fit the model:

model.fit(data[predictors],data[outcome])

#Make predictions on training set:

predictions = model.predict(data[predictors])

#Print accuracy accuracy = metrics.accuracy_score(predictions,data[outcome])

print ("Accuracy : %s" % "{0:.3%}".format(accuracy))

#Perform k-fold cross-validation with 5 folds

kf = KFold(data.shape[0], n_folds=5)

error = []
for train, test in kf:

# Filter training data

train_predictors = (data[predictors].iloc[train,:])

# The target we're using to train the algorithm.

train_target = data[outcome].iloc[train]

# Training the algorithm using the predictors and target.

model.fit(train_predictors, train_target)

#Record error from each cross-validation run

error.append(model.score(data[predictors].iloc[test,:], data[outcome].iloc[test]))

print ("Cross-Validation Score : %s" % "{0:.3%}".format(np.mean(error)))

#Fit the model again so that it can be referred outside the function:

model.fit(data[predictors],data[outcome])

#Combining both train and test dataset

#Create a flag for Train and Test Data set

df['Type']='Train'

test['Type']='Test'
```

```

fullData = pd.concat([df,test],axis=0, sort=True)

#Look at the available missing values in the dataset
fullData.isnull().sum()

#Create a new column as Total Income
fullData['TotalIncome']=fullData['ApplicantIncome'] + fullData['CoapplicantIncome']
fullData['TotalIncome_log'] = np.log(fullData['TotalIncome'])

#Histogram for Total Income fullData['TotalIncome_log'].hist(bins=20)

#create label encoders for categorical features

for var in cat_cols:

    number = LabelEncoder()

    fullData[var] = number.fit_transform(fullData[var].astype('str'))
    train_modified=fullData[fullData['Type']=='Train']
    test_modified=fullData[fullData['Type']=='Test']
    train_modified["Loan_Status"] =
    number.fit_transform(train_modified["Loan_Status"]).astype('str')

from sklearn.linear_model import LogisticRegression
predictors_Logistic=['Credit_History','Education','Gender']

x_train = train_modified[list(predictors_Logistic)].values
y_train = train_modified["Loan_Status"].values
x_test=test_modified[list(predictors_Logistic)].values

# Create logistic regression object
model = LogisticRegression()

# Train the model using the training sets
model.fit(x_train, y_train)

#Predict Output
predicted= model.predict(x_test)

#Reverse encoding for predicted outcome
predicted = number.inverse_transform(predicted)

#Store it to test dataset
test_modified['Loan_Status']=predicted

outcome_var = 'Loan_Status'

classification_model(model, df,predictors_Logistic,outcome_var)
test_modified.to_csv("Logistic_Prediction.csv",columns=['Loan_ID','Loan_Status'])

Accuracy: 80.945% Cross-Validation Score: 80.946%

```