A person shaking hands with another person

Description automatically generated

1st July 2024

**Loan Application Status Prediction**

* FlipRobo Technologies DataTrained

By:-

Priti Varma

Table of Contents

1. **Project Description………………………………………………………………………………………...02**
2. **Problem Statement…………………………………………………………………………………………02**
3. **Tech Stack Used…………………………………………………………………………………………….03**
4. **Objective……………………………………………………………………………………………………...03**
5. **Importing Libraries……………………………….…………………………………………………………04**
6. **Data Collection and Loading………….…………………………………………………………………04**
7. **Data Preprocessing.………………………………………………………………………………………..04**
8. **Finding Null Values.…………………………………………………………………………………………05**
9. **Exploratory Data Analysis.……………………………………………………………………………….07**
10. **Data Visualization and Interpretation.……………………………………………………………….09**

10.1 Univariate Analysis…………………………………………………………………………...**09**

10.2 Bivariate Analysis…………………………………………………………………….………..**20**

10.2.1 Data Visualization……………………………………………………….……….**20**

10.2.2 Descriptive Statistics………………………………………………….………..**23**

10.2.3 Finding Correlation……………………………………………………………….**24**

**11.** **Multivariate Analysis……………………………………………………………………………………….25**

**12. Outliers…………………………………………………………………………………………….….……….26**

12.1 Finding Outliers…………………………………………………………………………………**26**

12.2 Removal of Outliers by Z-score Method………………………………………….……..**27**

12.3 Removal of Outliers by IQR-Method………………………………………………………**28**

**13. Skewness.……………………………………………………………………………………………………..29**

13.1 Finding Skewness………………………………………………………………………….…..**29**

13.2 Removal of skewness…………………………………………………………………………**30**

**14. Separating Features and Label/target variables.…………………………………………………31**

**15. Feature Engineering: using Standard Scaling Method (Normalization).………………..…32**

**16. Checking Variance inflation factor (VIF).…………………………………………………………....32**

**17. Balancing the Target Variable.………………………………………………………………………....33**

**18. Modelling or Model Building and Cross\_valuation.……………………………………………...33**

**19. Model Evaluation…………………………………………………………………………………………....35**

**20. Hyper Parameter tuning………………………………………………………………………………..…36**

20.1 Explanation………………………………………………………………………………….…...**36**

20.2 Fit the Model: Fit the model using random\_search.fit (X, y)……………………..…**36**

20.3 Best Model Evaluation.……………………………………………………………………….**37**

20.3.1 Print Best Parameters and Score…………………………………………….**37**

20.3.2 Evaluate the Best Model…………………………………………………….....**37**

**21. ROC -AUC CURVE.…………………………………………………………………………………………..38**

21.1 Plotting ROC and compare AUC for the Best Model.………………………………..**43**

21.3 Plotting the ROC Curve for the Final Model and getting the AUC value………..**44**

**22. Saving the Model…………………………………………………………………………………………….45**

**22. Predicting The Saved Model……………………………………………………………………………..45**

**23. Comparison…………………………………………………………………………………………………..46**

1. **Project Description:**

This dataset includes details of applicants who have applied for loan. The dataset includes details like credit history, loan amount, their income, dependents etc. This project aims to analyse the risk appetite of financial organizations when deciding whether to approve loan applications based on applicant profiles. There are two primary risks associated with these decisions:

1. **Risk of Not Approving a Loan to a Creditworthy Applicant**:

If a bank rejects a loan application from an applicant who would have repaid the loan successfully, the bank loses potential business.

1. **Risk of Approving a Loan to a Default-Prone Applicant**:

If a bank approves a loan for an applicant who is likely to default, it could lead to financial losses for the bank.

1. **Problem Statement:**

You have to build a model that can predict whether the loan of the applicant will be approved (Loan\_status) or not on the basis of the details provided in the dataset.

**Features:**

1. Loan\_ID - This refers to the unique identifier of the applicant's affirmed purchases
2. Gender - This refers to either of the two main categories (male and female) into which applicants are divided on the basis of their reproductive functions
3. Married - This refers to applicant being in a state of matrimony
4. Dependents - This refers to persons who depends on the applicants for survival
5. Education - This refers to number of years in which applicant received systematic instruction, especially at a school or university
6. Self\_Employed - This refers to applicant working for oneself as a freelancer or the owner of a business rather than for an employer
7. Applicant Income - This refers to disposable income available for the applicant's use under State law.
8. CoapplicantIncome - This refers to disposable income available for the people that participate in the loan application process alongside the main applicant use under State law.
9. Loan\_Amount - This refers to the amount of money an applicant owes at any given time.
10. Loan\_Amount\_Term - This refers to the duration in which the loan is availed to the applicant
11. Credit History - This refers to a record of applicant's ability to repay debts and demonstrated responsibility in repaying them.
12. Property\_Area - This refers to the total area within the boundaries of the property as

set out in Schedule.

**Target Variable:**

Loan\_Status - This refers to whether applicant is eligible to be availed the Loan requested.

**Dataset Link** : <https://github.com/FlipRoboTechnologies/ML_Datasets/blob/main/Loan%20Application%20Status/loan_prediction.csv>

1. **Tech Stack Used:**

**Python and Jupiter Notebook-** The purpose is to perform comprehensive analysis and create graphical representations of the results to enhance understanding of the dataset. Later model Building and Finding the best fitted Model to train the data for further Predictions.

#### To build a predictive model for loan approval status using the given dataset, we can follow these steps:

* Data Collection and Loading: Load the dataset from the provided link.
* Data Preprocessing: Clean and preprocess the data, handle missing values, and encode categorical variables.
* Exploratory Data Analysis (EDA): Understand the data distribution, visualize relationships, and identify important features.
* Outliers: Finding Outliers and using Removal techniques
* Working on skewness to bring the values nearer to +/- 0.5.
* Feature Engineering: Create or transform features to improve the model.
* Model Building: Train and evaluate different machine learning models.
* Cross Validation Score.
* Model Evaluation: Evaluate the performance of the models and select the best one.
* ROC AUV Curve: Plotting ROC AUC Curve.
* Prediction: Use the trained model to make predictions on new data.

#### **Objective:**

The main objective of this project is to develop a predictive model that can accurately determine the loan approval status based on the given features of the applicants.

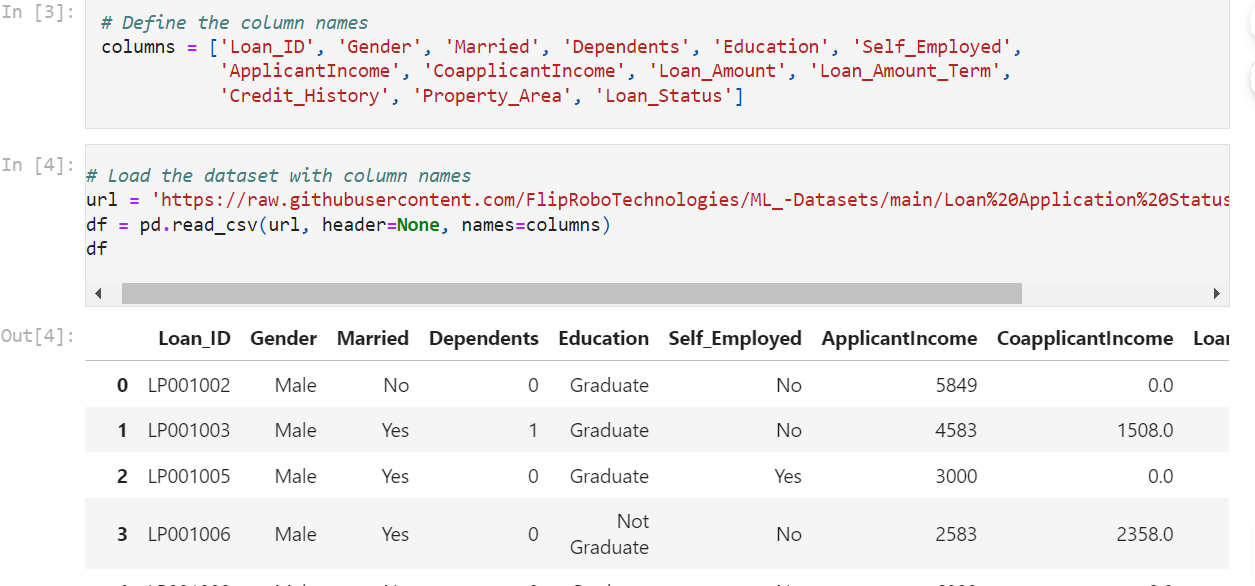
1. **Importing Libraries:**

All the required Libraries have been imported before starting to anayse.so that we don’t face any errors while performing Data Analysis and Model Building. Following were the Libraries imported.

A white rectangular object with a black border

Description automatically generated

1. **Data Collection and Loading :**

Before proceeding further we found that column header were not provided to the data. Hence lets first provide column headers to the dataset. Later load the dataset from the provided link. 

After loading the dataset with the correct column names, we can proceed with the data preprocessing,

1. **Data Preprocessing:**

Here we will clean and preprocess the data, handle missing values, and encode categorical variables.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* We can see that null values are present in the data and we will have to check and treat them.

1. **Finding Null Values:**

A screenshot of a computer

Description automatically generated

sns**.**heatmap(df**.**isnull()) *# we can clearly visualize that there are null value present*

A black and white chart with white lines

Description automatically generated with medium confidence

A screen shot of a computer code

Description automatically generated

* Hence we have filled all the null values and the data is ready for further process of Statistics.

1. **Exploratory Data Analysis(EDA):**

Understand the data distribution, visualize relationships, and identify important features. Following is the analysis.

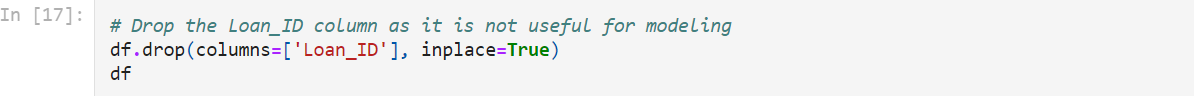
#### Before working on statistics we will have to check Loan\_ID column where we found that all the values are unique. Hence we drop the column Loan\_ID.

#### 

A rainbow colored background with black border

Description automatically generated

**Since all the Loan\_ID are unique we can drop the column**



A screenshot of a computer

Description automatically generated

* **Let’s drop columns ApplicantIncome and CoapplicantIncome as we got their Total\_Income columns**

A screenshot of a computer

Description automatically generated

# **Data Visualization and Interpretation:**

**10.1 Univariate Analysis**

**Univariate analysis** involves the examination of a single variable. The primary purpose is to describe and understand the distribution, central tendency, and variability of that variable.

A white background with text

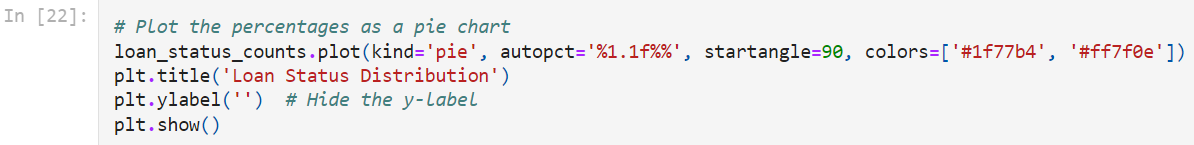
Description automatically generated

A graph of a loan status

Description automatically generated

A computer screen shot of a computer code

Description automatically generated

**A blue and orange pie chart

Description automatically generated

Y 68.729642

N 31.270358

Name: Loan\_Status, dtype: float64

* **Insights:- So overall 68.7% of Clients got their Loans Approved and 31.3% Clients were rejected.**

*A screen shot of a computer code

Description automatically generated*

A blue and orange rectangular bar graph

Description automatically generated

Male 81.758958

Female 18.241042

Name: Gender, dtype: float64

* **Insights:- 81.76% of Males have applied for Loans and only 18.24% are Female Applicants.**

A computer code with text

Description automatically generated

A graph of a person and person

Description automatically generated

Yes 65.309446

No 34.690554

Name: Married, dtype: float64

* **Insights:- Out of Total Applicants 65.31% are Married and 34.69% are Unmarried Applicants**

*A computer code with text

Description automatically generated with medium confidence*

A graph with different colored squares

Description automatically generated

0 58.631922

1 16.612378

2 16.449511

3+ 8.306189

Name: Dependents, dtype: float64

**Insights:-**

* **Most of the Applicants (58.63%) applying for Loans had no Dependent.**
* **16.61% had 1 Dependent**
* **16.45% have 2 Dependents**
* **Only 8.31% had more than 3 dependents**

*A computer code with text

Description automatically generated with medium confidence*

A graph of a graduate and not graduate

Description automatically generated

Graduate 78.175896

Not Graduate 21.824104

Name: Education, dtype: float64

* **Insights: - Out of the Total Applicants 78.18% were graduates and 21.82% were Non Graduates.**

*A computer code with colorful text

Description automatically generated*A blue and orange rectangular column

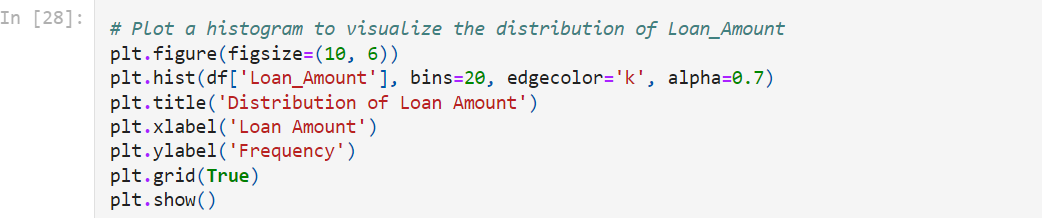
Description automatically generated

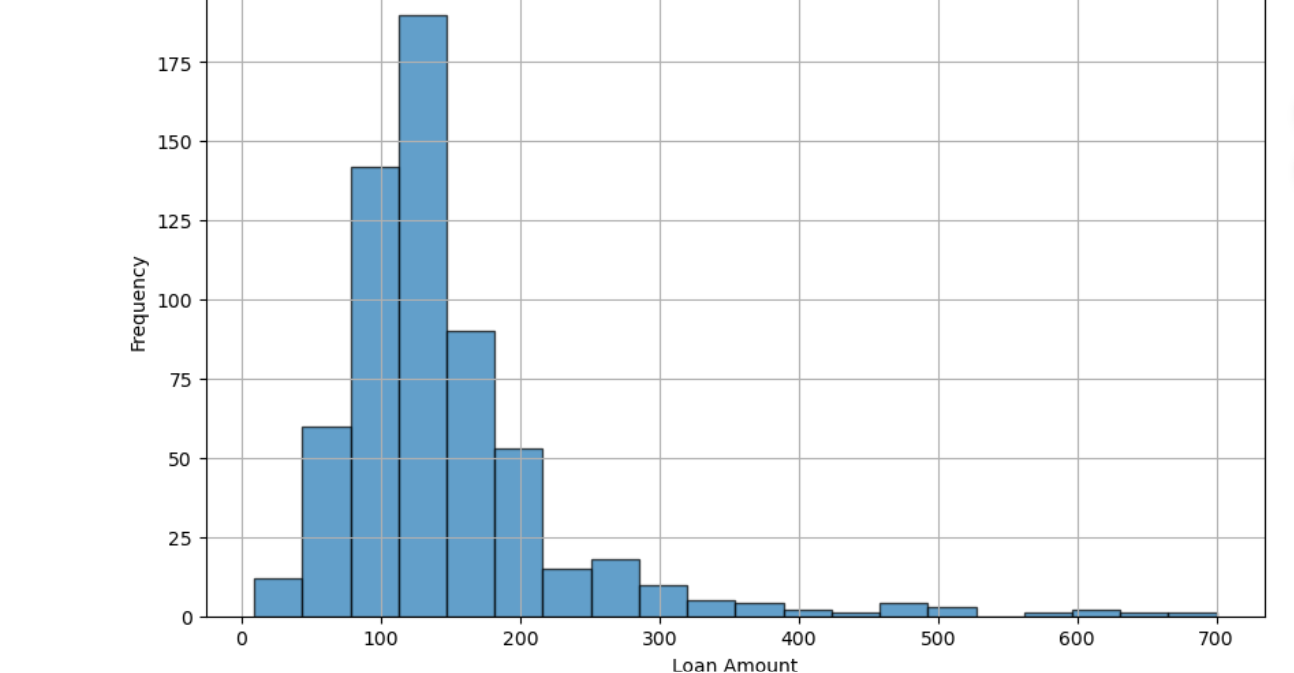
No 86.644951

Yes 13.355049

Name: Self\_Employed, dtype: float64

* **Insights :- Out of the total Applicants 86.64% are not Self\_Employed and 13.36% are Self-Employed Applicants.**

**



By running the above code, we get a histogram showing the distribution of loan amounts and a box plot to visualize the spread and detect any outliers in the loan amounts.

*A screenshot of a computer program

Description automatically generated*

**This code will give the maximum and minimum values of the loan amounts with the frequent loan amount in your dataset.**

*A computer code with colorful text

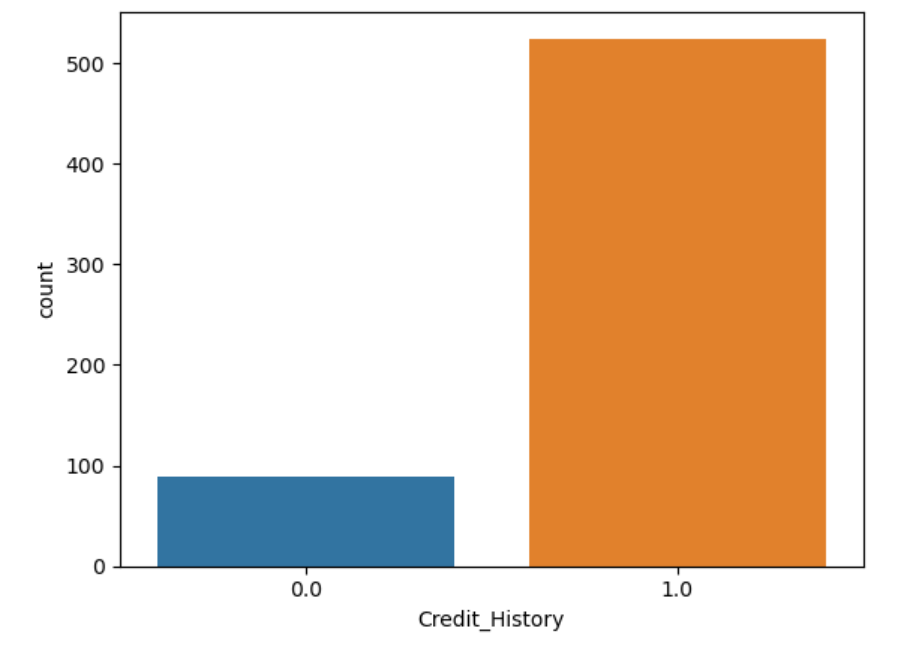
Description automatically generated*A graph with a bar graph

Description automatically generated

*A screenshot of a computer program

Description automatically generated*

*A close-up of a credit history

Description automatically generated*

* **Insights:- Most of the Applicants approaching for loans have Good Credit\_History.**

*A computer code with text

Description automatically generated with medium confidence*

A graph of different colored rectangular shapes

Description automatically generated

Semiurban 37.947883

Urban 32.899023

Rural 29.153094

Name: Property\_Area, dtype: float64

* **Insights: Out of the Total Applicants 37.95% belongs to Semiurban , 32.9% belongs to Urban and only 29.15% of Applicants belongs to rural Area**

**Now let’s try to Segregate the Numerical columns and Categorical column to get a proper Data Analysis.**

A screen shot of a computer code

Description automatically generated

*A screenshot of a computer code

Description automatically generated*

A group of graphs showing the amount of loan amount

Description automatically generated

* Insights:- Loan Amount and total Income are right skewed and loan\_Amount\_term and Credit History are left skewed and we will have to treat them.
  1. **Bivariate Analysis**

**10.2.1 Data Visualization**

After Univariate Analysis, Bivariate Analysis is necessary to understand the comparison between two variables. This is to understand if the variable is directly proportional / indirectly proportional to another variable. The correlation between the two variables can be understood.

A screenshot of a computer program

Description automatically generated

A graph of different colored squares

Description automatically generated

A graph of a number of people

Description automatically generated with medium confidence

A graph of different colored squares

Description automatically generated

A graph of a bar graph

Description automatically generated with medium confidence

A graph of different colored squares

Description automatically generated

A green and orange rectangular bars

Description automatically generated

A graph of a bar graph

Description automatically generated with medium confidence

### **OBSERVATION:**

* **Males are the maximum number of applicants.**
* **Maximum number of applicants seems to be married.**
* **Most of the applicants have no dependents on them.**
* **Majority of the applicants who have applied for loans are Graduates.**
* **Most of the applicants applying for Loans are not Self-employed. They may be Salaried.**
* **We observe that most of the Customers whose Credit\_History is 0, their Loans are not approved. If applicant's ability to repay debts and demonstrated responsibility in repaying them is 0 , they will; find it difficult to get Loans approved.**
* **Most of the Customers from Semiurban got their Loans approved. But the Loans got Rejected more for clients from Urban Area.**

#### To understand more statistics, we will have to convert all the categorical data into numerical data

A screenshot of a computer

Description automatically generated

**10.2.2 Descriptive Statistics**

The describe method provides a summary of the Loan\_Amount column, including count, mean, standard deviation, min, max, and quartiles. This gives an idea of the spread and central tendency of the data.



### **Observations:**

* **Mean is greater than median for the columns, Self\_Employed, Loan\_Amount, Property\_Area, Total\_Income.**
* **For rest of the columns the the mean is lesser than median.**
* **There is different between 75% and Max for Dependents,Loan\_Amount, Loan\_Amount\_Term and Total\_Income and we will have to treat them.**

**10.2.3 Finding Correlation:**

**Correlation** is a statistical measure that describes the strength and direction of the relationship between two variables. It quantifies how much one variable changes in response to changes in another variable.

A screenshot of a graph

Description automatically generated

Finding Correlation through Heatmap

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

### **Observation:**

* **There is a moderate correlation between Loan\_Amount and total\_income, Credit\_history and loan\_Status.**
* **Apart from this, no much correlation found between variables.**

1. **Multivariate Analysis:**

**Multivariate analysis** refers to the statistical analysis techniques used to understand relationships between multiple variables simultaneously. Unlike univariate (single variable) or bivariate (two variables) analysis, multivariate analysis allows us to analyze the influence and interaction of more than two variables at once.

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

**After the completion of Analysis, we must find the Outliers in order to avoid overfitting/underfitting of data.**

# **Outliers:**

**Outliers** are data points that significantly differ from the majority of the data. They can be much higher or lower than the rest of the data points. Outliers can arise due to variability in the data, errors in data collection, or rare occurrences. It's crucial to identify and handle outliers as they can distort statistical analyses and model predictions.

**12.1 Finding Outliers**

A computer code on a white background

Description automatically generated

A graph of a graph of a graph

Description automatically generated with medium confidence

* **Insights : Loan\_Amount, Loan\_Amount\_Term and Total\_Income has lots of outliers and we will have to treat them**
  1. **Removal of Outliers with Z-score method:**

The **Z-score** method measures how many standard deviations a data point is from the mean of the dataset. It's a standardized way to identify outliers.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**After removal of Outliers the data is left with 581 rows**

* 1. **Removal of Outliers using IQR Method:**

The **IQR method** identifies outliers based on the spread of the middle 50% of the data. It's less sensitive to extreme values compared to the Z-score method.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

After removal of outliers we are left with 299 rows

* **Insights: Data loss is less in Z-Score Outliers removal method. Hence we will move further with Z-Score method.**

1. **Skewness:**

**Skewness** is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. In other words, it indicates the extent to which the data deviates from a normal distribution.

**13.1 Finding Skewness**

A screenshot of a computer

Description automatically generated

* 1. **Removal of skewness**

The **Power Transform Method** is a statistical technique used to stabilize variance, make the data more closely approximate a normal distribution, and reduce skewness.Using Power Transform Method to skew the columns 'Loan\_Amount', 'Loan\_Amount\_Term', 'Total\_Income', 'Credit\_History'

**A close-up of a computer screen

Description automatically generated**

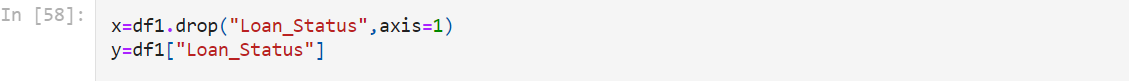
Checking the Skewness again

A screenshot of a computer

Description automatically generated

* **Insights: We have skewed that numerical columns and now its ready for scalarization**

# **Separating Features and Label/target variables:**



A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

# **Feature Engineering: using Standard Scaling Method (Normalization):**

**Feature engineering** is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy. One of the common techniques in feature engineering is **standard scaling** (also known as normalization), which transforms the data to have a mean of zero and a standard deviation of one. Let us move forward to scale the data in order to Normalize it.

A close-up of a computer code

Description automatically generated

A table with numbers and a blue line

Description automatically generated

1. **Checking Variance inflation factor(VIF):**

**Variance Inflation Factor (VIF)** is a measure used to detect multicollinearity in a set of multiple regression variables. Multicollinearity occurs when predictor variables are highly correlated, which can inflate the variance of the coefficient estimates and make the model unstable.

A screenshot of a computer

Description automatically generated

* In Standard Scaler technique we see that the vif values have minimized and are less than 5. Hence the multicollinearity issue is solved.

1. **Balancing the Target Variable:**

Let’s move forward to see if the Target variable is balanced or not.

A screenshot of a phone

Description automatically generated

We installed scikit-learn**,** imblearn to balance the target variable. Post Installation the target variable has been balanced.

A screenshot of a computer

Description automatically generated

Once the data seems to be Balanced, we now have to move further for Modelling or Model Development.

# **Modelling or Model Building and Cross\_valuation:**

# **Modelling**, also known as **model development**, is the process of creating a mathematical representation (model) of a real-world process or system using data. In the context of machine learning and data science, it involves building algorithms that can learn patterns from data and make predictions or decisions based on new, unseen data.

# We will have to call all libraries used for Model Development and Cross validating all the models.

A computer screen shot of a program

Description automatically generated

Before using Train-test-split method it is important to know the Best accuracy score with a certain random state. Hence following code is used

A screenshot of a computer program

Description automatically generated

This is how a best random state is found, so that same can be entered for all the classification Model and the find the best Model to train the data.

The data is divided into 80:20 ration for train:test respectively

A screenshot of a computer code

Description automatically generated

And this is how prediction of values can be done. We have used all the type of classification model like RandomForestClassifier, SVC(), GradientBoostingClassifier(), AdaBoostClassifier(), LogisticRegression(), BaggingClassifier(), ExtraTreesClassifier(), DecisionTreeClassifier(), KNeighborsClassifier(), to find the accuracy\_score, confusion\_matrix, classification\_report and cross validation score.

Later a dataframe has be prepared using

**A computer code with many text

Description automatically generated with medium confidence**



1. **Model Evaluation:**

Evaluate the performance of the models and select the best one based on accuracy, precision, recall, and F1-score.

Through this method we got all the score in a single data frame and we were able to analyse which is the best model.

A screenshot of a graph

Description automatically generated

From the above observation it is very clear that Extra Tree Classifier is the best Model. The accuracy score is 0.89 and the cross-validation score is 0.77.

1. **Hyper Parameter tuning:**

**20.1 Explanation:**

I**nitialize the Classifier**: Create an instance of ExtraTreesClassifier.

**Define the Hyperparameter Grid**: Specify a dictionary where the keys are the hyperparameters and the values are lists of settings to try.

A screen shot of a computer code

Description automatically generated

**Initialize RandomSearchCV**: Pass the classifier, parameter grid, and other parameters like cv (number of folds for cross-validation) to RamdomizedSearchCV.

***# Initialize RandomizedSearchCV with 5-fold cross-validation***

**random\_search = RandomizedSearchCV(estimator=ET, param\_distributions=param\_dist, n\_iter=100, cv=5, n\_jobs=-1, verbose=2, random\_state=39)**

**20.2 Fit the Model: Fit the model using random\_search.fit(X, y)**

A screenshot of a computer code

Description automatically generated

**20.3 Best Model Evaluation:**

**20.3.1 Print Best Parameters and Score**: Output the best hyperparameters found and the corresponding score.

**20.3.2 Evaluate the Best Model**: Use the best estimator to make predictions and evaluate the model's performance.

A screenshot of a computer program

Description automatically generated

These methods will help you find the best hyperparameters for your Extra Trees Classifier, improving the model's performance.

Finally, we have got the best Model with the best Parameters. Moving further a final Model has been made ready.

A close-up of a computer screen

Description automatically generated

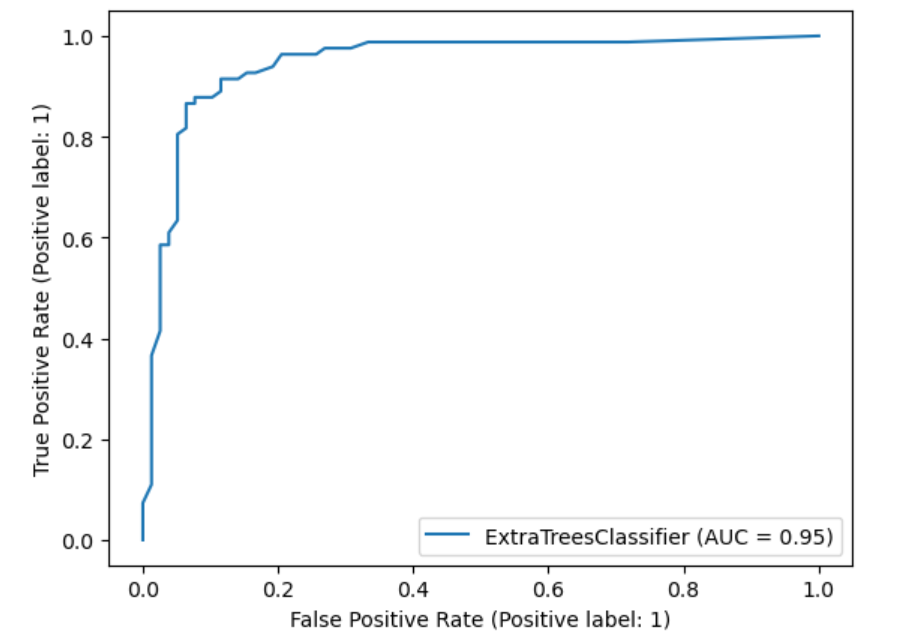
1. **ROC -AUC CURVE :**

The **ROC AUC curve** (Receiver Operating Characteristic - Area Under the Curve) is a graphical representation used to evaluate the performance of a binary classification model. It helps in understanding the trade-offs between the true positive rate (sensitivity) and the false positive rate (1-specificity) across different thresholds.

Now is the time to Plot the ROC AUC Curve to check if the Chosen model has given the right curve.

A screen shot of a computer program

Description automatically generated



A graph of a curve

Description automatically generated with medium confidence



A graph of a positive rate

Description automatically generated with medium confidence

A graph of a positive result

Description automatically generated with medium confidence

A graph of a positive result

Description automatically generated

A graph with a blue line

Description automatically generated

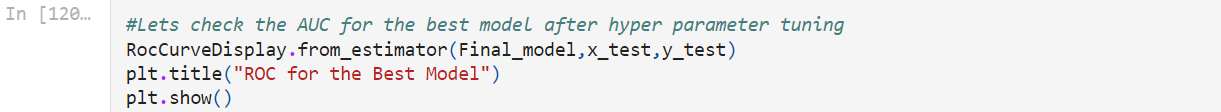
A graph with a blue line

Description automatically generated

A graph of a line

Description automatically generated

# **21.1 Plotting ROC and compare AUC for the Best Model:**



A graph of a model

Description automatically generated

**21.2 Plotting the ROC Curve for the Final Model and getting the AUC value:**

A close-up of a computer code

Description automatically generated

A graph with a line

Description automatically generated

#### Insights: Here we have plotted the ROC curve for the final model and the AUC value for the best model is 88%

1. **Saving the Model:**

A screen shot of a computer

Description automatically generated

1. **Predicting The Saved Model:**

Using the trained model to make predictions on new data.

A screenshot of a computer code

Description automatically generated

1. **Comparison:**

Comparing the present values with the predicted values.

A white background with black and white clouds

Description automatically generated with medium confidence

A screenshot of a number

Description automatically generated

# **We have predicted for 160 rows using the best suitable model Extra Tree Classifier**.