Report on Loan Eligibility Analysis

-R Madhupreetha, 110122084

1) The primary goal of this analysis is to identify the factors influencing loan eligibility decisions, develop different methods to predict the loan status of a given set of people.
2) As an extended task develop a model to predict the maximum loan amount the customer can obtain for the specified loan duration and minimum loan duration required for eligibility (if term ≤ 20 years)

Dataset

- **Number of Records in Training Dataset:** 614
- **Number of Features in Training Dataset:** 12
- **Target Variable:** Loan_Status (1 for eligible, 0 for not eligible).
- **Number of Records in Test Dataset:** 367
- **Number of Features in Test Dataset:** 11
- **Categorical features:** These features have categories (Gender, Married, Self_Employed, Credit_History, Loan_Status)
- **Ordinal features:** Variables in categorical features having some order involved (Dependents, Education, Property_Area)
- **Numerical features:** These features have numerical values (ApplicantIncome, Co-applicantIncome, LoanAmount, Loan_Amount_Term)

Exploratory Data Analysis (EDA)

Univariate Analysis

Visualized each variable separately using the Bar Graph.

It can be inferred from the bar plots that:

- 81% of applicants in the dataset are male.
- Around 65% of the applicants in the dataset are married.
- Around 14% of applicants in the dataset are self-employed.
- Around 78% of the applicants in the dataset are Graduates.
- Around 85% of applicants have repaid their doubts.
- Most of the applicants don't have any dependents.
- Most of the applicants are from the Semiurban area.
- Around 69% of the applicants have an approves Loan_Status.

Bivariate Analysis

Visualized each variable with respect to the target variable (Loan_Status) using the Bar Graph

It can be inferred from the bar plots that:

- Almost equal amount of male and female applicants has approved loan status.
- Almost equal amount of married and unmarried applicants has approved loan status.
- Graduates have a slightly more chance of approved loan status.
- Irrespective of whether they are self-employed or not the loan_status is approved.
- **ApplicantIncome:** Applicants with low income have lower loan approval rates, while approval rates increase with income, reaching 100% for the 'High' income group.
- **CoapplicantIncome:** Applicants with low coapplicant income have a higher loan approval rate, while those with very high coapplicant income have no approvals.
- **LoanAmount**: The loan approval rate is higher for moderate loan amounts and decreases for higher loan amounts, with 'Medium' and 'High' loan amounts showing the highest and lowest approval rates, respectively.
- **Loan_Amount_Term:** Loan approval rates are relatively consistent across 'Short', 'Medium', and 'Long' terms, with 'Very Long' loan terms showing a notable drop in approval rates.
- **Credit_History:** Applicants with 'Good' credit history are much more likely to get approved, while those with 'Bad' credit history have a very low approval rate.

Missing Value Handling

There are missing values in Gender, Married, Dependents, Education, Self_Employed, ApplicantIncome, CoapplicantIncome, LoanAmount, Loan_Amount_Term, and Credit_History features.

- **For numerical variables:** Filled the missing values using the linear interpolation method.

- **For categorical variables:** Filled the missing values the mode of the respective columns.

Converted the categorical columns to numerical for easy training purposes.

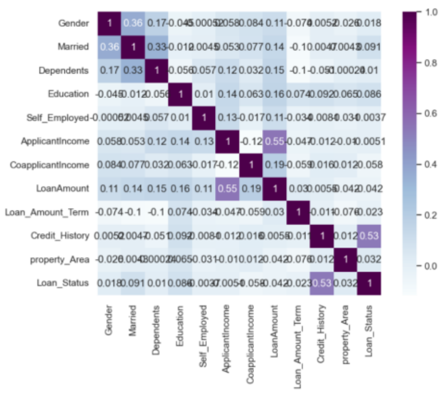
For example, replace Y values with 1 and N values with 0 and same for other Boolean types of columns.

Correlation

Next, find the correlation between all the numerical variables. Then used the heat map to visualize the correlation. Heatmaps visualize data through variations in colouring. The variables with darker colour mean their correlation is more.

- (**ApplicantIncome - LoanAmount) with correlation coefficient of 0.55**

- **(Credit_Hstory - Loan_Status) with correlation coefficient of 0.53**



Outlier Treatment

Due to these outliers' bulk of the data in the loan amount is at the left and the right tail is longer. This is called right skewness (or positive skewness).

One way to remove the skewness is by doing the **log transformation**. As we take the log transformation, it does not affect the smaller values much, but reduces the larger values. So, we get a distribution similar to normal distribution.

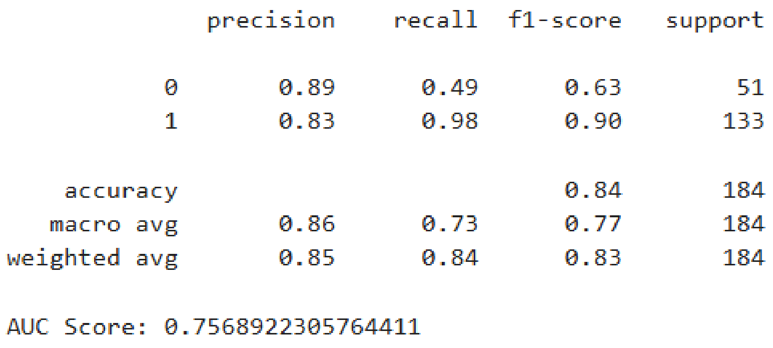
Model Development and Evaluation

Drop the Loan_ID variable as it does not have any effect on the loan status.

Did the same changes to the test dataset as for the training dataset.

First, we'll use the **Logistic Regression model** to predict the Loan_Status and evaluate it using the following metrics:

- → Accuracy
- → Recall
- → Precision
- → F1-Score
- → ROC Curve

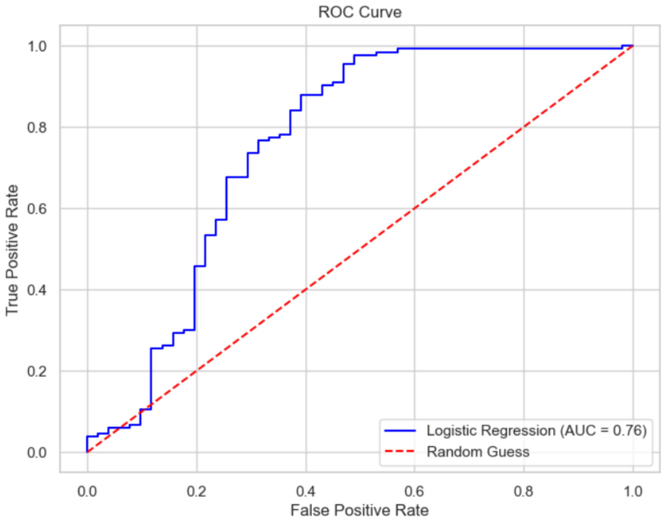


```
              precision    recall  f1-score   support

           0       0.89      0.49      0.63        51
           1       0.83      0.98      0.90       133

    accuracy                           0.84       184
   macro avg       0.86      0.73      0.77       184
weighted avg       0.85      0.84      0.83       184

AUC Score: 0.7568922305764411
```



Then performed **Hyper-parameter tuning using GridSearchCV.** It systematically tests combinations of hyperparameters to find the best configuration.

It uses cross-validation to evaluate the performance of each hyperparameter combination. This means the model is trained and validated multiple times on different subsets of the data, ensuring the results are not dependent on a single train-test split.

By fine-tuning hyperparameters, we can significantly improve model performance, ensuring that the model is not underfitting or overfitting.

This model is then evaluated using **R², MAE, MSE, RMSE, and plot for the residuals and actual vs. predicted values** is visualized.

This same process is followed for the following models:

- → **Random Forest**
- → **Gradient Boost**
- → **Decision Tree**
- → **k-Nearest Neighbour**

The predicted results are then **written into a csv file** with an extra column added and saved to a file named **'test_with_predictions.csv'**. This file is then used further for the 2nd task.

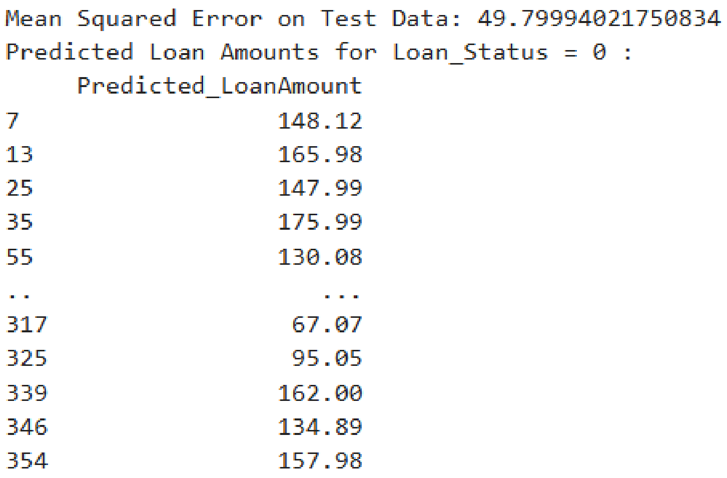
- ❖ We find that the Logistic Regression had the best accuracy of 84% for prediction among the above models after hyper-parameter tuning.

Task 2

Next, we use the **GradientBoosting Regression model** to predict the max Loan amount that the Not approved people will be able to get based on their Loan Duration.

Out of the 367 people in the test dataset, 61 people have not been approved the loan as found using the Gradient Boost model.

Using the GradientBoosting Regression model we are able to predict the max Loan which can be obtained by them.



```
Mean Squared Error on Test Data: 49.79994021750834
Predicted Loan Amounts for Loan_Status = 0 :
     Predicted_LoanAmount
7                  148.12
13                 165.98
25                 147.99
35                 175.99
55                 130.08
..                    ...
317                 67.07
325                 95.05
339                162.00
346                134.89
354                157.98
```

- → **Key features influencing the loan eligibility:** “Credit_score” has the maximum influence and the “Applicant_income” up to a certain extent.

Recommendations for improving loan eligibility

- → A strong credit score increases the likelihood of loan approval. Ensuring timely payments of credit card bills, utility bills, and EMIs can aid this process.
- → Higher income makes it easier to qualify for loans. Taking up part-time jobs, freelance projects, or other income opportunities can facilitate the loan approval process.
- → Opting for a longer loan term reduces the EMI, improving affordability and the likelihood of approval.