**Project: - Data Science**

**Group Members: -**

|  |  |  |
| --- | --- | --- |
| **Name** | **Roll No** | **GR.No.** |
| Sanket Sonje | 322053 | 17u176 |
| Shardul Pathak | 322057 | 17u666 |

**Division: - COMP B**

**COURSE FACULTY: Prof. Leena A. Deshpande.**

**Title: - Loan Prediction System**

1. **Problem: -**

* Suppose some company wants to automate the loan eligibility process (real time) based on customer detail provided while filling online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others.
* To automate this process, they have given a problem to identify the customers segments, those are eligible for loan amount so that they can specifically target these customers. Here they have provided a data set.

1. **Objectives: -**

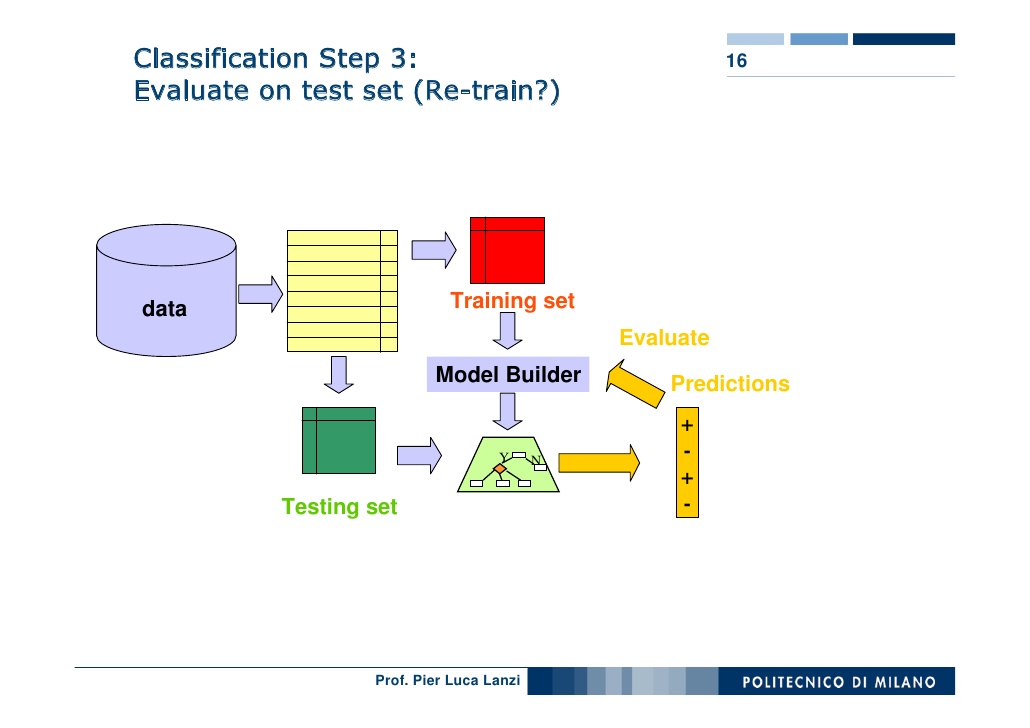
* Main Objective: -
  + Automate the Loan Eligibility Process (Real Time).
* Sub-Objectives: -
  + Extract the data from Datahack as a Resource.
  + Analyse the data and fit into the classification model.
  + Perform and Evaluate the model and plot the outcome of the project graphically**.**

1. **Data (Variable Description): -**

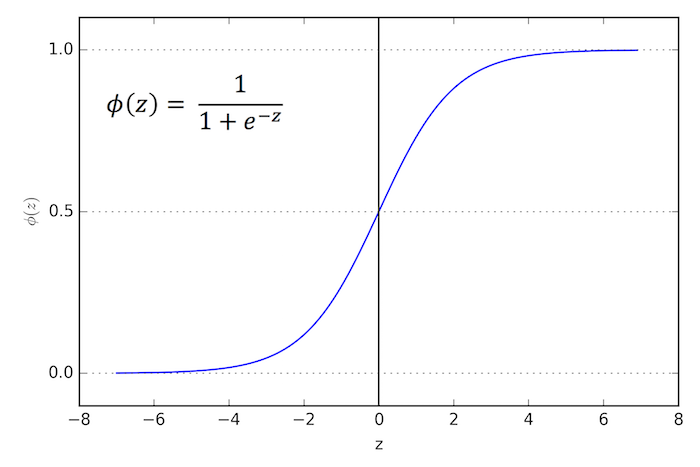
|  |  |  |
| --- | --- | --- |
| **SR.NO.** | **Variables** | **Description** |
| 1 | Loan\_ID | Unique Loan ID |
| 2 | Gender | Male/ Female |
| 3 | Married Applicant | married (Y/N) |
| 4 | Dependents | Number of dependents |
| 5 | Education | Applicant Education (Graduate/ Under Graduate) |
| 6 | Self\_Employed | Self-employed (Y/N) |
| 7 | ApplicantIncome | Applicant income |
| 8 | CoapplicantIncome | Co-applicant income |
| 9 | LoanAmount | Loan amount in thousands |
| 10 | Loan\_Amount\_Term | Term of loan in months |
| 11 | Credit\_History | credit history meets guidelines |
| 12 | Property\_Area | Urban/ Semi Urban/ Rural |
| 13 | Loan\_Status | Loan approved (Y/N) |

* Rows: 615
* Source: Datahack

1. **Analyse problem and convert it into data: -**
   1. How data collected: -
      * Datahack (Resource)
   2. How data stored:
      * Data is stored in the format of csv file.
        1. Test.csv
        2. Train.csv
2. **Flow of the Project: -**

****

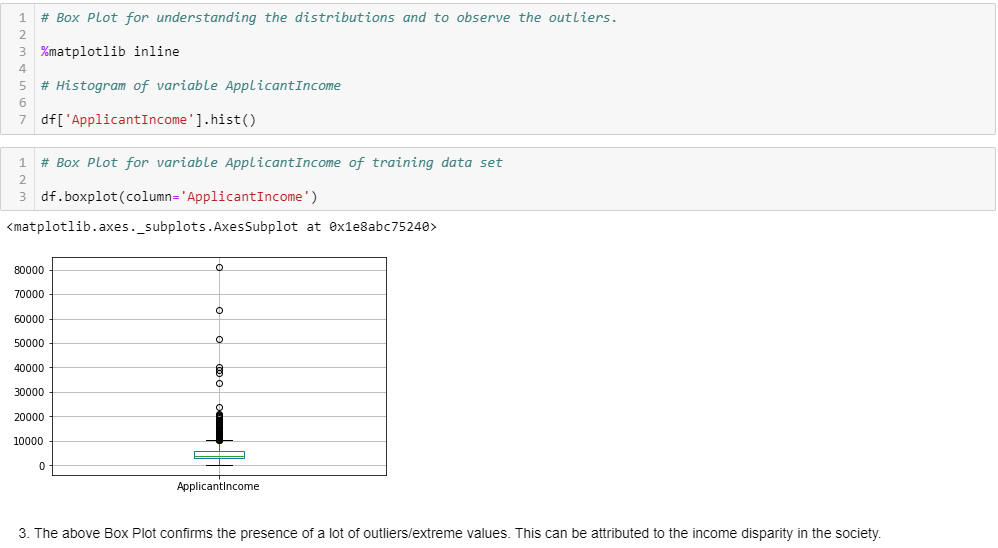
1. **Flow of the Project: -**
   1. Importing all libraries and storing data sets.
   2. Understanding various features of the datasets.
   3. Understanding the Distribution of Non-Categorial values.
   4. Understanding the Distribution of Categorial values.
   5. Replacing the Null values of Self-Employed column with maximum occurrence of the data.
   6. Nullifying Outliers of Loan Amount and Amount Income.
   7. Data preparation and model building.
   8. Testing and Validating score.
2. **Logistic Regression: -**
   1. Logistic Regression is used when the dependent variable(target) is categorical.
   2. For example,
      * To predict whether an email is spam (1) or (0)
      * Whether the tumor is malignant (1) or not (0)
   3. In our example, Our Targeted Value is “Loan Status” and it has values Yes or No.
   4. To validate our Accuracy, we calculate the Cross-Validation Score.
3. **Types of Logistic Regression: -**
   1. Binary Logistic Regression: -
      * The categorical response has only two 2 possible outcomes. Example: Spam or Not. We use this type.
   2. Multinomial Logistic Regression: -
      * Three or more categories without ordering. Example: Predicting which food is preferred more (Veg, Non-Veg, Vegan)
   3. Ordinal Logistic Regression: -
      * Three or more categories with ordering. Example: Movie rating from 1 to 5
4. **Logistic Regression Graph: -**

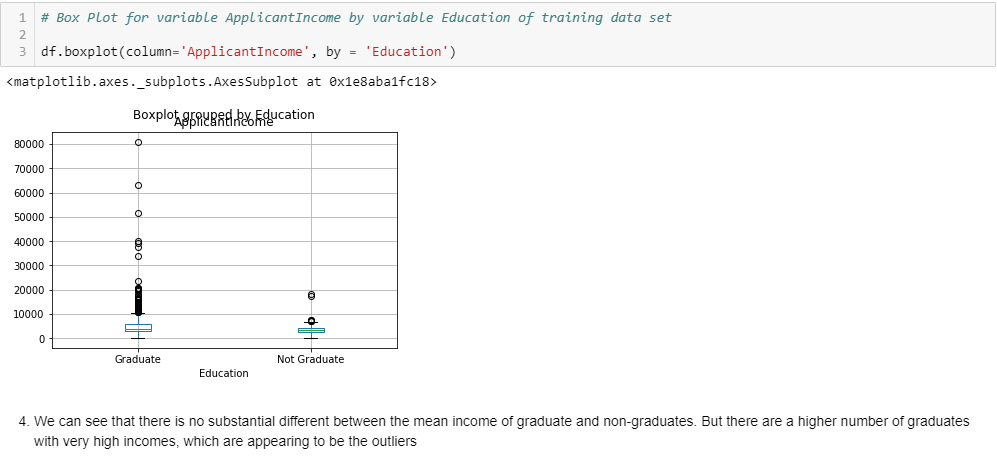
****

1. **Analyze problem and convert it into data: -**
   1. How data represented using statistical model: -
      * Data distribution: Binary Logistic Regression Model.
   2. Statistical operations performed: -
      * Min (), Avg (), describe (), Mode (), etc.
   3. Dealing with missing values: -
      * NAN, zero values /variable conversion.
2. **Model building and training: -**
   1. Model used: -
      * Logistic Regression Model
      * Binary output is expected.
   2. Libraries to be used: -
      * SKLEAN, SCIKIT, PANDAS, NUMPY, etc.
3. **Code: - Importing Libraries: -**

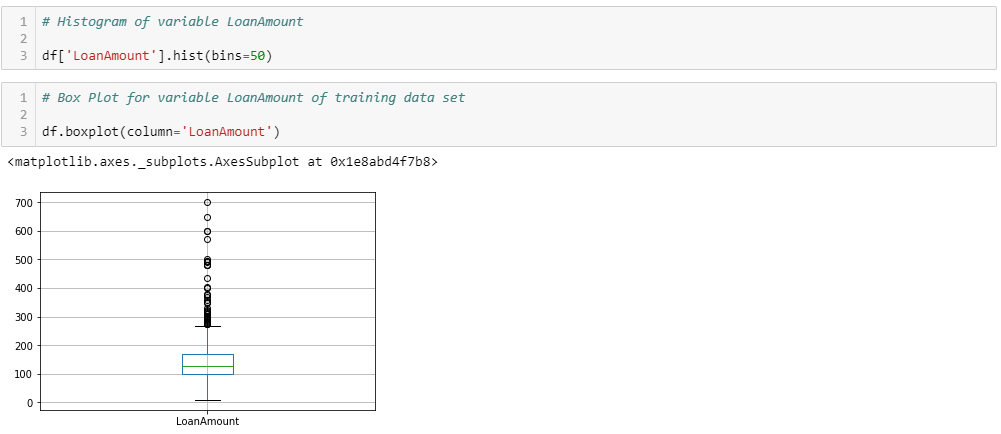
****

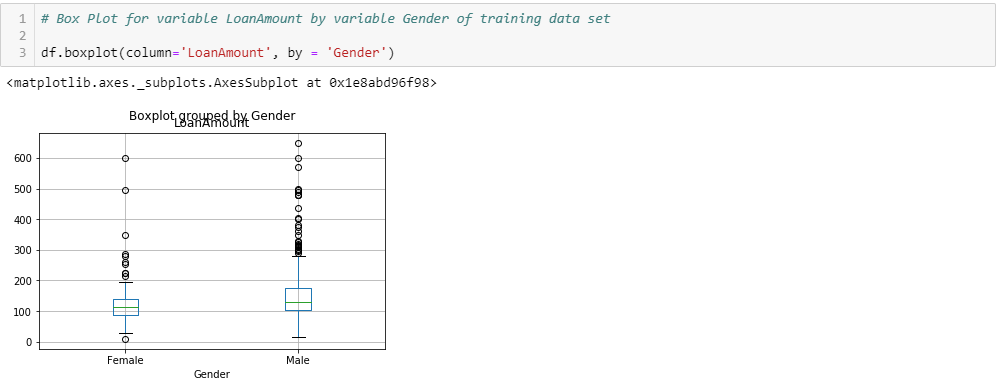
1. **Understanding Distribution of Numerical Values: -**
   1. **ApplicationIncome: -**

****

****

* 1. **LoanAmount: -**

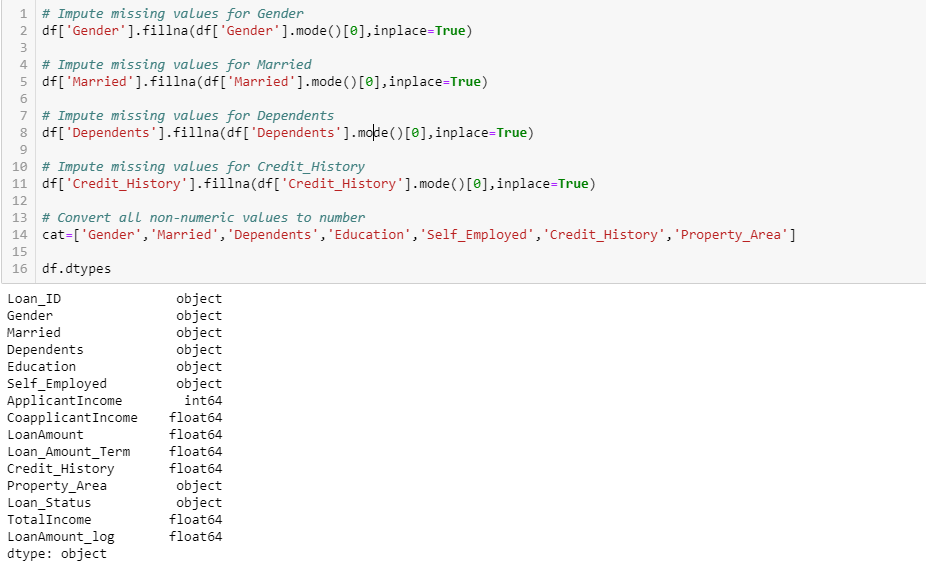
****

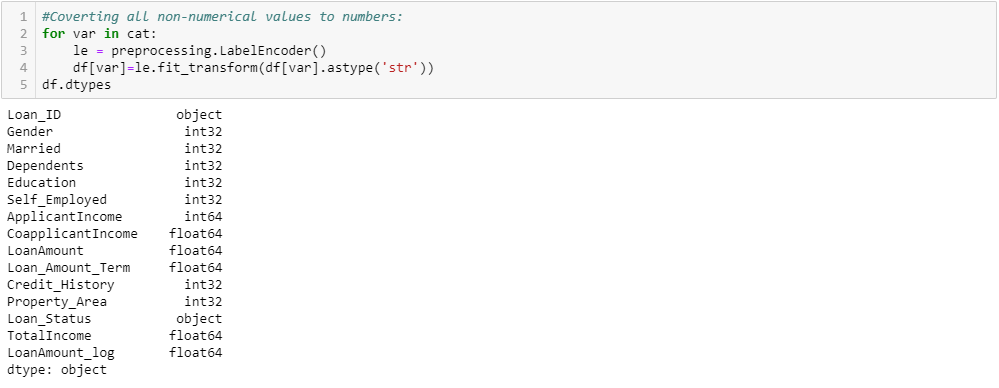
****

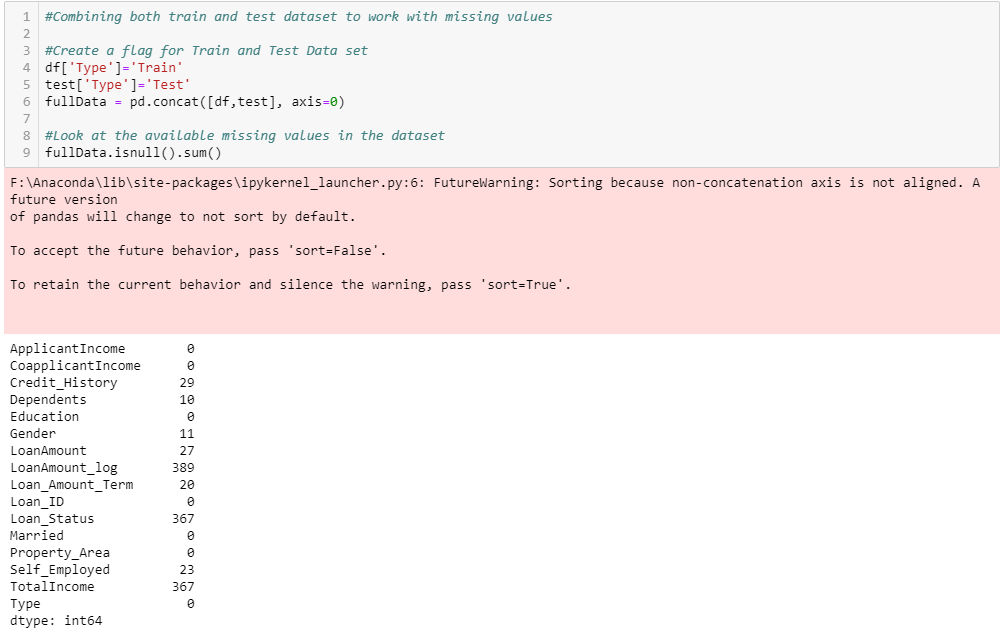
1. **Understanding Distribution of Categorical Variables: -**

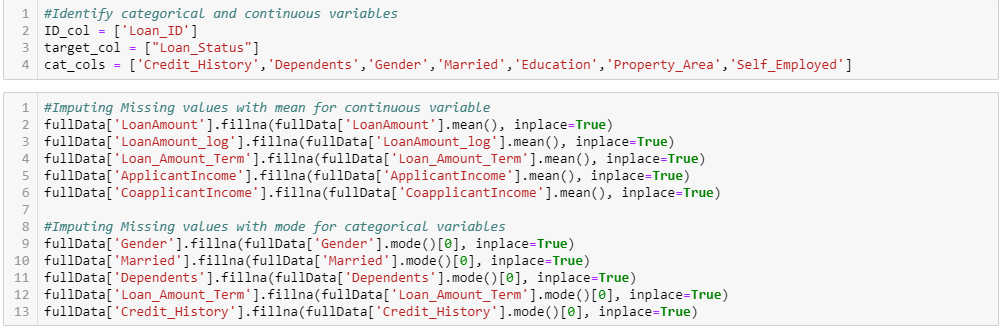
****

1. **Data Preparation And Model Building: -**

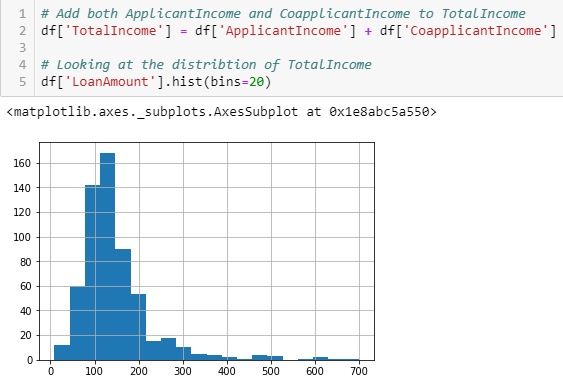
****

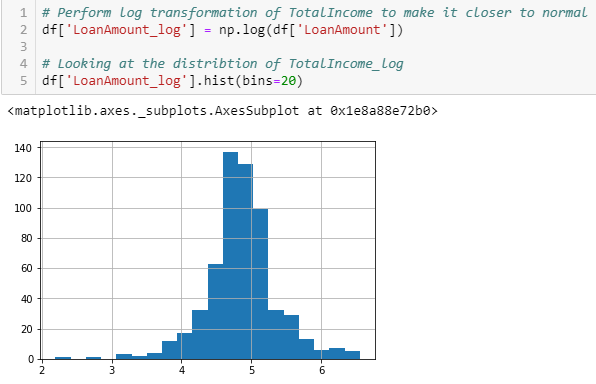
****

****

****

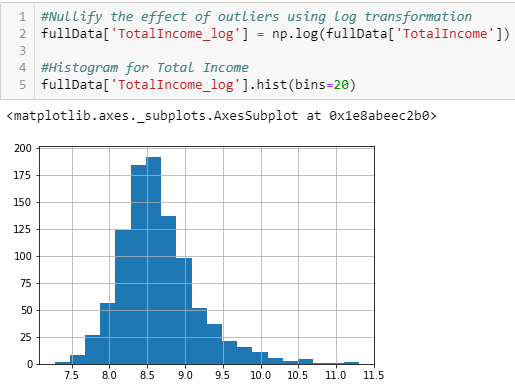
1. **Nullify the Effect of Outliers: -**
   1. **Loan Amount: -**

****

****

* 1. **Total Income: -**

****

****

1. **Data Preparation And Model Building: -**

****

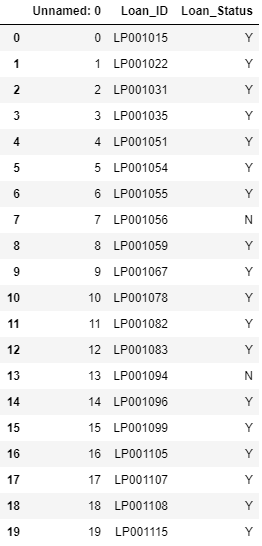
1. **Classification Model: -**

****

1. **Logistic Regression Model and Output: -**

****

1. **Output: -**

****

1. **Conclusion And Observations: -**
   1. 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.
2. **Analysing and interpreting the results: -**
   1. Test the result on different samples: -
      * Training data
      * Testing dataset
   2. Accuracy: - **80.945%**
   3. Cross – Validation Score: - **80.945%**