**LOAN APPROVAL PREDICTIONS**

**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.

You can download CSV File from the below link:

<https://github.com/dsrscientist/DSData/blob/master/loan_prediction.csv>

**ABSTRACT**

Loan approval is a very crucial process for banking organizations. The core activity of banks is lending of funds. The ultimate profit directly comes from the loan’s interest. Companies grant loan after a long process of verification and validation. In this notebook, I'll build a model to predict if an applicant get approval for loan or not. Machine Learning techniques are very useful in predicting outcomes for large amount of data. In this paper three machine learning algorithms, Logistic Regression (LR), Support Vector Classifier (SVC) and Random Forest (RF) are applied to predict the loan approval of customers. The experimental results conclude that the accuracy of Random Forest Classifier machine learning algorithm is better as compared other machine learning approaches.

**INTRODUCTION**

Distribution of the loans is the core business part of almost every bank. The main portion the bank’s assets are directly coming from the profit earned from the loans distributed by the banks. The prime objective in banking environment is to invest their assets in safe hands where it is. Today many banks/financial companies approve loan after proper verification and validation but still there is no surety whether the chosen applicant is the deserving right applicant or not.

Our task is to predict whether a customer get loan approval or not using various classification algorithms. Exploratory data analysis is done on the dataset to achieve insights and the pre-processing pipeline is done to get the data ready for the training.70% of the data is used for training purpose and 30% for the testing purposes. Three Classification models are trained and their performances are compared with various performance metrics like confusion metrics, accuracy score, cross validation score and the Receiver operating characteristic curve.

**LOADING DATASET**

The Loan Approval data set is of shape (614,13) i.e. It has 13 attributes and 614 rows.

The dataset provides 12 input variables and 1 target variable that are a mixture of ordinal, categorical and numerical data types. Following are the variables is our dataset:

1. Dependents: Number of dependents
2. Education\_Qualification: Graduate/ Under Graduate
3. Self\_Employed: Self-employed (Y/N)
4. Applicant\_Income: Income of Applicant
5. Co\_Applicant\_Income: Income of Co-applicant
6. Loan\_Amount: Loan amount in thousands
7. Loan\_Amount\_Term: Term of loan in months
8. Credit\_History: Credit history meets guidelines
9. Property\_Area: Urban/ Semi Urban/ Rural

**TARGET VARIABLE:**

1. Loan\_Status: Loan Approved(Y/N)

We have three kinds of data types:

**Object:** It means variables are categorical.

Following are the Categorical variables in our dataset: Loan\_ID, Gender, Married, Dependents, Education, Self\_Employed, Property\_Area, Loan\_Status.

**int64:** It represents the integer variables. "ApplicantIncome" is given in int64 format.

**float64:** It represents the variable that has some decimal values. They are also numerical. Following are the float64 variables in our dataset: CoapplicantIncome, LoanAmount, Loan\_Amount\_Term, Credit\_History.

**Exploratory Data Analysis**

Now, we will do exploratory data analysis to get the insight about the data and how target variable depends on various attributes.

First, we are analyzing our target variable i.e. Loan\_Status

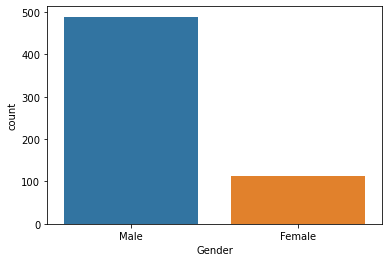
Figure.1

This figure shows the analysis of our target variable.

Y 422

N 192

This means that the count of customers whose loan approved is 422 out of 614.

**UNIVARIATE ANALYSIS OF INDEPENDENT VARIABLES:**

1. **GENDER**:

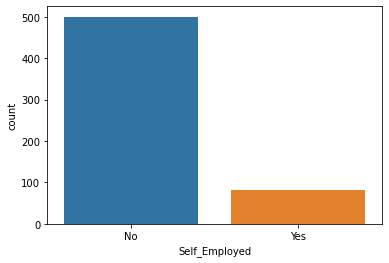
This figure shows that around 490 are males in our dataset and remaining are females.

Figure.2

1. **MARRIED**:

This figure shows that most of customers are married.

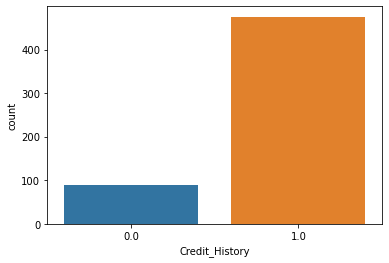
Figure.3



1. **SELF\_EMPLOYED:**

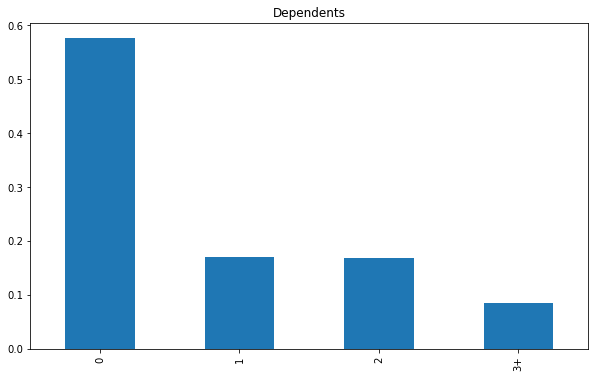
This figure shows that most of customers are not self-employed.

Figure.4

1. **Credit History:**

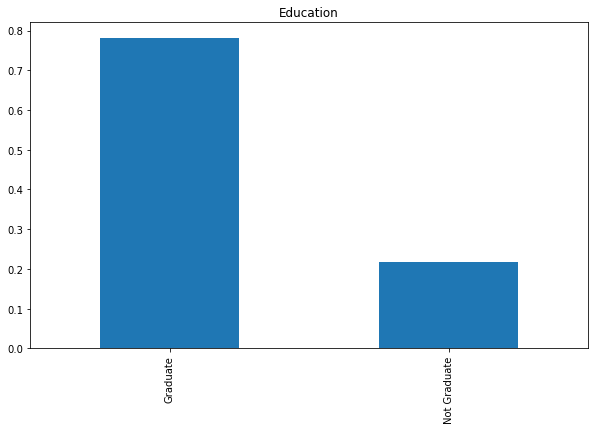
This figure shows that most of customers don’t have credit history.

Figure.5

1. **DEPENDENTS**:

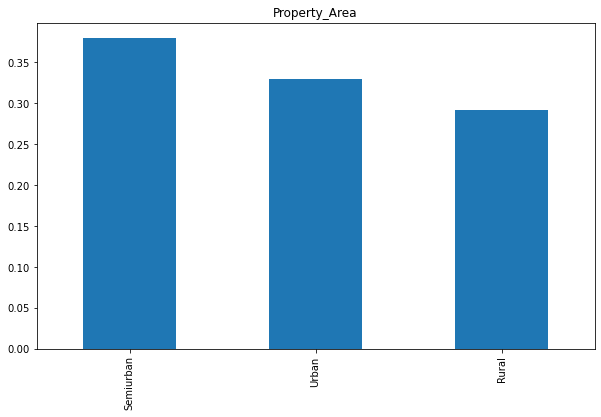
This figure shows that about 58% of customers don’t have dependents.

Figure.6

1. **EDUCATION**:

We can visualize from this figure that 78% of customers are graduated.

Figure.7

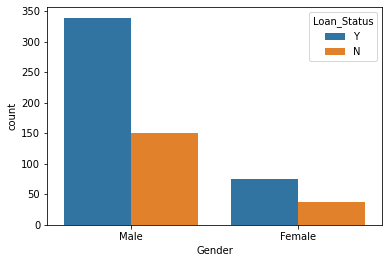
1. **PROPERTY AREA**:

Around 40% of customers are living in Semi-urban, 32% are living in urban and remainig 28% are living in rural area.

Figure.8

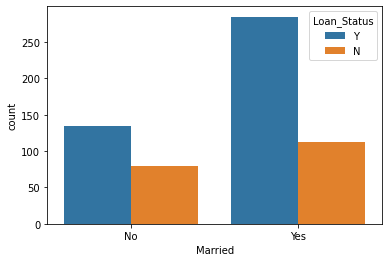
### Analysis of Categorical Independent Variable with Target variable:

### Analysis of “Gender” WRT “Loan Status”:

It is shown that 70% customers are males and around 330 are getting loan approval.

### Figure 9

### Analysis of “Gender” WRT “Loan Status”:



We earlier analysed that most of customers are married and we get to know that around 300 married customers are getting loan approval and 180 unmarried customers got approval for loan.

Figure. 10

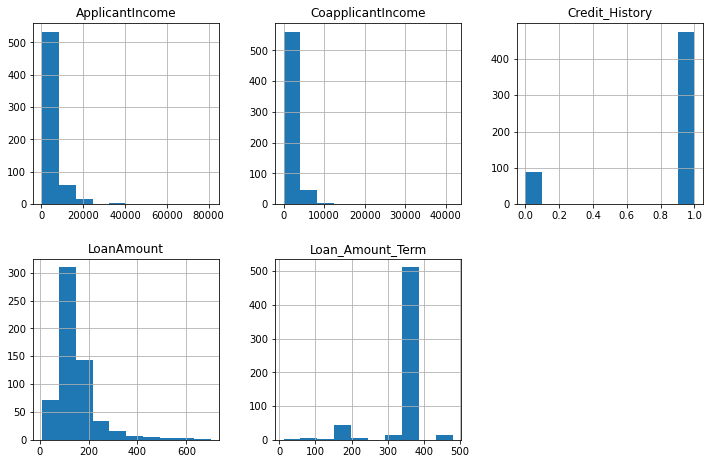
### Analysis of “Self-employed” WRT “Loan Status”:

### Most of customers are not self -employed and around 330 customers are getting approval for loan who are not self-employed.

Figure. 11

### 

**ANALYSIS OF CONTINOUS VARAIBLES:**

 Figure.12 shows the attributes which are continuous in nature. Applicant Income, co-applicant income, credit history, loan amount, loan amount term are given in numeric form.

We can clearly analyse that Applicant Income, co-applicant, loan amount is rightly skewed.

### Loan amount term is normally distributed. Figure.12

### DATA PREPROCESSING

**Data preprocessing** is very essential step in any **data** mining process. It directly impacts the predictions of the model. If data is unclean, have missing vales, missing attributes or contains outliers, if skewness is present, then all these factors degrade the quality of our results and our predictions will be biased.

First, we will check for missing or NaN values through heat-map:

There are missing values present in Gender, Married, Dependents, Self\_Employed, LoanAmount, Loan\_Amount\_Term, and Credit\_History features.We will treat the missing values one by one.

We will consider these methods to fill these missing values:

1. For numerical variables: using mean or median
2. For categorical variables: using mode.

### There are very few missing values in Gender, Married, Dependents, Credit\_History, and Self\_Employed features so we treated these using the mode as these are categorical features.

### Now we have treated the LoanAmount variable. As it is a numerical variable, we can use mean or median to treat the missing values. We have used the median to fill the null values as earlier we saw that the loan amount has outliers so treating missing values with the mean will not be the proper approach as it will be highly affected by the presence of outliers in dataset. Figure.13

## **Drop irrelevant columns:**

### We dropped attribute Loan ID because high number of unique values present. Irrelevant to our predictions.

**LABEL ENCODING:**

Label Encoding means to convert all the categorical values into numerical values so as to convert it into the machine-readable form.

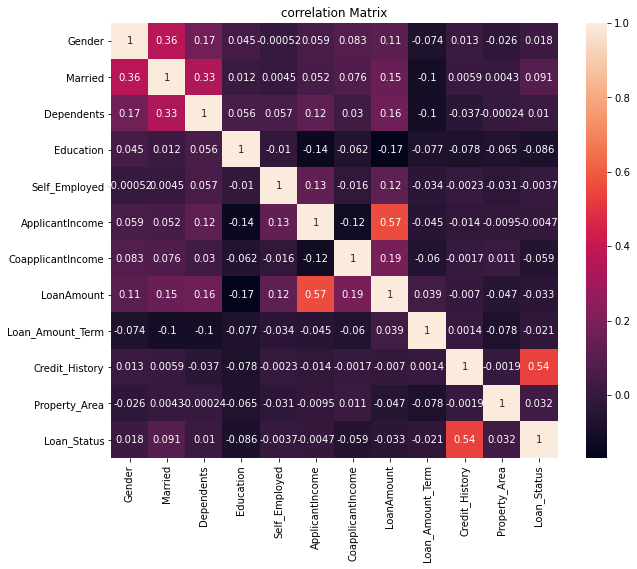
First, we have imported the label encoder and then a looped Label Encoder is used to convert all the labels in every column to the numeric form for the training and testing purpose. The attributes that has categorical values are ‘Loan\_ID', 'Gender', 'Married', 'Dependents', 'Education', 'Self\_Employed', 'Property\_Area', 'Loan\_Status'. The target variables i.e. Loan Status had only two categories Y & N. That’s why our target variable has only two values 1 and 0.

**Correlation matrix heatmap:**

Checking correlations is very important to analyse data. A heatmap has been plotted to check the correlation between the attributes, if there is positive or negative relationship. This is one of the methods to decide which attributes affect the target variable the most.

The figure.14 is the heatmap of the correlation matrix of the attributes.

.

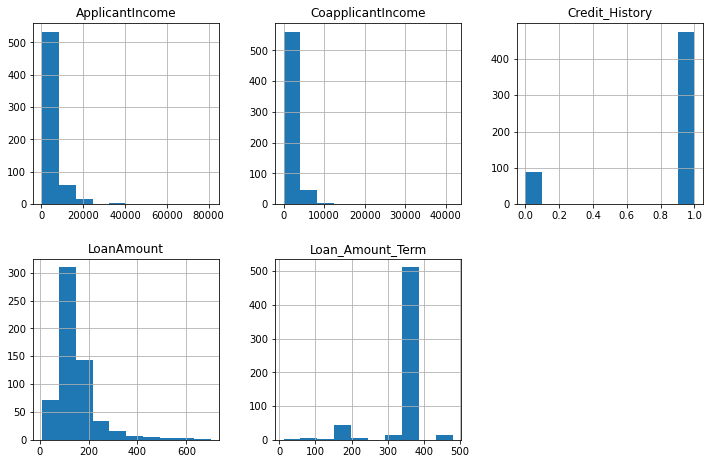
. It can be inferred that Credit\_History is positively correlated with Loan\_status. There is no relationship between Loan\_status & Dependents, Loan\_status & Self\_employed, Loan\_status & Applicant Income. Education & Loan\_status shows negative relationship.   
 Figutre.14

**CHECKING OUTLOIERS:**

An outlier means an observation that falls outside the overall pattern or we can say an abnormal distance from other values in a random sample from a population.

We have outliers present in various attributes named ApplicantIncome, CoapplicantIncome, LoanAmount, Loan\_Amount\_Term. From scipy.stats we imported Z-score and drop all the rows in which threshold value is greater than 3. But by dropping these rows we lost our data and the results came from our predictions will be biased. Therefore, we will not drop these outlier values as these values are important for our predictions.

**TREATING SKEWNESS:**



If the skewness is between -0.5 and 0.5 then the data is fairly symmetrical and represent normal distribution. If the skewness is between 0.5 and 1 or -1 and -0.5 then the data is moderately skewed. If the skewness is less than -1 or greater than 1then the data is highly skewed.

As we earlier analysed that skewness is present in ApplicantIncome, CoapplicantIncome, LoanAmount, Loan\_Amount\_Term. Therefore, we will treat this with power transformation.

**SPLITTING DATASET**

The shape of the dataset after dropping of the irrelevant columns is (614, 12). We split the dataset where 70% is used for training the model and 30% for testing the model. Hence out of 614 data entries, 430 are used for training and 184 are used for testing the model.

**FINDING BEST RANDOM STATE**

Our model best performs at random state 8 and we are achieving 0.84 accuracy score.

Four Classification Algorithms are used.

1. Logistic Regression
2. RandomForestClassifer
3. SupportVectorClassifier

**Logistic Regression**

We are achieving 84% accuracy with Logistic Regression.

**SupportVectorClassifier**

We are achieving 85% accuracy with SVC

**RandomForestClassifer**

We are achieving 82% accuracy with RandomForestClassifier.

**Cross Validation Score:**

Imported cross validation score to check the over-fitting and under fitting in our predictions.

**from** **sklearn.model\_selection** **import** cross\_val\_score

#### We proceed with RandomForestClassifier as it is giving highest accuracy\_score and there is minimum difference between accuracy\_score & cross\_validation\_score

We get best accuracy\_score from RandomForestClassifier.

**TUNNING WITH BEST PARAMETERS:**

Imported GridSearchCV from sklearn.model\_selection and find out the best parameters of RandomForestClassifier which performed best on our model.

Following are the best parameters for our model:

{'criterion': 'gini', 'max\_features': 'log2'}

**ROC (Receiver Operating Characteristic) Curve**

Receiver Operating Characteristic curve or ROC curve represents a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the TPR against the FPR at various threshold settings. The true-positive rate (TPR) is also known as recall, sensitivity, or probability of detection in machine learning. The false-positive rate (FPR) is also known as the probability of false alarm or fall-out.

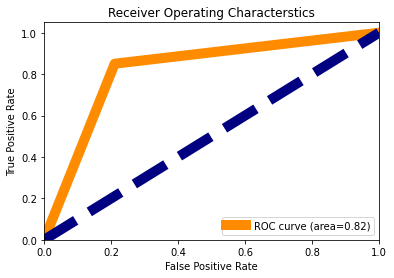


Figure.15:

The area under the curve is 82%.

**SAVE THE MODEL**

We finally save our best model by importing pickle. The use of pickle is widespread as they allow us to easily transfer data from one server or system to another and then store it in a file or database.

**CONCLUSION**

We successfully predict whether a customer get loan approval or not using various classification algorithms. Exploratory data analysis is done on the dataset to achieve insights and the pre-processing pipeline is done to get the data ready for the training.70% of the data is used for training purpose and 30% for the testing purposes. Three Classification models are trained and their performances are compared with various performance metrics like confusion metrics, accuracy score, cross validation score and the Receiver operating characteristic curve. The RandomForestClassifier comes out to be the best performing algorithm above all other models with an accuracy of 82.9% and over all generalizing well.

**DOWNLOAD JUPTYER NOTEBOOK**

Click on the below link to find my juypter notebook in GitHub:

<https://github.com/riturani2403/Data-Trained-practice-projects/blob/main/LOan_prediction.ipynb>