Report of Bank Loan Classification

Background:

In the banking sector, it is necessary and important to identifying potential loan applicants who are likely to accept loan offer is crucial . To modernize this process, a predictive models can be developed to classify customers based on their probability of accepting a loan offer.

Objective:

This model aims to develop predictive model which is capable of determine whether a loan applicant will accept a loan offer ,based on the applicants other attributes and historical data.

Data Description:

The dataset provided contains information on loan applicants, including demographic details, financial profiles, loan applicants history and whether they ultimately accepted the loan offer. Categorical variables and numerical variables were included in the dataset.

Approach:

The following approaches were used:

1)Data Preprocessing:

- The irrelevant ID column was dropped.
- Missing values in columns such as Gender, Income, Home Ownership, and Online were addressed using Simple Imputer.
- Special symbols in the Gender column were treated as noise and replaced.
- Missing values in numerical columns were replaced with the median, while categorical columns were filled with the mode.
- Missing values in the target column (Personal Loan) were replaced with the minority class value.

2) Exploratory Data Analysis (EDA):

- No duplicate rows were found.
- Unrealistic values and outliers in the Age column were replaced with the median within a reasonable range.
- Negative values in the Experience column were converted to positive values for interpretability.
- Correlation analysis revealed insights such as higher income and CD Account ownership correlating with loan acceptance.

3) Data Transformations:

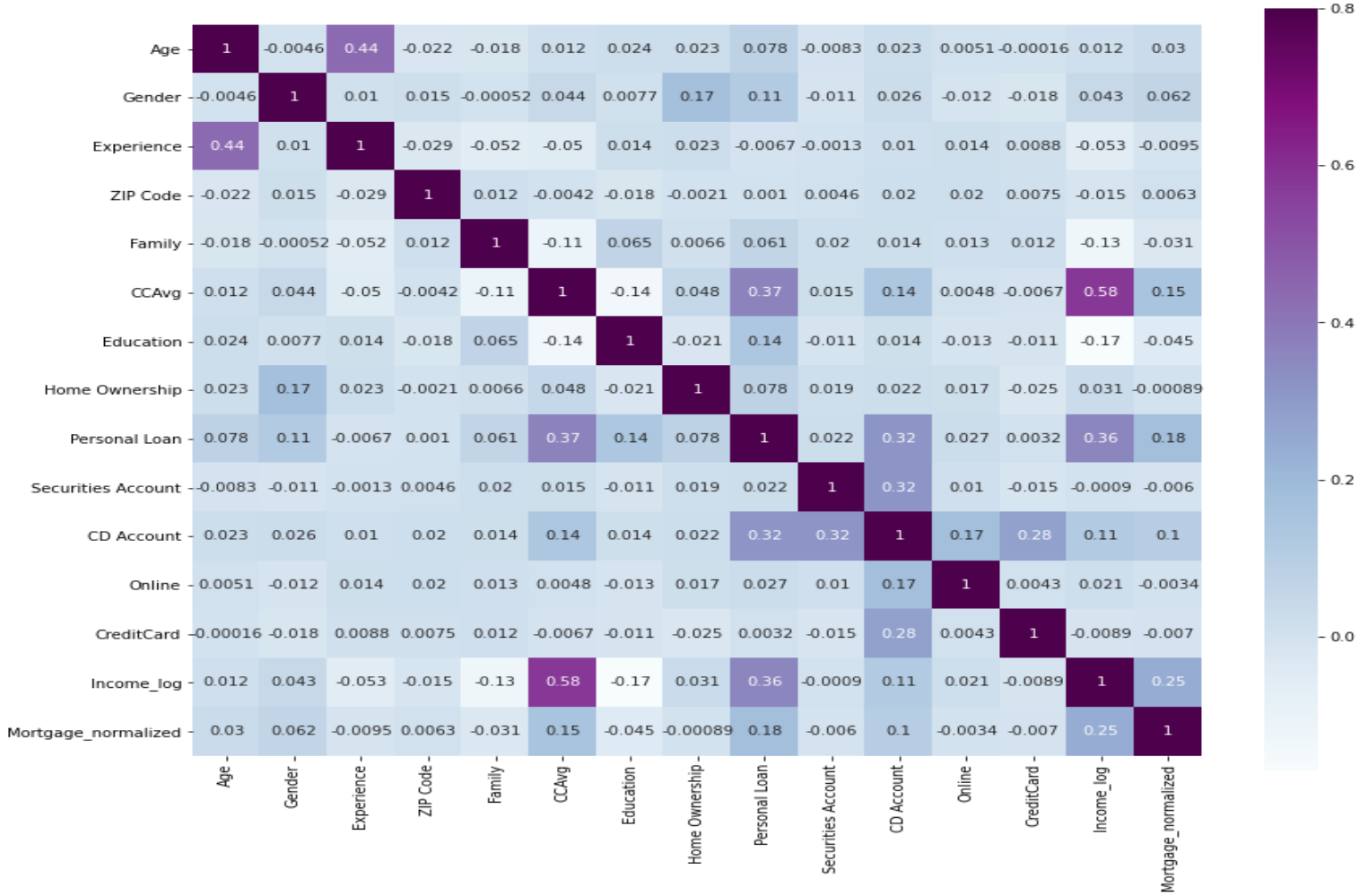
- The dataset was split into dependent (target) and independent (predictor) variables.
- A pipeline handled missing values and prepared the data for modeling.
- Categorical variables were one-hot encoded, and numerical features were standardized using StandardScaler.
- A stratified train-test split was employed due to the imbalanced nature of the target variable.

4) Training and Evaluation:

- Multiple algorithms (LogisticRegression, RandomForest, Adaboost) were trained and evaluated.
- Performance metrics such as accuracy, precision, recall, and f1 score were considered.
- RandomForest emerged as the top-performing algorithm based on the f1 score and was selected as the final model.

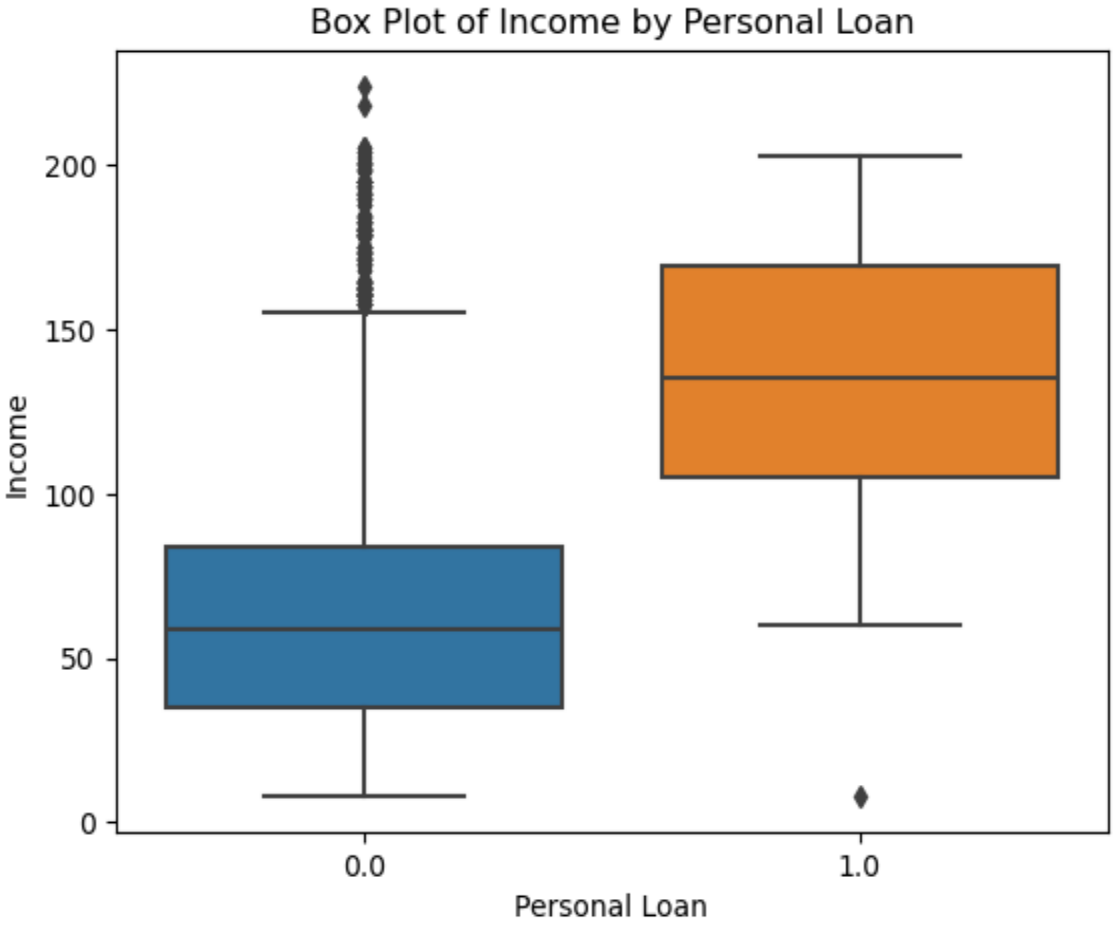
Key Insights:

1) Using Heatmap:
   From the correlation matrix, it looks like people with high income, high CCAvg and people with CD Account in the bank are likely to accept loan.



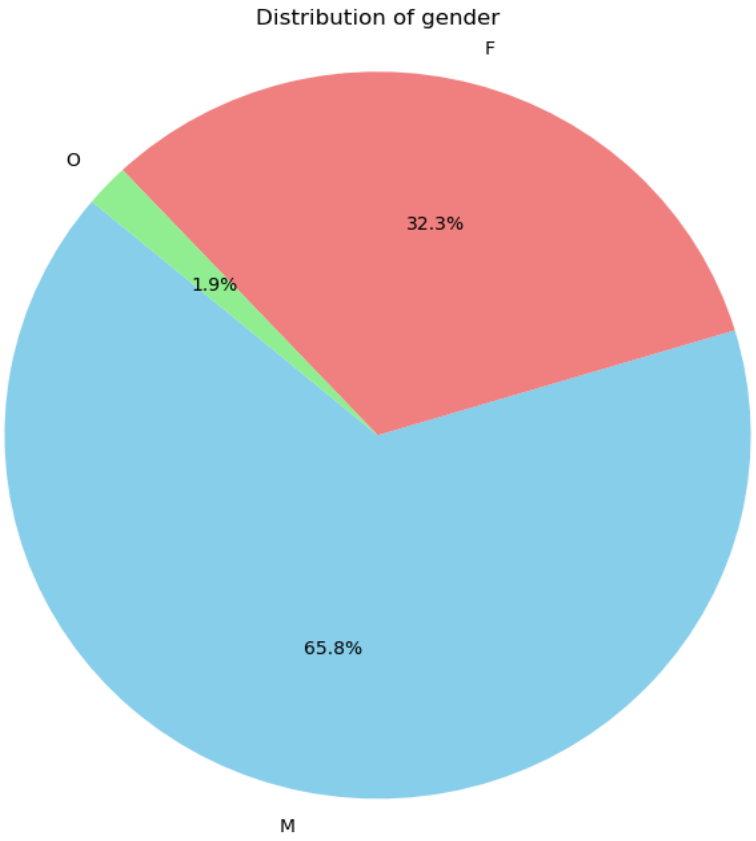
2)Box Plot:

From the box plot of Income and Personal Loan we can see that having more than 50000 income are likely to accept loans.



3)Pie chart:

From the below gender pie chart we can say that male percentage is purely dominant.



Thank You !