Loan Application Status Prediction

Dated: 08.06.2021

**Prepared By**: Bhumik P. Shah

**Data Trained Batch No**: 1828

**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. The main profit of the banks comes directly from the loan’s interest. The loan companies grant a loan after intensive process of verification and validation. However, they still don’t have assurance if the applicant is able to repay the loan with no difficulties.

**Dataset:**

The data has 615 rows and 13 columns.

**Dataset Description:**

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Loan\_ID | Unique Loan ID |
| Gender | Male/Female |
| Married | Applicant married(Y/N) |
| Dependents | Number of dependents |
| Education | Applicant Education(Graduate/Under Graduate) |
| Self\_Employed | Self employed (Y/N) |
| ApplicantIncome | Applicant Income |
| CoapplicantIncome | Coapplicant Income |
| Loan\_Amount | 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) |

**Libraries used:**

* NumPy
* Scikit Learn
* MatplotLib
* Seaborn
* Pandas
* Github

**Data Types:**

**There are three formats of data types:**

* **Object-** The variables are categorical in nature. The features named as Loan\_ID, Gender, Married, Dependents, Education, Self\_Employed, Property\_Area, Loan\_Status are of object datatype.
* **Int64-** It represents the integer variables. The feature named as Applicant Income is having int64 as a datatype.
* **Float64-** It represents the float variables. The features named as Co-applicant Income Loan Amount, Loan\_Amount\_Term is having float64 as datatype.

**Null Values in the Dataset:**

There are null values present in the features named as Gender, Dependents, Self\_Employed, Loan Amount, Loan\_Amount\_Term and Credit\_History.

Loan\_ID 0

Gender 13

Married 3

Dependents 15

Education 0

Self\_Employed 32

ApplicantIncome 0

CoapplicantIncome 0

LoanAmount 22

Loan\_Amount\_Term 14

Credit\_History 50

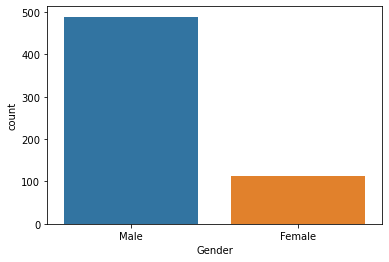
Property\_Area 0

Loan\_Status 0

**Univariate Analysis:**

For Categorical features count plot from seaborn library is used which will calculate the number

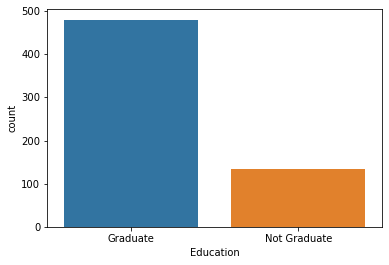
of each category in a particular variable.



There are 489 records (80%) for Male and 112 records (20%) for Female in the dataset.



There are 398 records (65%) who have applied for loan are married and rest 213 (35%) are not married.



There are 480 records(66%) who have applied for loan are Graduate and rest 134 records (34%) are Not Graduate.



The loan of 422(around 69%) people out of 614 was approved.

**Missing value Imputation**:

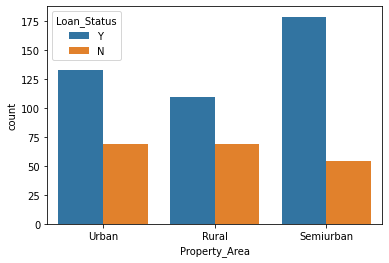
After exploring all the variables in the dataset, we can now impute the missing values as missing data can have adverse effect on model performance.

**For Numerical variables:** Imputation using mean.

**For Categorical variables**: Imputation using mode.

**Bivariate Analysis:**

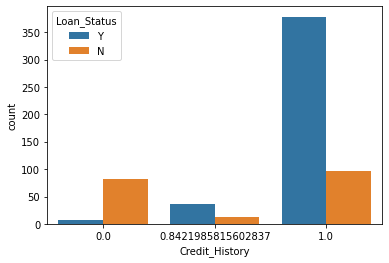
After looking into the variables individually in univariate analysis, we will now explore them again with respect to target variable.



Most of the loan applications i.e 233 out 614 are from semiurban areas. The loan approval percentage is 77% for semiurban areas.

The loan applications i.e 202 out of 614 are from urban areas. The loan approval percentage is 66% for urban areas.

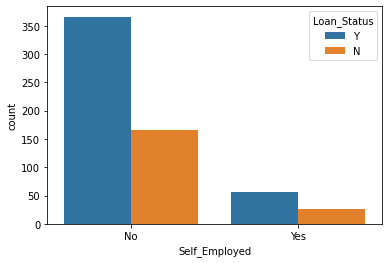
The least applications i.e 179 out of 614 are from rural areas. The loan approval percentage is 61% for rural areas.



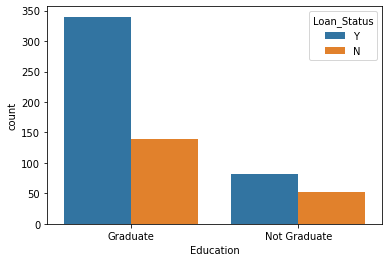
Most of the loan applications i.e. 475 out of 614 are from those whose credit rating is 1. The loan approval percentage i.e. 378 out of 475 applications is 80% with credit rating 1.

The loan applications i.e. 89 out of 614 are from those whose credit rating is 0. They are not having any loan history or the case may be the data is not available. The loan approval percentage i.e. 7 out of 89 applications is just 8% with credit rating 0.

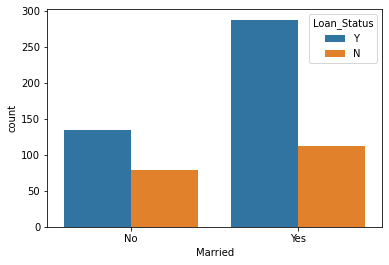
The least loan applications i.e. 50 out of 614 are from those whose credit rating is 0.842199. The loan approval percentage i.e. 37 out of 50 applications is 74% with credit rating 0.842199.



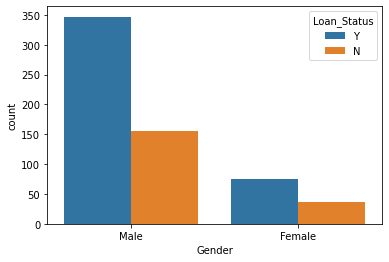
Most of the loan applications i.e.532 out of 614 are from those who are not self-employed. The loan approval percentage i.e. 366 out of 532 applications is 69% for those who are not self-employed.



Most of the loan applications i.e 480 out of 614 are from those who are graduate. The loan approval percentage i.e 340 out of 480 applications is 71% having graduation as a qualifying degree.

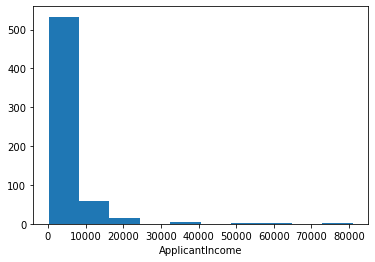


Most of the loan applications i.e.502 out of 614 are from those who are married. The loan approval percentage i.e 288 out of 502 applications is 68% for the ones who are married.

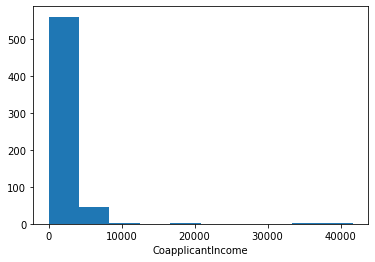


Most of the loans applications a i.e 502 out of 614 are from Male category. The loan approval percentage i.e 347 out of 502 applications is 69% for Male category.

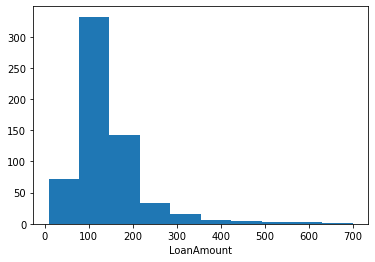
**Understanding Distribution of Numerical Variables:**



The Applicant-Income is lying in the range from 0 to 10000 rupees for most of the records available in this dataset. Further, few of the records in this dataset are having income between 30000 to 40000 ,50000 to 60000 and 75000 to 80000 i.e very small blocks can be seen in the above histogram, the chances are they are either outliers or they actually are the application-income.



The Co-Applicant-Income is lying in the range from 0 to 5000 rupees for most of the records available in this dataset. Further, few of the records in this dataset are having income of 10000, 20000 , 35000 to 41000 i.e very small blocks can be seen in the above histogram, the chances are they are either outliers or they actually are the the Co- Applicants-income.



Most of the records for the feature Loan-Amount lies in the range of 80 to 150. Further, small blocks can be seen in the above histogram from 300 to 700 is nothing but a skewness(data is not symmetrical), the chances are they are outliers or it could be that the sanctioned loan amount is greater based on the applicant as well as co-applicant income.

**Correlation:**

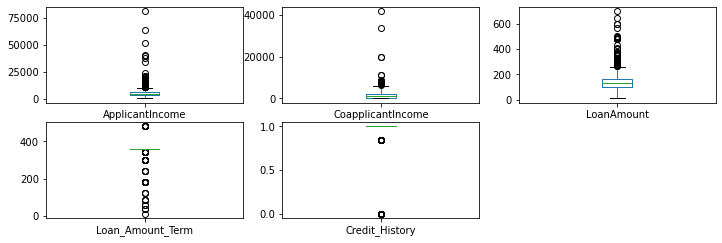
As per the heatmap attached below, it is observed that the features Application-Income, Co-Applicant-Income, Loan-Amount are having very less correlation with Credit History.

There is a very good correlation i.e 0.57 between Applicant-Income and the loan-Amount.

.



**Outlier Detection:**

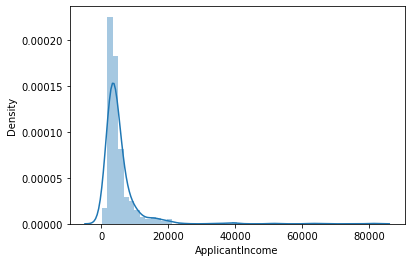
****

There are outliers present in features named as Applicant-Income, Co-applicant-Income, Loan amount , Loan-Amount-Term, Credit-history as per the boxplot attached above.

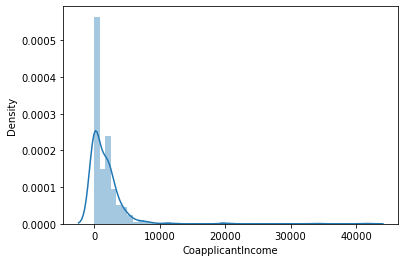
I am not treating any of the outliers, as the case may be that an Applicant-Income and Co-Applicant Income can be higher than the average income of other applicants and co-applicants.

The Loan-Amount sanctioned can be more depending on the applicant-income and co-applicant income.

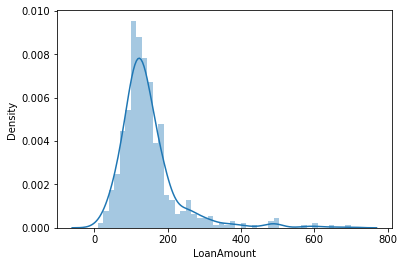
**Skewness:**

****

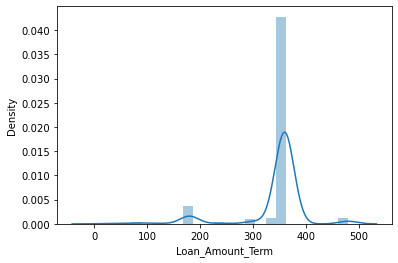
The feature Applicant-Income is having right tailed skewness where the mean is greater than the median as per the distribution plot attached above.



The feature Co-Applicant-Income is having right tailed skewness where the mean is greater than the median as per the distribution plot attached above.



The feature Loan-Amount is having right tailed skewness where the mean is greater than the median as per the distribution plot attached above.



The feature Loan-Amount-Term is having left tailed skewness where the median is greater than the mean as per the distribution plot attached above.

ApplicantIncome 0.479580

CoapplicantIncome 7.491531

LoanAmount 2.726601

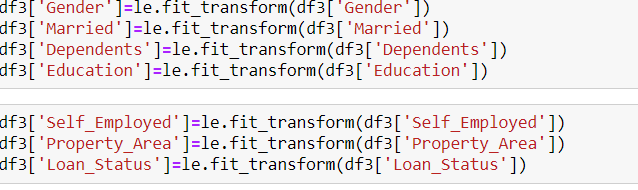
Loan\_Amount\_Term -2.389680

Credit\_History -1.963600

The Skewness in numerical variables is attached above. Further, to treat skewness log transformation is applied as the skewness greater than 0.5 is not generally acceptable .

**Label Encoding:**

The Label Encoding is applied on the categorical variables as the machine learning models only understand the numerical values i.e integer and float values.



The Label Encoding is applied on above mentioned features.

**Model building:**



The data is segregated into x which will have all independent features and y with a dependent feature i.e in other words it is the target variable.

Before splitting the data into train\_test\_split standard scaler is applied to independent variables(input variables). The idea behind StandardScaler is that it will transform the data such that its distribution will have a mean value 0 and standard deviation of 1 so that the machine learning models will not be biased to any particular values.

In Train Test Split validation we split the training data into training and validating data as to evaluate our model.

Splitting the data into test\_train\_split as per code attached below.



**Logistic Regression:**

The model is build using Logistic Regression as one of the supervised machine learning algorithms.

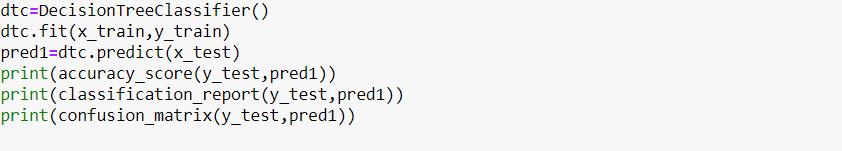




**The accuracy score is 0.772 for this model**.

**Decision Tree Classifier:**

The model is build using Decision Tree Classifier as one of the supervised machine learning algorithms.

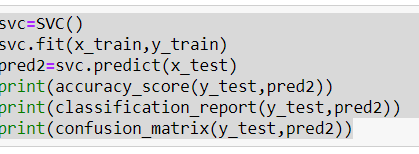




**The accuracy score is 0.731 for this model.**

**Support Vector Machine (SVC):**

The model is build using SVC as one of the supervised machine learning algorithms.

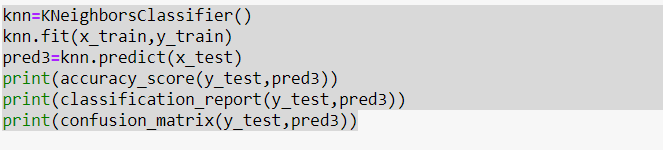


****

**The accuracy score is 0.756 for this model.**

**K Neighbours Classifier:**

The model is build using K Neighbours Classifier as one of the supervised machine learning algorithms.

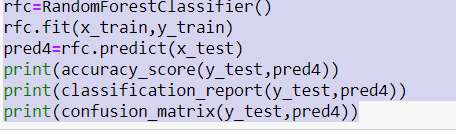




**The accuracy score is 0.731 for this model.**

**Random Forest Classifier:**

The model is build using Random Forest Classifier as one of the supervised machine learning algorithms.

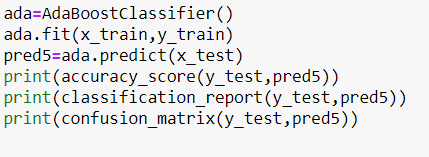




**The accuracy score is 0.756 for this model.**

**AdaBoost Classifier:**

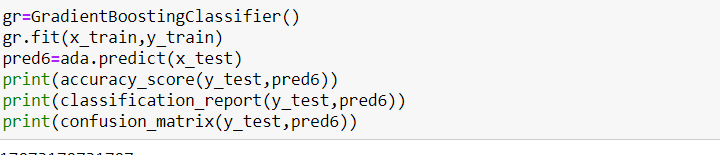
The model is build using AdaBoost Classifier as one of the supervised machine learning algorithms





**The accuracy score is 0.731 for this model.**

**Gradient Boosting Classifier:**





**The accuracy score is 0.731 for this model.**

**Cross validation score:**

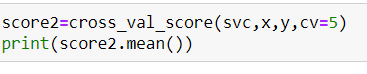
The cross validation score for logistic Regression, Decision tree classifier, SVC, K Neighbours Classifier, Random Forest Classifier, Ada boost Classifier and Gradient Boosting Classifier is **0.770,0.706,0.765,0.731,0.778,0.778,0.775.**











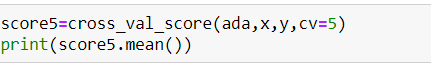
















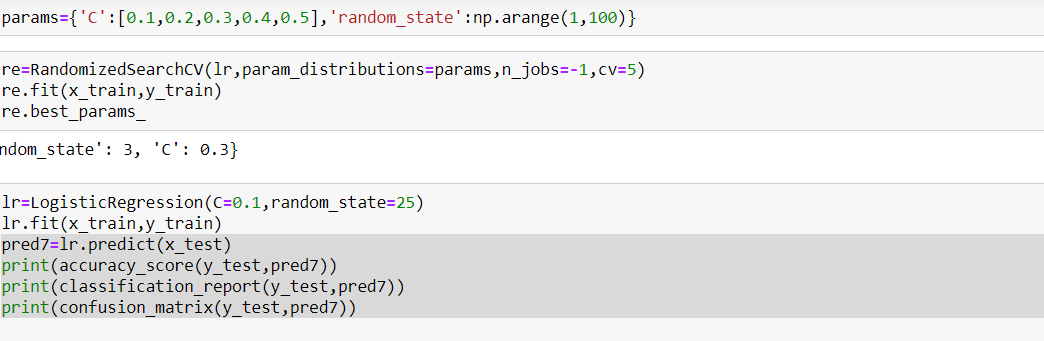


**Observations:**

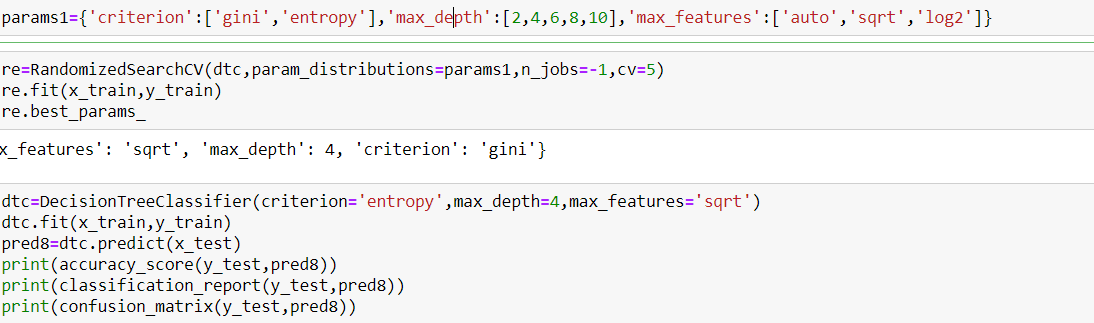
The difference between cross validation score and the accuracy score is minimum for Logistic Regression, Decision Tree Classifier, K Neighbours Classifier and Random Forest Classifier. To find the best model out of all these hyper tuning is required**.**

**Hyper Tuning:**

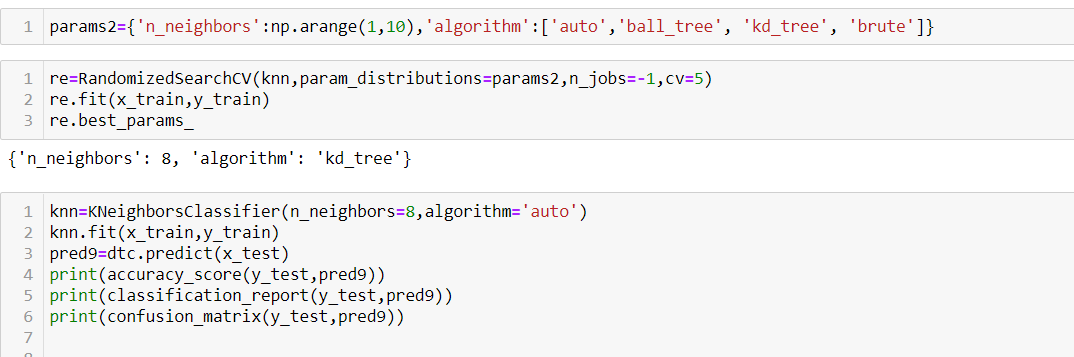
After performing hyper tuning on Logistic Regression, Decision Tree Classifier, K Neighbours Classifier and Random Forest Classifier the accuracy score is 0.780,0.63,0.63 and 0.788.

















**Conclusion**

Out of all the classification algorithms used on the dataset, Random Forest Classifier algorithm gives the best overall prediction accuracy i.e. accuracy score of 79%.

In near future this module of prediction can be integrated with module of automated processing system**.**

**Future Scope**

An app with proper UI can be built, which can take various inputs from the user like, name, address, loan amount, Education, loan duration etc and give a prediction of whether their loan application can be approved by the banks or not.