Project Report

On

Loan Application Status Prediction

Submitted By

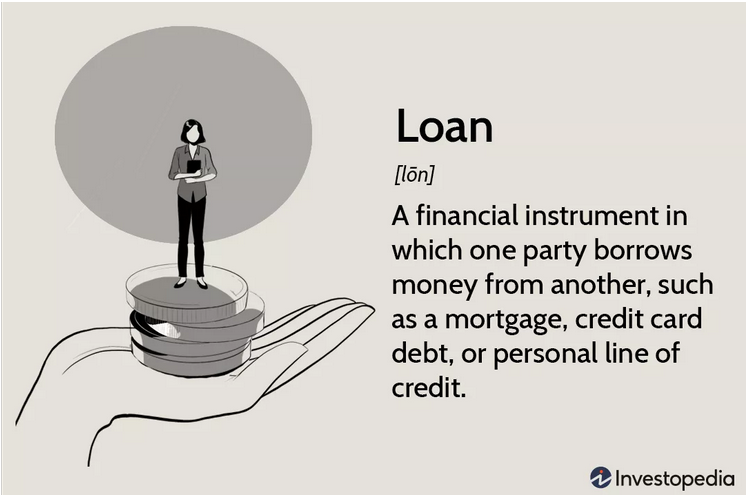
Mrs. Swati Amit Motugade

Introduction

Loan:



A loan is a sum of money that an individual or company borrows from a lender. It is a form of debt incurred by an individual or other entity. The lender – usually a corporation, financial institutions or banks advances a sum of money to the borrower. In return, the borrower agrees to a certain set of terms including any finance charges, interest, repayment date and other conditions.



Types of Loans:

The loans are classified into two types based on the purpose they are used for.

* Secured Loans
* Unsecured Loans

1. Secured Loans :

Secured loans are the loans which require the borrower to pledge an asset or security to avail the loan. Home loans and Vehicle loans are the most common examples of secured loans where the borrower will be required to pledge the vehicle or house to be purchased as collateral, which then become a secured debt. In case the borrower defaults on their loan repayment, the lender has full right to take possession of the collateral/secured debt.

A secured loan is one of the best and assured sources of obtaining a high volume of funds.

Following are the secured loans:

* Home Loan
* Loan Against Property
* Loan Against Insurance Policies
* Gold Loans
* Loan Against Mutual Funds and Shares
* Loan Against Fixed Deposit

2. Unsecured Loans :

Unsecured loans are the loans that don’t require collateral. They are also referred to as Signature loans because a signature is all that’s needed if you meet the lenders borrowing requirements. Because lenders take on more risk when loans are not backed by collateral, they might charge higher interest rates and require good or excellent credit.

If a borrower stops making payments and defaults on the unsecured loans, there is no collateral for the lender to take to recover the outstanding debt.

Following are some Unsecured loans:

* Personal Loan
* Short-term Business Loan
* Flexi Loan
* Education Loan

Requirements



The things that are to be considered before approving your loan application are

1. Purpose of loan

Most of the lenders wants to know the purpose of your loan.

2. Credit History

While approving your loan application lender will conduct a credit check. They will do this to determine your personal and business credit score. Before applying for a bank loan, make sure that you know both scores.

3. Personal Information

Even though you will be borrowing money for your business, some personal information could affect your ability to qualify. In addition to the personal credit score lenders usually also requests the following personal information in loan application:

* Addresses
* Criminal Record
* Information on your education
* Tax returns
* Financial Statements
* Assets
* Personal Loan balances

4. Financial Statements

In addition to personal information, you will also need to submit your financial statements. The number of statements will vary depending on the lender you are applying to and their requirements. Most of the lenders will require the following documents:

* Balance sheet
* Accounts receivables
* Profit and Statements
* Cash Flow statements
* Income Statements
* Other financial projections

5. Collateral

Even if your business or personal credit history falls below lender loan requirements, you could still receive financing by submitting collateral. Lender define collateral as business or personal property that you put up to guarantee the repayment of the loan. The lender will match collateral with the value of the loan you want to obtain.

6. Income Sources

The lender will consider through out the income sources before approving your loan application. If you are self-employed the lender will check for profit and loss of your business. If you are an employee the lender requests you for the salary sleep to make sure about the loan repayment.

7. Personal Guarantee

If you have applied for a personal loan, the lender may have asked you to sigh a personal guarantee. In some cases, you may need to sign this document to secure a small business loan from a bank. This document is a written promise stating that you will repay your loan using your personal assets if you can’t pay it’s loan balance.

Machine Learning for Prediction

Prediction in Machine Learning refers to the output of an algorithm after it has been trained on a historical dataset and applied to new data when forecasting the likelihood of a particular outcome.



Machine Learning for Loan Status Prediction

Banks are making major part of profits through loans. Though lot of people are applying for loans, it’s hard to select the genuine applicant who will repay the loan. While doing the process manually, lot of misconceptions may happen to select the genuine applicant. Therefore, there is a need to develop loan prediction system using machine learning, so the system automatically selects the eligible candidates. This is helpful to both bank staff and applicant. The time period for the sanction of loan will be drastically reduced.

To deal with the problem, the automatic loan prediction using machine learning techniques is developed. Here we will train the machine with the provided previous dataset, so machine can analyse and understand the process. Then machine will check for eligible applicants and give us result.

Advantages of Using Machine Learning Techniques

1. Time period for loan sanctioning will be reduced.
2. Whole process will be automated, so human error will be avoided.
3. Eligible applicant will get loan approval without any delay.

Problem Statement

This dataset includes details of applicants who have applied for loan. The dataset includes details like credit history, loan amount, their income, dependents etc.

We have to build a model that can predict whether the loan of the applicant will be approved or not on the basis of the details provided in the dataset.

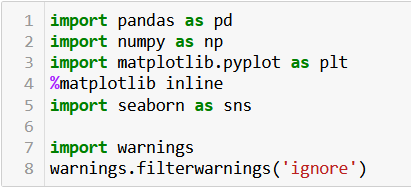
About the Dataset

|  |  |
| --- | --- |
| Variable | Description |
| Loan\_id | Unique Loan application Id |
| Gender | Male/Female |
| Married | Applicant Married (Yes/No) |
| Dependents | Number of Dependents |
| Education | Applicant’s Education (Graduate/Not Graduate) |
| Self\_Employed | Self Employed (Yes/No) |
| Applicant Income | Applicant’s income (in Thousands) |
| Co-applicant Income | Co-applicant’s income (in Thousands) |
| Loan\_Amount | Loan amount required (In Thousands) |
| Loan\_Amount\_Term | Tenure of Loan (in Months) |
| Credit History | Credit history of Applicant |
| Property\_Area | Property area of applicant (Rural/Urban/Semiurban) |
| Loan\_Status | Loan Approved (Y/N) |

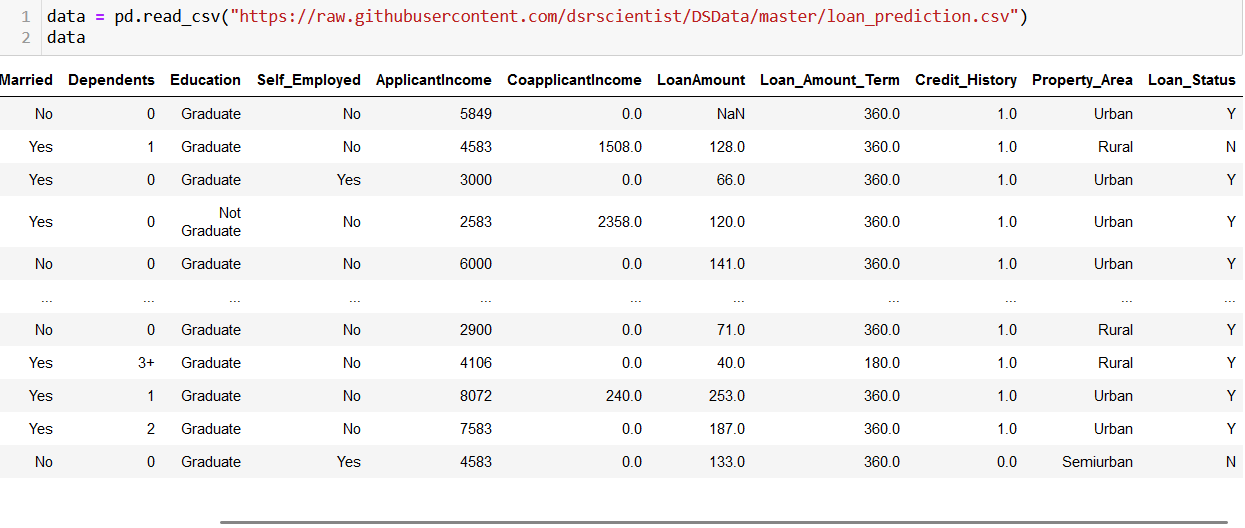
Importing necessary libraries

Let’s start with importing necessary libraries and loading the dataset from the link given below

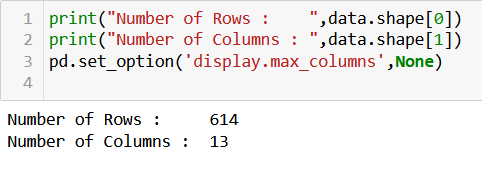
https://github.com/dsrscientist/DSData/blob/master/loan\_prediction.csv



Dataset Loading

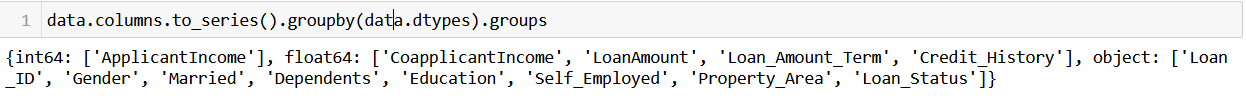


Check for Shape of Dataset



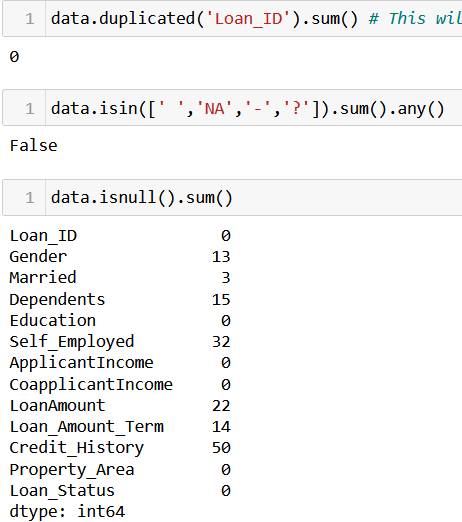
Provided dataset contains 614 rows and 13 columns.

Check for various datatypes



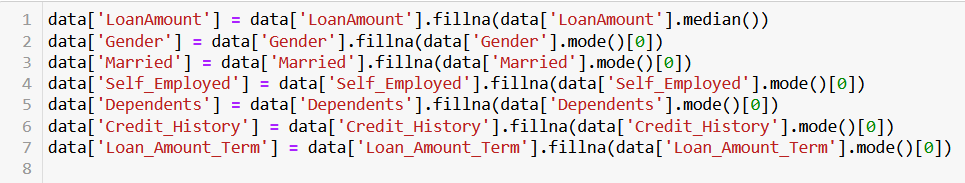
Here we have 8 variables with object datatype and rest are of int64 and float64 datatype.

Data Integrity Check



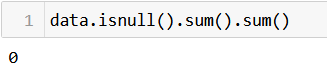
Here we can see that there are no duplicate entries and no any [ ‘ ‘, ‘ -’, ’NA’, ’?’, ‘,’] But our dataset contains missing values in some features.

Filling Null values

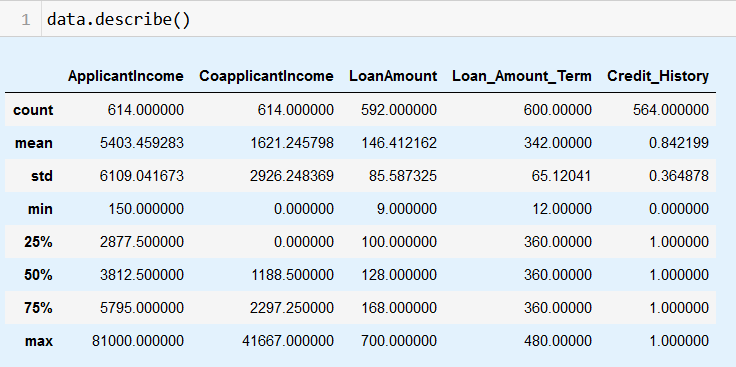


Since LoanAmount is a numerical feature and it contains outliers so we have imputed its null values with median and the remaining features with missing values are of categorical type so we used mode for imputation.

Checking for null value imputation



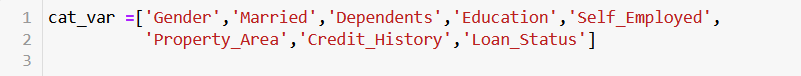
Check for Descriptive Statistics of Dataset



Observations

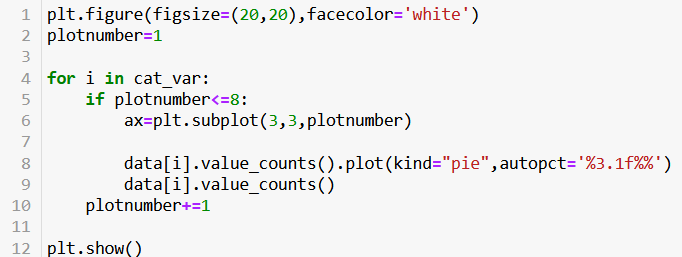
1. By observing the mean and 50% values we can say that the features are not normally distributed.
2. Also, there are outliers present in our dataset.
3. Applicant’s minimum income is 150 and maximum income is 81000.
4. Co-applicant's minimum income is 0 and maximum income is 41667.
5. Minimum loan amount is 9 and maximum is 700.
6. Minimum Tenure for loan is 12 months and maximum of this is 480 months.

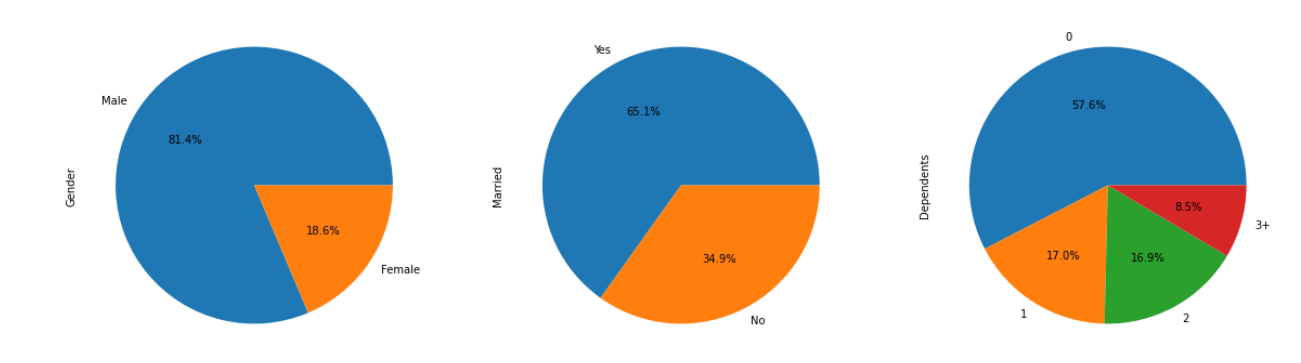
Separate Categorical and Numerical Features

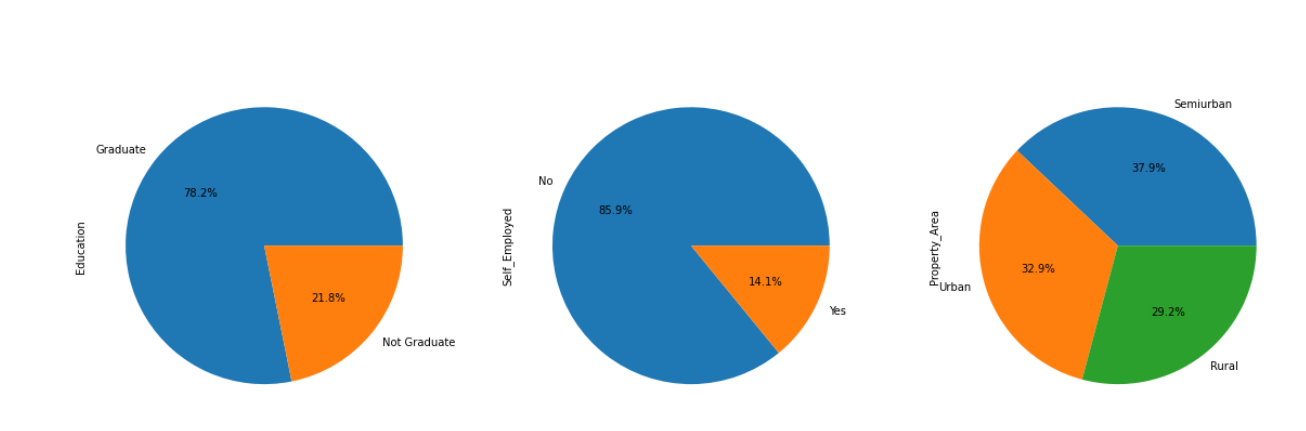


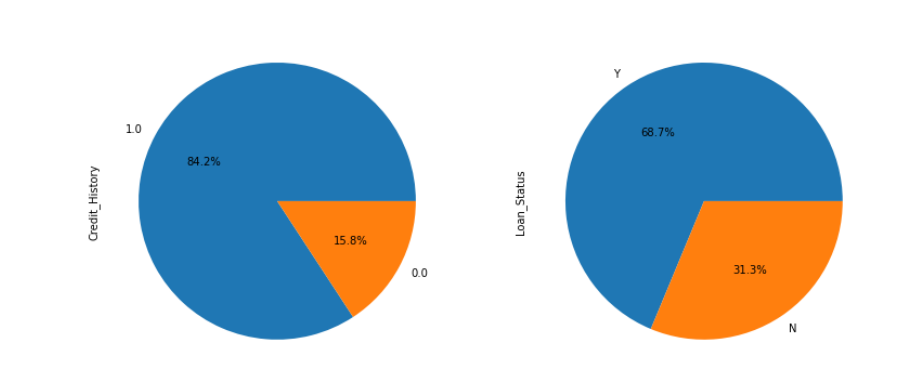


EDA (Categorical Features)





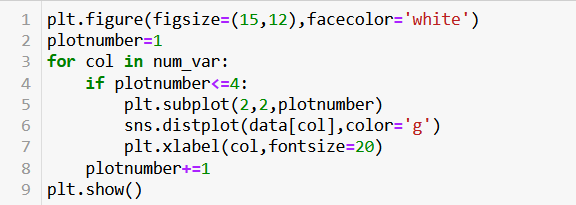


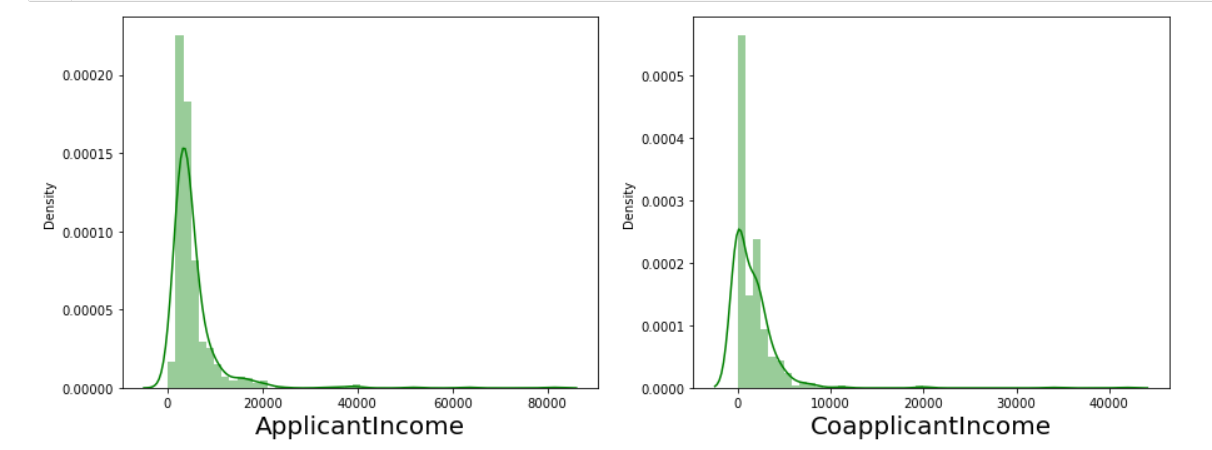


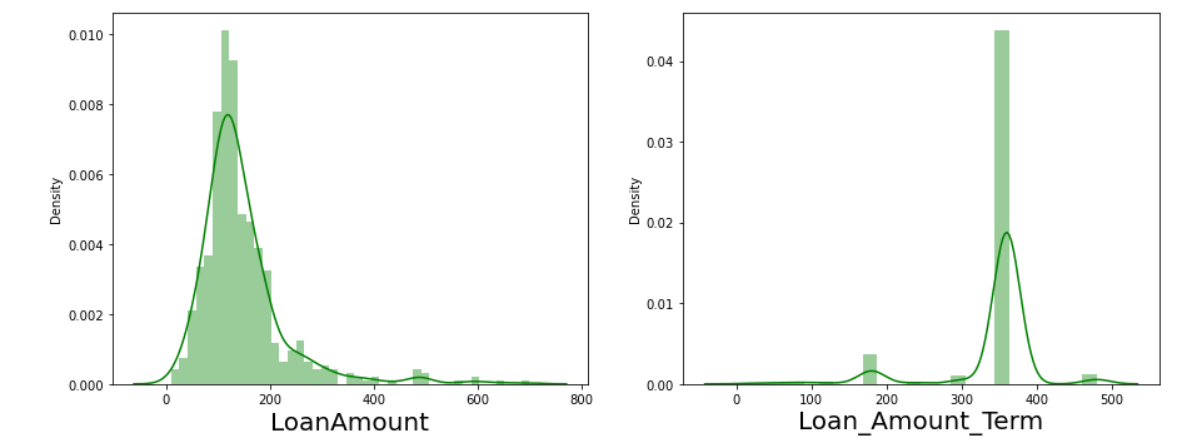
Observations

1. More than 80% loan applications are coming from Males.
2. Out of all the applicants 65% are married.
3. 57.6% applicants have 0 dependents, 17% applicants have 1 dependent, 17% applicants have 2 dependents and 8.5% applicants have 3+ dependents.
4. 78.2% applicants are Graduates where remaining 21.8% are Non graduates.
5. Only 14.1% applicants are self-employed.
6. Majority, 37.9% applicants are from Semi-Urban area, 32.9% are from Urban area and 29.2% applicants are from Rural area.
7. 68.7% application are approved and 31.3% are rejected.
8. 84% applicants have credit history 1.

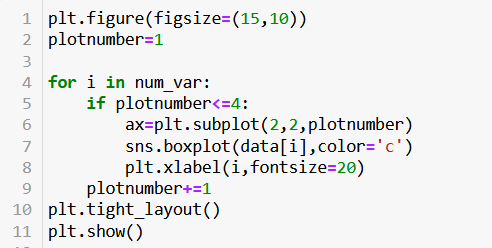
EDA (Numerical Features)

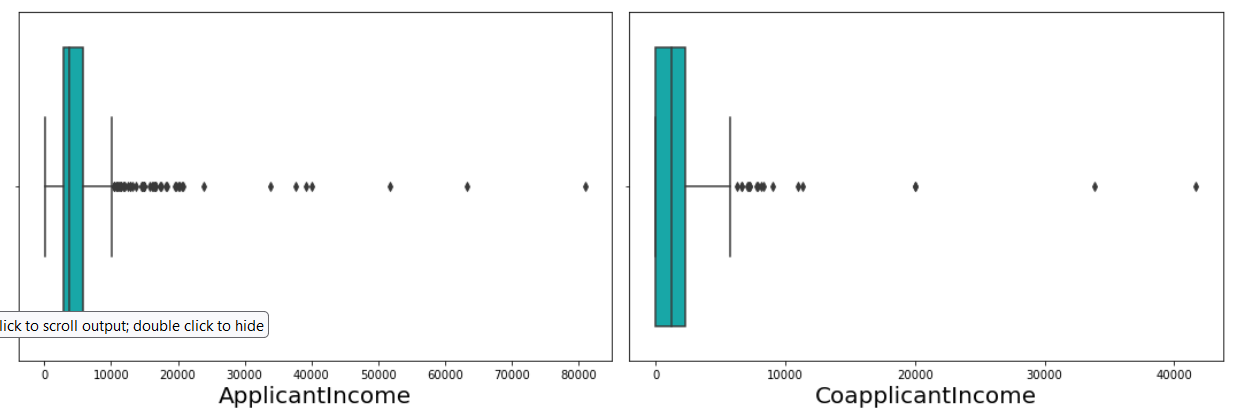


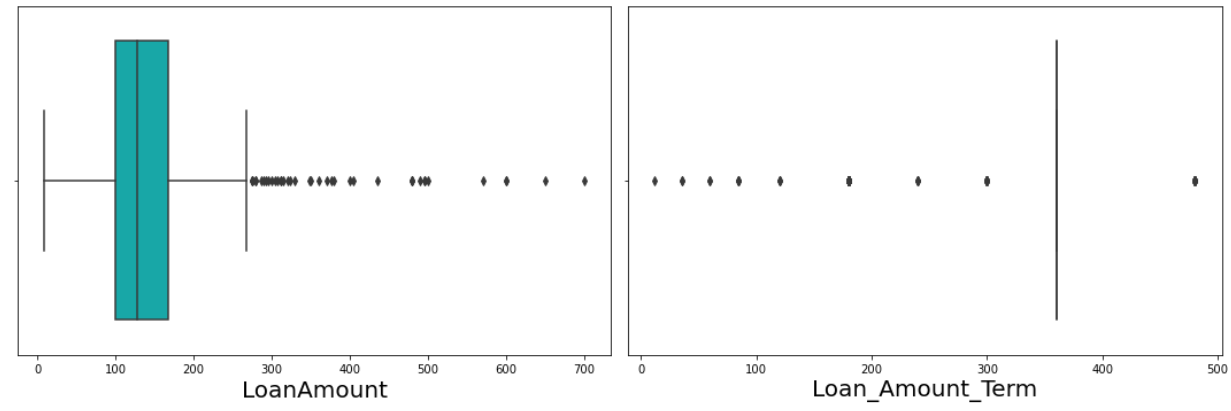




Outliers Detection







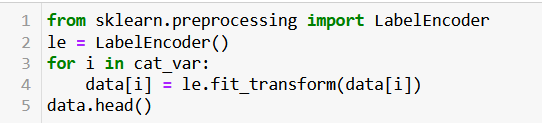
Observations

From the above boxplots we observed that all the numerical features contains outliers and we have to remove these outliers.

Before removing the outliers, we have to perform label encoding for categorical features.

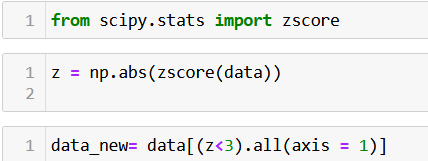
Label Encoding

We have to convert the encode the categorical features into numerical one. For this we will use the label encoding method.

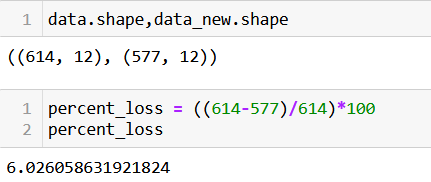


Outliers Removal

Here we will use z-score method to remove outliers from the features.

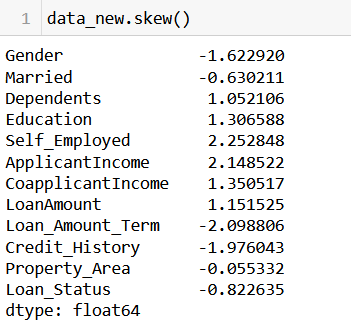


Check for Percentage of Data loss



Here, we lost 6% data while removing the outliers. The data loss percentage is not very high so that we can afford this data loss.

Skewness



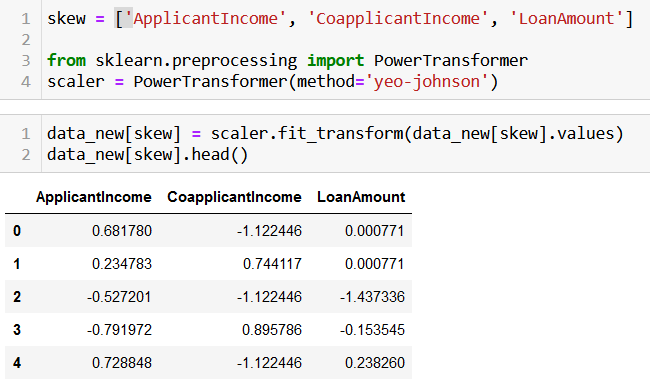
The features with skewness are:

* Gender
* Dependents
* Education
* Self\_Employed
* ApplicantIncome
* CoapplicantIncome
* LoanAmount
* Loan\_Amount\_Term
* Credit\_History
* Loan\_Status

Out of the above features 'Gender', 'Education', 'Self\_Employed', 'Credit\_History', 'Dependents' are categorical features and 'Loan\_Status' is target variable so we will not remove skewness from them.

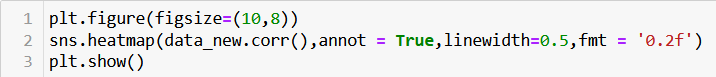
We will consider only the features 'ApplicantIncome', 'CoapplicantIncome', 'LoanAmount' for removing skewness.

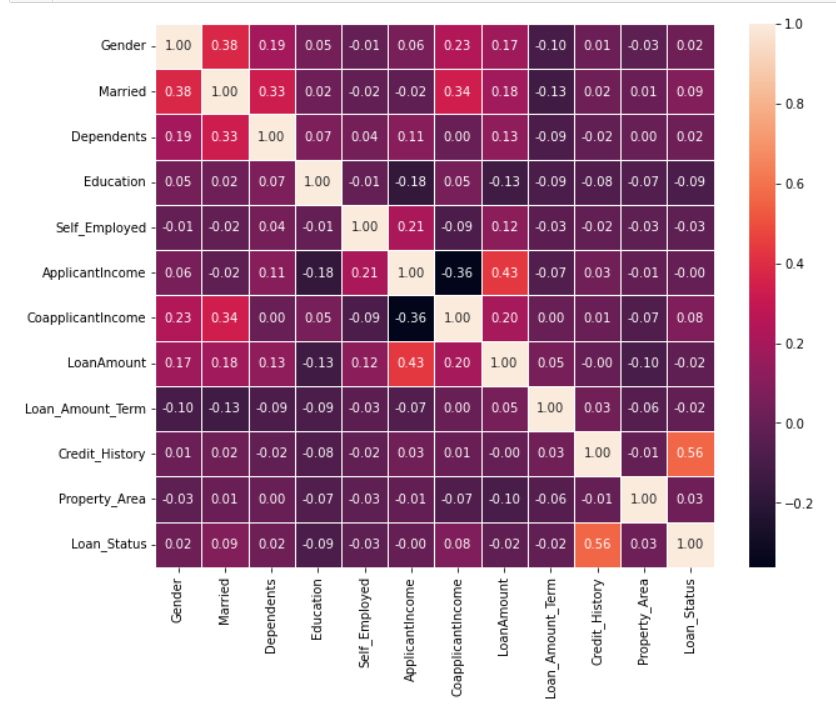
To remove skewness from the above-mentioned variables we will use yeo-Johnson transformation method.

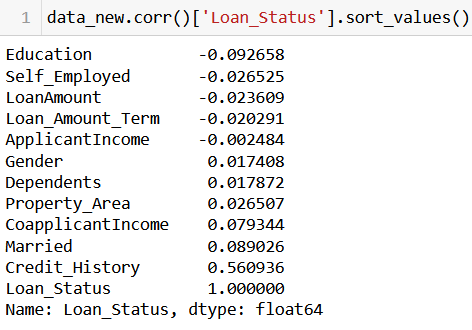


Correlation Heatmap

Here we are going to plot the correlation heatmap to get the overall idea about correlation between all features in one plot. Here we will also get the idea about the multicollinearity.







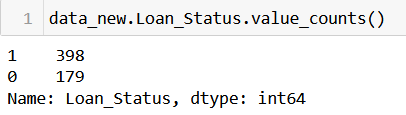
From the above correlation summary, we can see that Credit\_History has highest i.e., 56% correlation with target variable Loan\_Status.

Also, by observing the heatmap we can say that there is no multicollinearity present between the features.

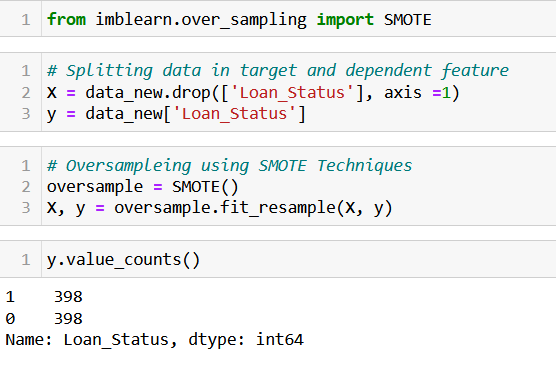
Separate Features and Target Variable



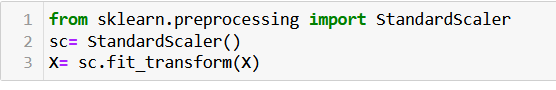
Balancing the Target Variable



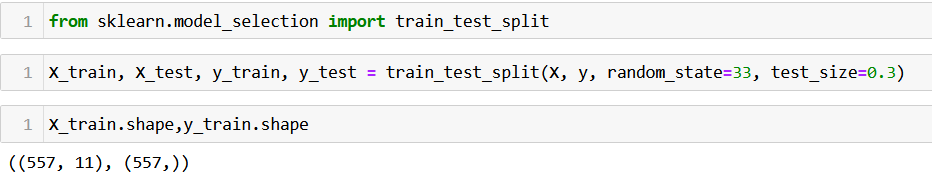
From the above value counts summary of target variable, it is clear that our target variable is imbalanced and we have to balance it. We will balance it by using SMOTE oversampling technique.



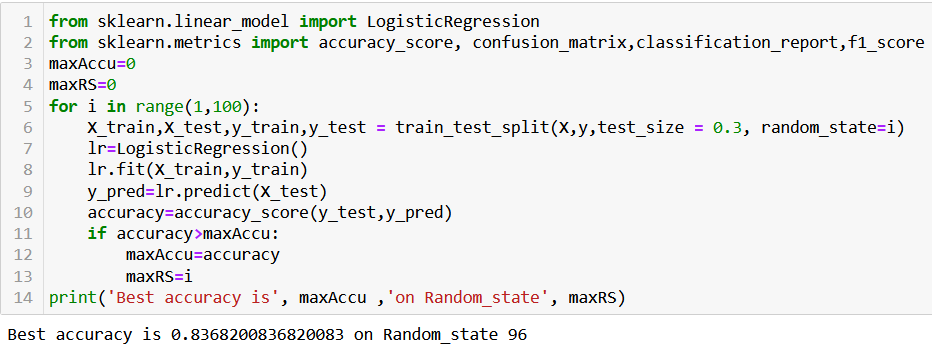
Feature Scaling



Split data into training and test data

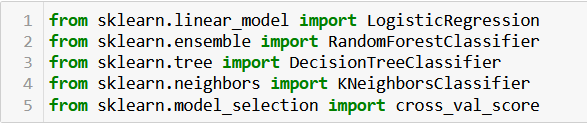


Finding best random state

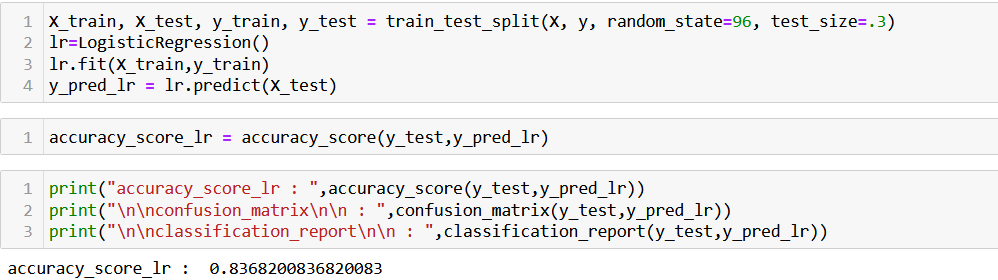


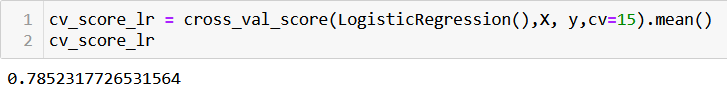
Here we got 96 as best random state and accuracy score at random state 96 is 0.8368200.

Model Building

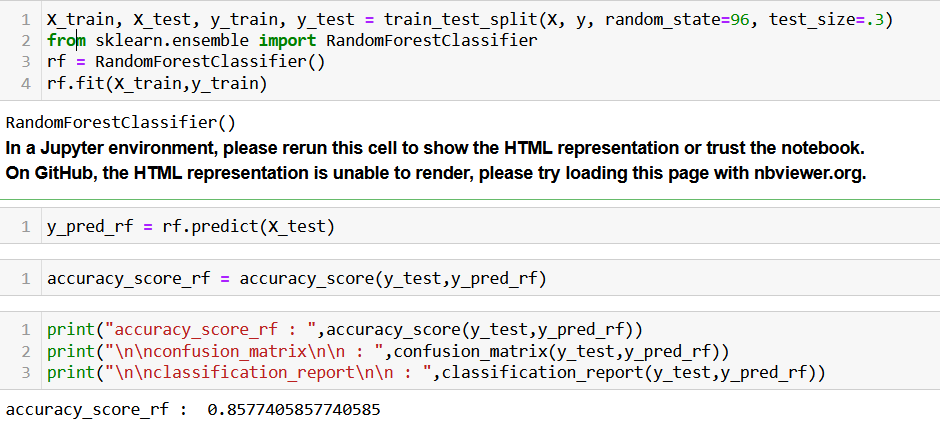


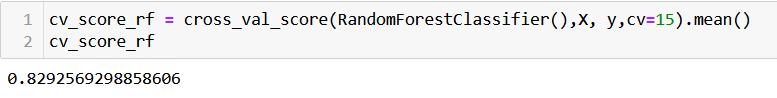
1. Logistic Regression Model



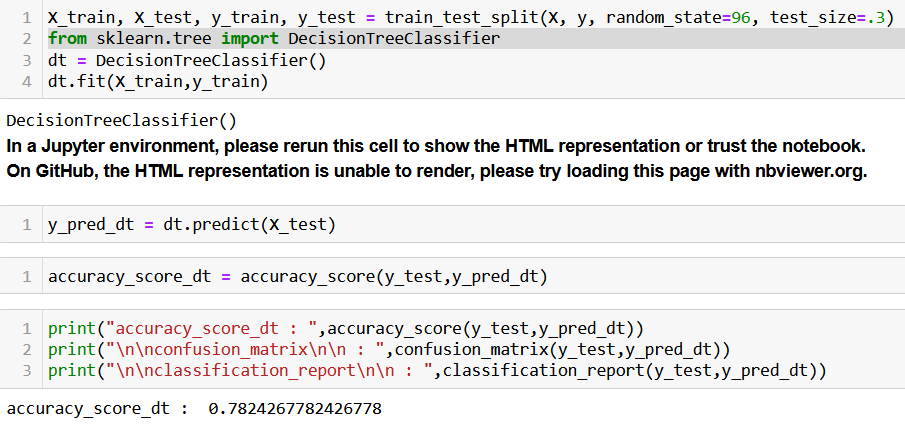


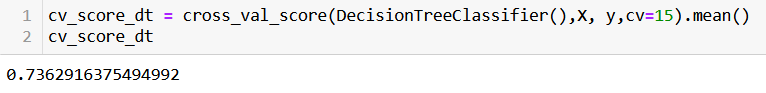
2. Random Forest Classifier



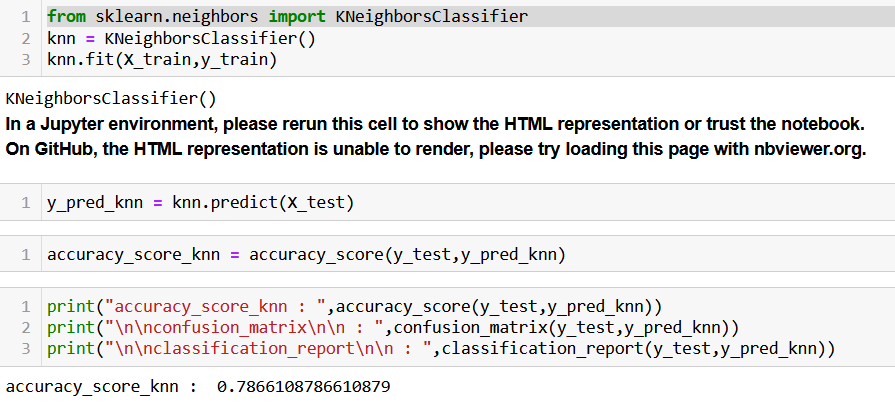


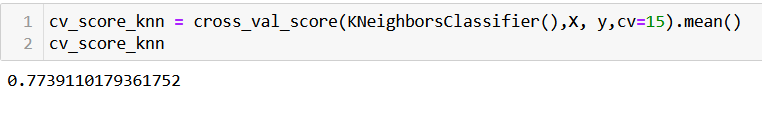
3. Decision Tree Classifier



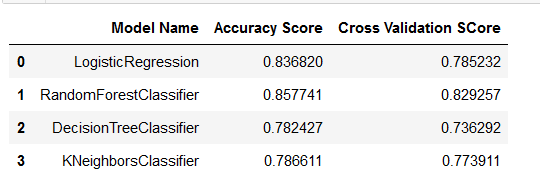


4. K-Neighbours Classifier





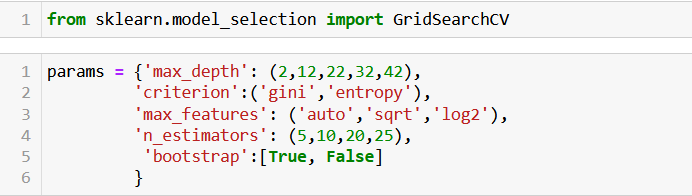
Accuracy Score and Cross Validation Score for used algorithms

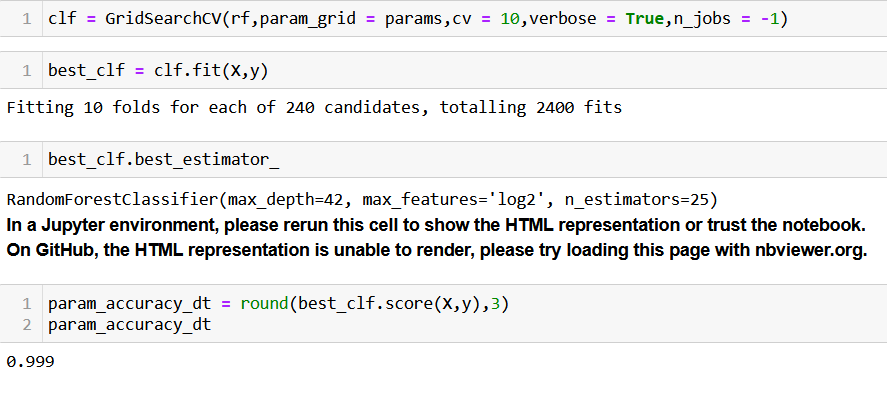


Since the model Random Forest Classifier gives us best accuracy score we will choose Random Forest Classifier as a final model.

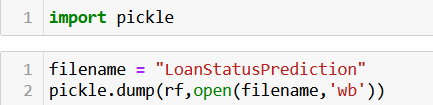
Hyperparameter Optimization

Since we have selected Random Forest Classifier as a final model, we will perform Hyperparameter optimization for this model.

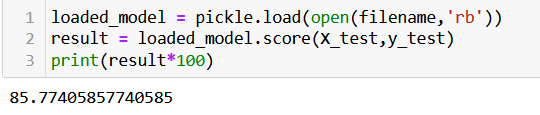




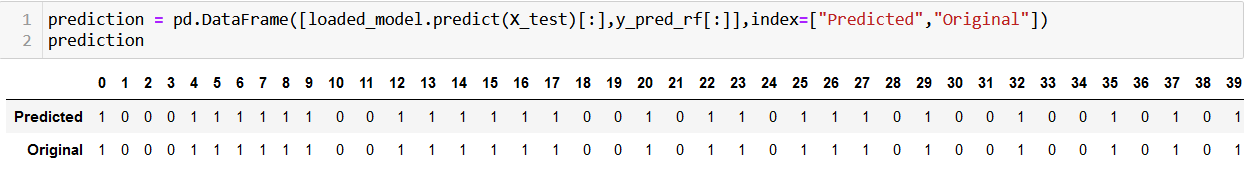
Model Saving



Conclusion



Prediction on Test data



Conclusion based on EDA:

* The number of approved loans is more for Male applicants than that of female.
* The married applicants are more likely to get loan approval.
* The applicants with 0 dependents are more likely to get loan approval.
* The graduate applicants are most probable for getting sanctioned loan.
* The probability of getting loan approval is more for the applicants who are not self-employed.
* The applicants whose property area is semi-urban are more likely to get loan approval.
* The applicants with credit history 1 are most probable to get loan approval.
* The random forest classifier algorithm gives the best accuracy score. Hence we have selected this as final model.