LOAN PREDICTION USIN MACHINE LEARNING ALGORITHMS

Submitted By,

BINI VARGHESE

(NOV OL Batch)

LUMINAR TECHNOLAB

I. Introduction

1.1 Machine Learning

Machine learning is the subarea of artificial intelligence (AI). It enables a machine to automatically learn from data,improve performances from experience and predict things without being explicitly programmed. It is a vast area and it is quite beyond the scope of the tutorial to cover all its features. Machine learning combines data with statistical tools to predict an output. This output is then used by corporate to makes actionable insights. Machine learning is closely related to data mining and bayesian predictive modeling. The machine receives data as input, use an algorithm to formulate answers.

1.2 Types of Machine Learning

At a broad level machine learning can be classified into three types.

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Supervised Learning

This algorithm consist of a target / outcome variable (or dependent variable) which is to be predicted from a given set of predictors (independent variables). Using these set of variables, generate a function that map inputs to desired outputs. The training process continues until the model achieves a desired level of accuracy on the training data. Examples of Supervised Learning: Regression, Decision Tree, Random Forest, KNN, Logistic Regression etc.

Unsupervised Learning

In this algorithm, we do not have any target or outcome variable to predict / estimate. It is used for clustering population in different groups, which is widely used for segmenting customers in different groups for specific intervention. Examples of Unsupervised Learning: Apriori algorithm, K-means.

Reinforcement Learning

Using this algorithm, the machine is trained to make specific decisions. It works this way: the machine is exposed to an environment where it trains itself continually using trial and error. This machine learns from past experience and tries to

capture the best possible knowledge to make accurate business decisions. Example of Reinforcement Learning: Markov Decision Process.

II. Introduction to Project

Loan Prediction is very helpful for employee of banks as well as for the applicant also. The aim of this Project is to provide quick, immediate and easy way to choose the deserving applicants. Dream housing Finance Company deals in all loans. They have presence across all urban, semi urban and rural areas. Customer first apply for loan after that company or bank validates the customer eligibility for loan. Company or bank wants to automate the loan eligibility process (real time) based on customer details provided while filling application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and other. This project has taken the data of previous customers of various banks to whom on a set of parameters loan were approved. So the machine learning model is trained on that record to get accurate results. Our main objective of this project is to predict the safety of loan. To predict loan safety, the Logistic regression, Decision trees, Random Forest and XGBoost algorithms are used. First the data is cleaned so as to avoid the missing values in the data set.

III. Problem Statement

Dream Housing Finance company deals in all home loans. They have presence across all urban, semi-urban and rural areas. Customer first apply for home loan after that company validates the customer eligibility for loan. Company wants to automate the loan eligibility process (real time) based on customer detail provided while filling online application form.To automate this process, they have given a problem to identify the customer segments, those are eligible for loan amount so that they can specifically target these customers.

Libraries for Data Analysis

The models are implemented using Python 3.7 with listed libraries:

- **Pandas**

Pandas is a Python package to work with structured and time series data.The data from various file formats such as csv, json, sql etc can be imported using Pandas. It is a powerful open source tool used for data analysis and data manipulation operations such as data cleaning, merging, selecting as well wrangling.

- **Seaborn**

Seaborn is a python library for building graphs to visualise data. It provides integration with pandas. This open source tool helps in defining the data by mapping the data on the informative and interactive plots. Each element of the plots gives meaningful information about the data.

- **Sklearn**

This python library is helpful for building machine learning and statistical models such as clustering, classification, regression etc. Though it can be used for reading, manipulating and summarizing the data as well, better libraries are there to perform these functions.

IV. Proposed Model

This system predict whether the loan is approve or reject . This System refers the following things or ways:

a. Data Collection and Understanding the data

b. Exploratory Data Analysis

c. Missing value Handling and Outlier Treatment

d. Modelling using Machine Learning Algorithms

e. Evaluating the mode

A. Data Collection and Understanding the data

Loan Dataset is very useful in our system for prediction of more accurate result. Using the loan Dataset the system will automatically predict which costumer's loan it should approve and which to reject. System will accept loan application form as an input. Justified format of application form should be given as an input to get processed.The dataset have 13 attributes here.12 of them are input attributes and we are gonna predict the Loan Status approved or not.It will be yes or no. Loan_ID, Gender, Married, Self employed, Applicant income, Co-applicant income, Education, Loan Amount, Loan Amount term, Credit history, Dependents, Property area and Loan Status are the attributes. The attributes Loan_ID, Gender, Married, Self employed, Education, Dependents,Property area and Loan Status are categorical variables. Rest of them are numerical variables.Here we are having 2 datasets. One is for training the model and other is for testing the model. Initially we combine these two for handling missing values and outliers together.

Columns	Description
Loan_ID	A uniques loan ID
Gender	Male/Female
Married	Married(Yes)/ Not married(No)
Dependents	Number of persons depending on the client
Education	Applicant Education (Graduate /Undergraduate)
Self_Employed	Self emplyored (Yes/No)
ApplicantIncome	Applicant income
Coapplicant income	Coapplicant Income
LoanAmount	Loan amount in thousands
Loan_Amount_Term	Term of lean in months
Credit_Hostory	Credit history meets guidelines
Property_Area	Urban/Semi and Rural
Loan_Status	Loan approved (Y/N)

We have 614 rows and 13 columns in the train dataset and367 rows and 12 columns in test dataset.

B. Exploratory Data Analysis

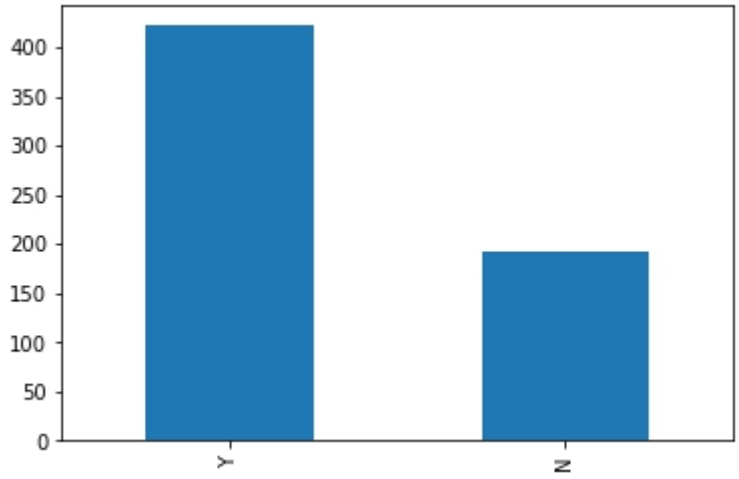
Two types of analysis are used:

- Univariate Analysis
- Bivariate Analysis

Univariate Visual Analysis

➢ **Target Variable - Loan Status**

We will start first with an independent variable which is our target variable as well. We will analyse this categorical variable using a bar chart as shown below. The bar chart shows that loan of 422 (around 69 %) people out of 614 was approved.



Predictor Variables

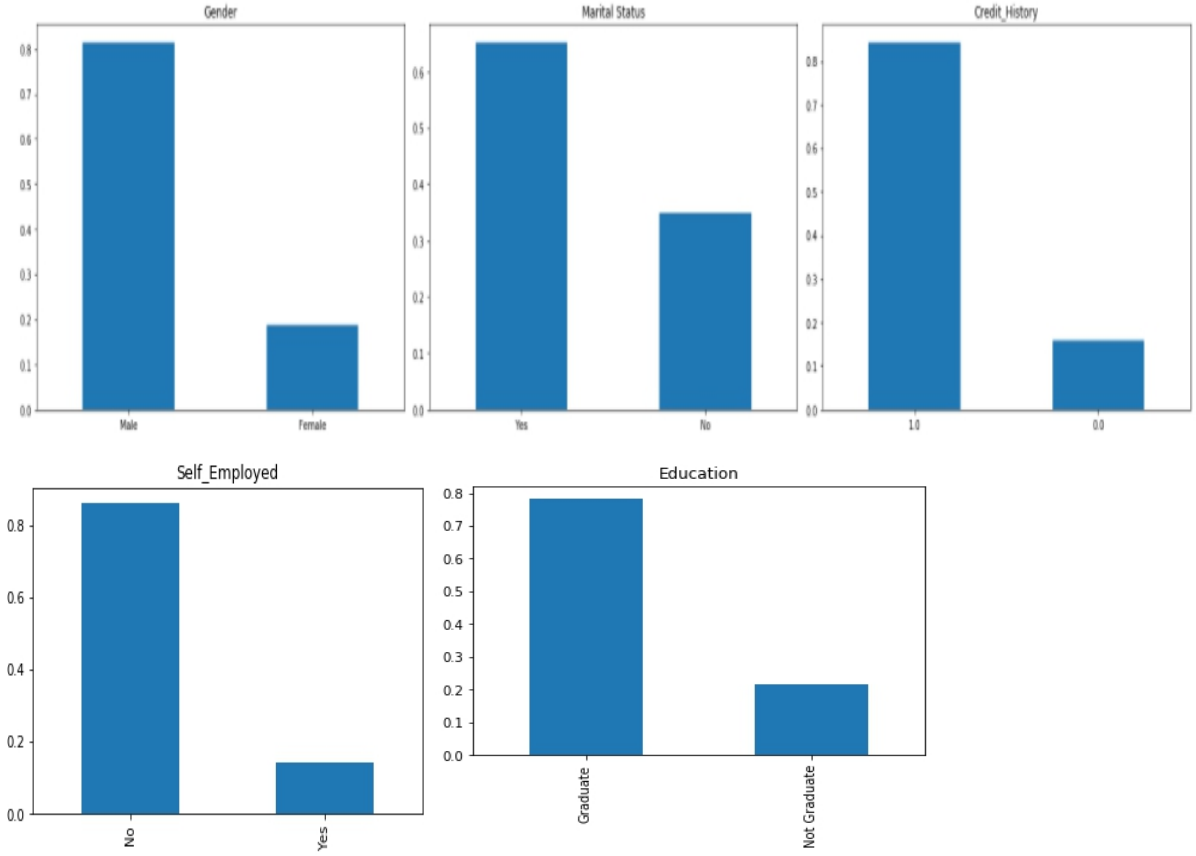
There are 3 types of Independent Variables: Categorical, Ordinal & Numerical.

Categorical Features

- Gender
- Marital Status
- Employment Type
- Credit History

It can be inferred from the below bar plots that in our observed data:

- 80% of loan applicants are male in the training dataset.
- Nearly 70% are married
- About 75% of loan applicants are graduates
- Nearly 85–90% loan applicants are self-employed
- The loan has been approved for more than 65% of applicants.

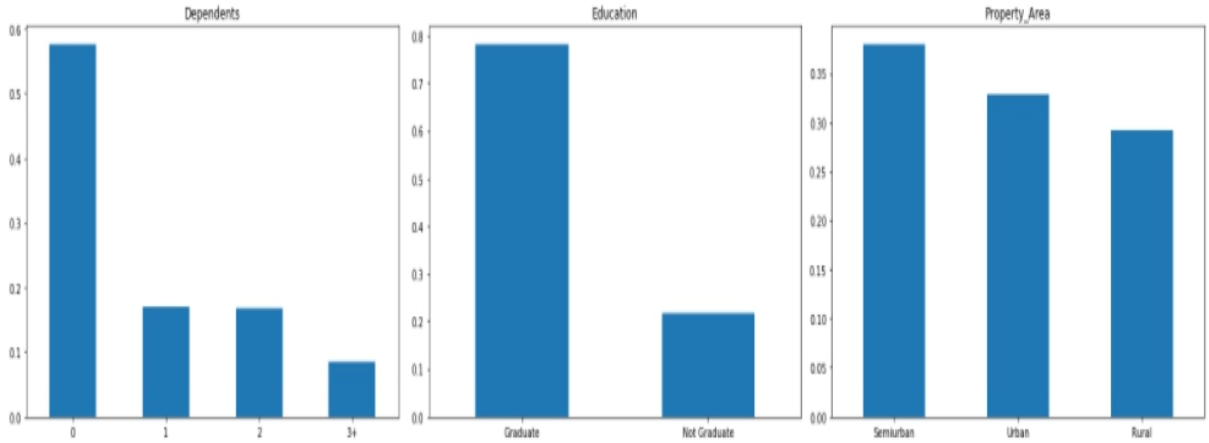


Ordinal Features

- Number of Dependents
- Education Level
- Property or Area Background

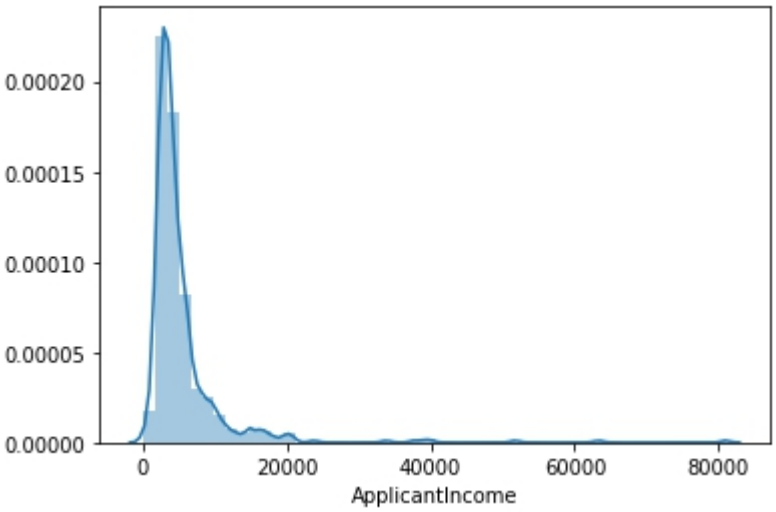
Our Visual Analysis below, indicates that:

- Almost 58% of the applicants have no dependents.
- Highest number of applicants are from Semi Urban areas, followed by urban areas.
- Around 80 % of the applicants are Graduate.

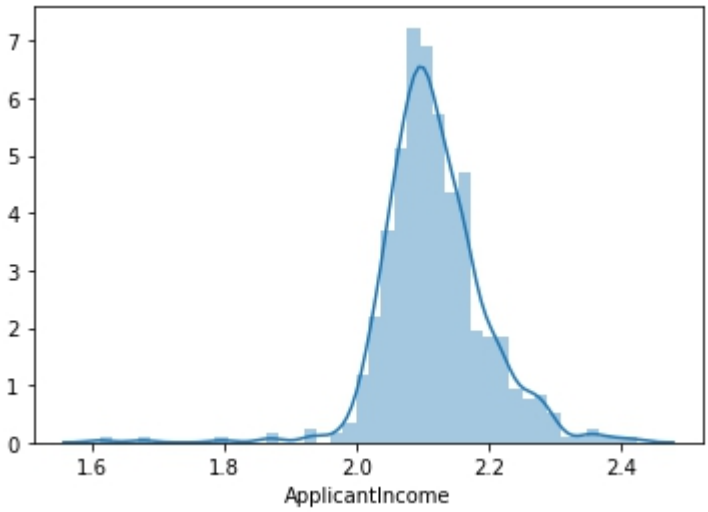


Numerical Features

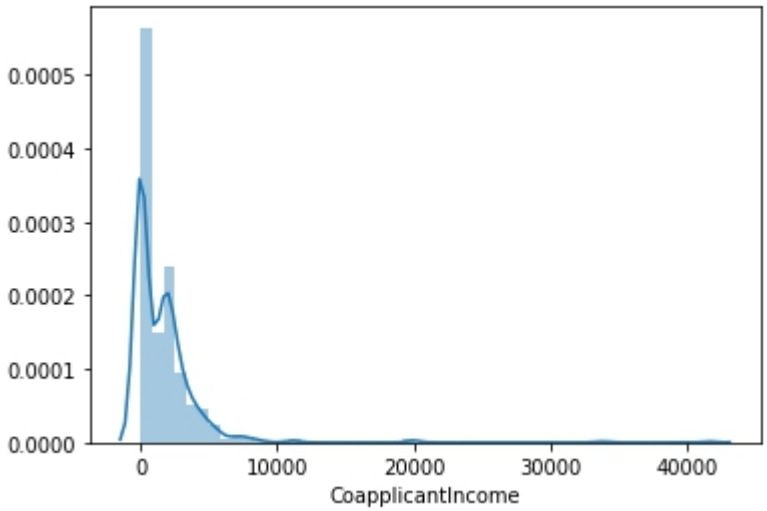
- The Applicant's Income
- The Co-Applicant's Income



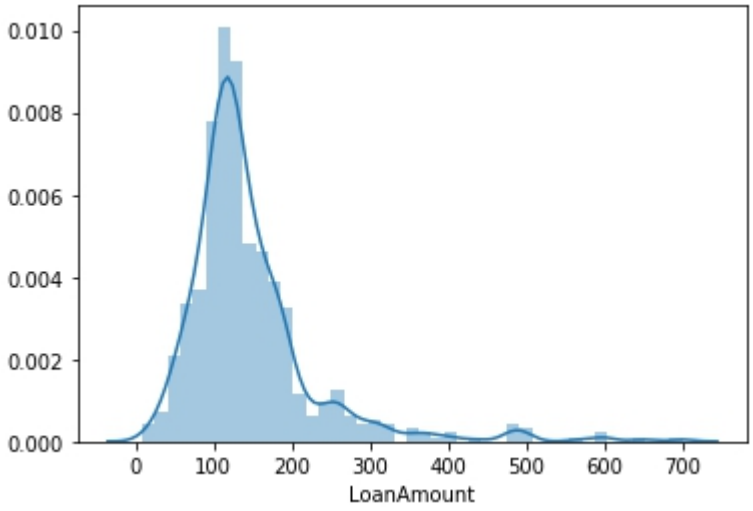
Here, majority of the applicant's income are in the range of 0 to 10000. Only few of them are in 40000 to 80000 and we can see the graph that are in left skewed. That is, most of the applicants are in left side. That is not suitable distribution for training the model. So we will apply the log function in the column to normalize the attributes in the form a bell curve.



Now,we can see in the form of a bell curve and it is normalized and now the mean is in the center instead of left skewed. This is good distribution for training the model.



Here, majority of the co applicant's income are in the range of 0 to 10000. Only few of them are in 20000 to 40000 and we can see the graph that are in left skewed. That is, most of the applicants are in left side. That is not suitable distribution for training the model. So we will apply the log function in the column to normalize the attributes in the form a bell curve.

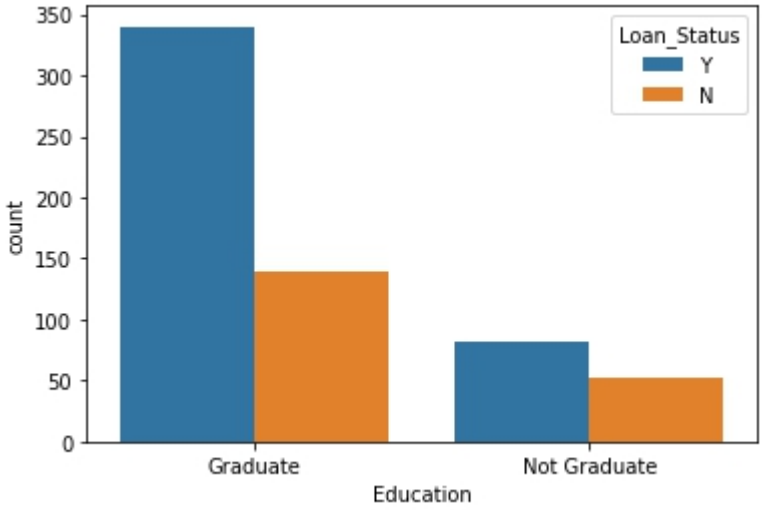


Now,we can see in the form of a bell curve and it is normalized and now the mean is in the center instead of left skewed. This is good distribution for training the model.

Bivariate Analysis

Bivariate analysis is finding some kind of empirical relationship between two variables. Specifically the dependent vs independent Variables.

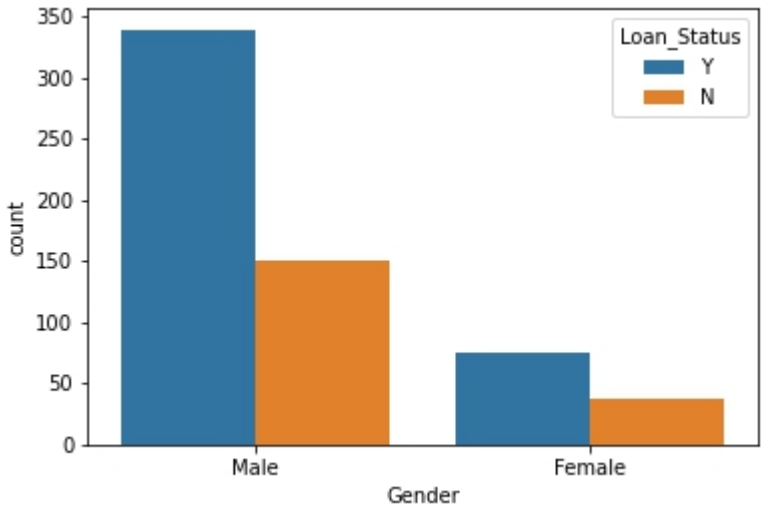
- **Education v/s Loan Status**



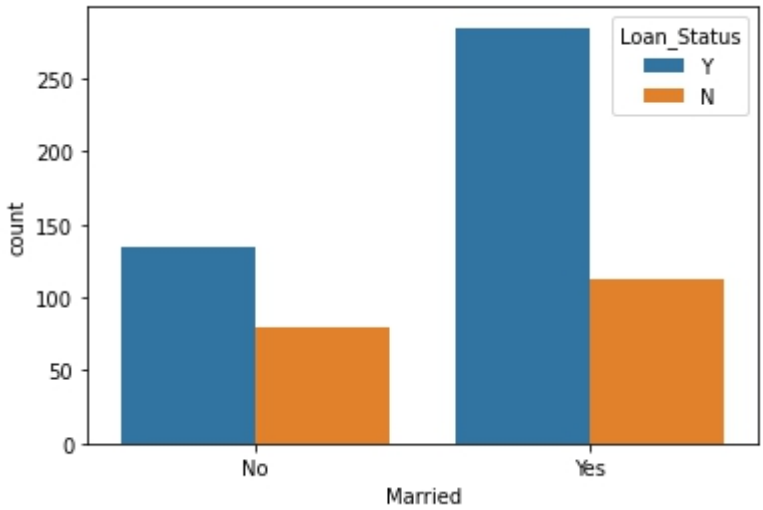
If the person is graduated, the company is giving loan for them in the most of the cases.

- **Gender v/s Loan Status**

More males are on loan than females. Also, those that are on loan are more than otherwise.



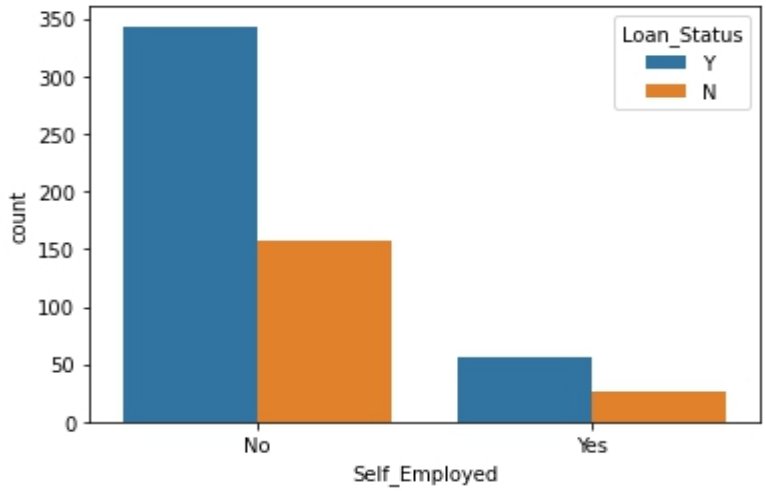
- **Marital status v/s Loan Status**



If a person is married, both husband and wife can contribute to settling the loan. So that is there is high chance for that finance company to get the money back.So ie. the reason high proportional of loan approval for married people.

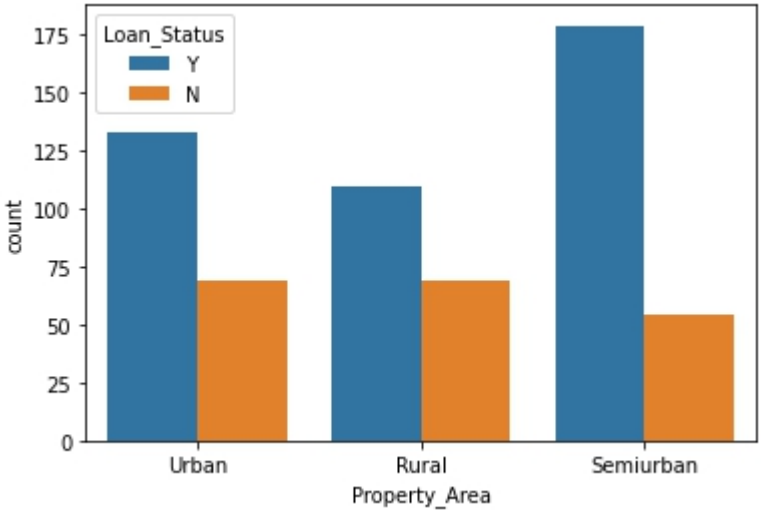
- **Self employed v/s Loan Status**

The category of those that take loans is less of self-employed people. That's those are not self-employed probably salary earners obtain more loan.



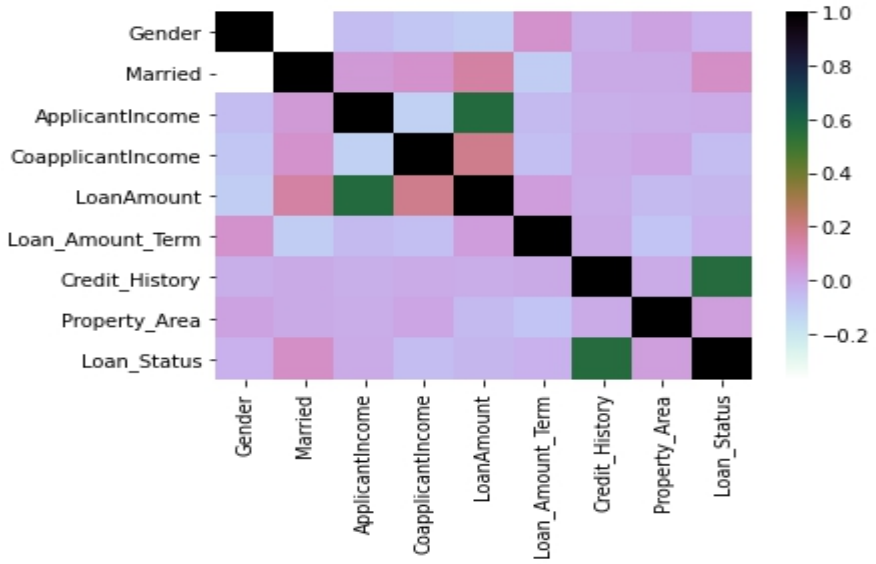
- **Property Area v/s Loan Status**

Semiurban obtain more loan, folowed by Urban and then rural. This is logical!



- **Correlation matrix**

Heat-map: Showing the correlations of features with the target. No correlations are extremely high. The correlations between Loan Amount and Applicant Income can be explained.



C. Missing value Handling and Outlier Treatment

After exploring all the variables in our data, we can now compute the missing values and treat the outliers because missing data and outliers can have adverse effect on the model performance.

Missing value Handling

There are missing values in Gender, Married,Dependents, Self_Employed, Loan Amount,Loan_Amount_Term, Credit_History and Loan Status.We will treat missing values in all features one by one.For handling ,we can choose mode,mean, median values.

For numerical columns: missing values handled by replacing it by mean or median.

For categorical columns: missing values handled by replacing it by mode.

Outlier Treatment

If data having outliers has a significant effect on the mean, and standard deviation and hence affecting the distribution. So we must take steps to remove outliers from our datasets. Hence we checked for outliers and replaced them with upper and limit values.Outliers can be detect by plotting box plots.

D. Modelling using Machine Learning Algorithms

We need to convert the categorical values into numerical values for modeling. So we convert it by using label encoding technique.Here modelling done by using 4 machine learning algorithms.They are:

- Logistic Regression
- Decision Trees
- Random Forest
- XGBoost

Logistic Regression

Let us make our first model predict the target variable. We will start with Logistic Regression which is used for predicting binary outcome.

- Logistic Regression is a classification algorithm. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables.
- Logistic regression is an estimation of Logit function. The logit function is simply a log of odds in favor of the event.
- This function creates an S-shaped curve with the probability estimate, which is very similar to the required stepwise function.

Now we will train the model on the training dataset and make predictions for the test dataset. We can train the model on this training part and using that make predictions for the validation part. In this way, we can validate our predictions as we have the true predictions for the validation part (which we do not have for the test dataset).

We will use the train_test_split function from sklearn to divide our train dataset. So, first, let us import train_test_split.The dataset has been divided into training and validation part. Let us import LogisticRegression and accuracy_score from sklearn and fit the logistic regression model.

Decision Trees

This is a supervised machine learning algorithm mostly used for classification problems. All features should be discretized in this model, so that the population can be split into two or more homogeneous sets or subsets. This model uses a different algorithm to split a node into two or more sub-nodes.With the creation of more sub-nodes, homogeneity and purity of the nodes increases with respect to the dependent variable.

Random Forest

This is a tree based ensemble model which helps in improving the accuracy of the model .It combines a large number of Decision trees to build a powerful predicting model.It takes a random sample of rows and features of each individual tree to prepare a decision tree model. Final prediction class is either the mode of all the predictors or the mean of all the predictors.

XGBoost

This algorithm only works with the quantitative variable.It is a gradient boosting algorithm which forms strong rules for the model by boosting weak learners to a strong learner. It is a fast and efficient algorithm which recently dominated machine learning because of its high performance and speed.

Stratified k-folds cross-validation

To check how robust our model is to unseen data, we can use Validation. It is a technique that involves reserving a particular sample of a dataset on which you do not train the model. Later, you test your model on this sample before finalizing it. Some of the common methods for validation are listed below:

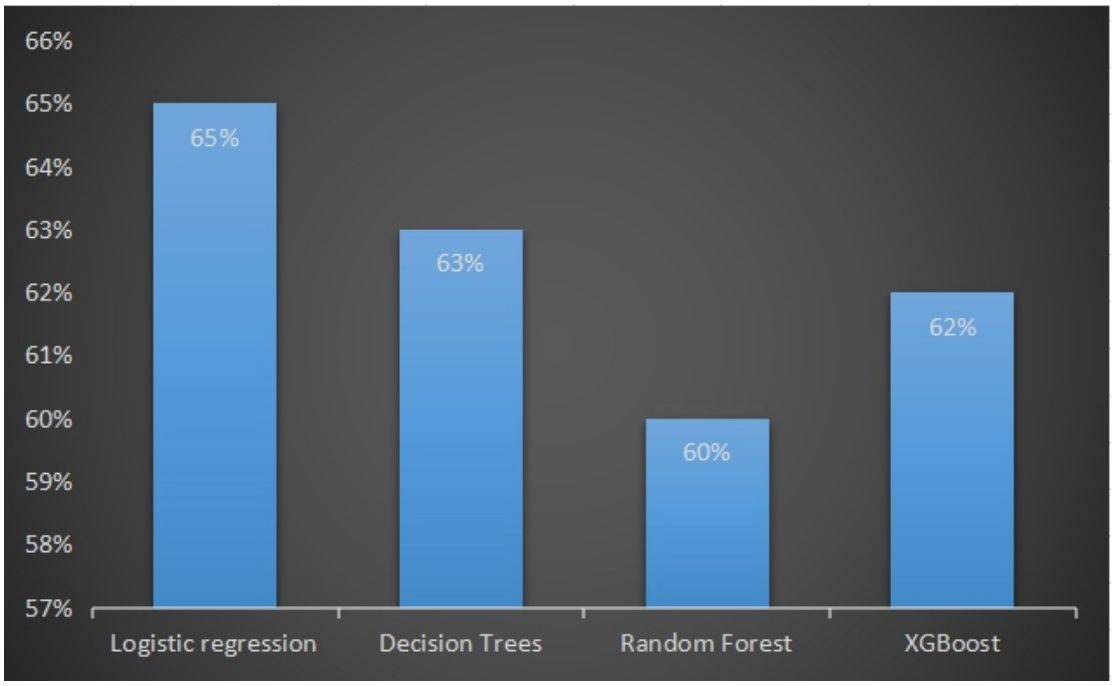
- The validation set approach
- k-fold cross-validation
- Leave one out cross-validation (LOOCV)
- Stratified k-fold cross-validation

Here we use Stratified k-fold cross-validation.

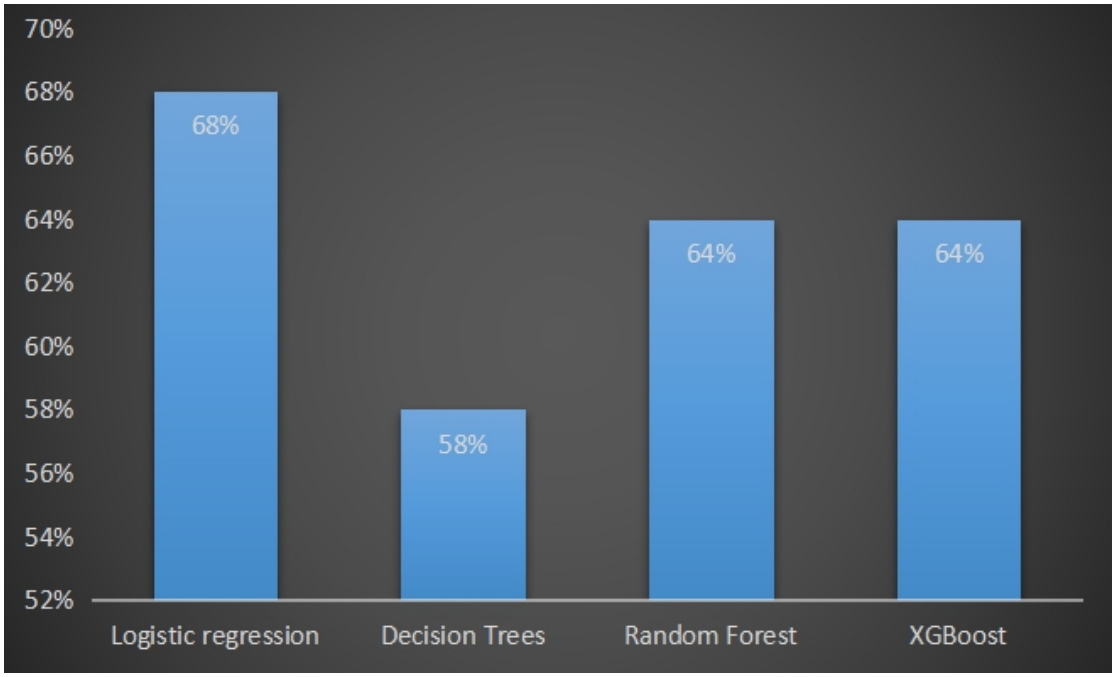
- Stratification is the process of rearranging the data so as to ensure that each fold is a good representative of the whole.
- For example, in a binary classification problem where each class comprises of 50% of the data, it is best to arrange the data such that in every fold, each class comprises of about half the instances.
- It is generally a better approach when dealing with both bias and variance.
- A randomly selected fold might not adequately represent the minor class, particularly in cases where there is a huge class imbalance.

E. Evaluation of Models

Below graphs shows the accuracy score of all the models.



This graph shows the accuracy score of the prediction done by using basic algorithms. Logistic regression algorithm provides 65 % accuracy score. Compared with others it is the suitable model. Logistic regression using stratified k-folds cross-validation provides 68 % accuracy score. Compared with others it is the suitable model.



The above graph shows the accuracy score of the prediction done by using stratified k-fold cross validation technique in each models.

Conclusion

From the Exploratory Data Analysis, we could generate insight from the data.Did Exploratory data Analysis on the features of this dataset and saw how each feature is distributed.We Did univariate and bivariate analysis to see impact of one another on their features using charts.Handled missing values and outlier treatment successfully

Compare accuracy of each classifiers and logistic regression using stratified K-fold Cross validation gives the highest accuracy (68%)