Report of Bank Loan Classification

Problem Description :-

The high-level task at hand is to provide personal loans to people efficiently and responsibly. This includes identifying the right individuals to approach and offering them suitable loan offers. By accurately classifying loan applicants, we aim to maximize loan approval for eligible people.

A bank loan dataset is used consisting of 5000 rows and 16 columns including the target column. The dataset consists of int, float and object data types. The technical task is to classify whether the personal loan was accepted or not based on the information provided.

Approach:-

1. Data preprocessing :-

The ID column which is irrelevant for the task was dropped from the dataset. The columns Gender, Income, Home Ownership, and Online had missing values. To address this issue, a Simple Imputer from the scikit-learn library was utilized for filling the missing values. For numerical columns, the missing values were replaced using the median value, while for categorical columns, the missing values were filled with the mode (most frequent category).
The Gender column contained special symbols, specifically '#' ' and '-', which were considered as noise. These symbols were replaced with the most frequent category in the Gender column.
As for the target column, Personal Loan, which had a missing value, it was replaced with the minority class value to ensure that no information was lost during the imputation process.

2. Exploratory Data Analysis :-

There were no duplicate rows in the dataset. However, the Age column contained some unrealistic values such as 978, 600, 0, and 2, which are considered outliers. To handle this issue, these outliers were replaced with the median age value, limited to a reasonable range between 20 and 100.
Additionally, the Experience column had some negative values, which do not have a meaningful interpretation. To address this, the negative values were converted to their positive counterparts, making the data more interpretable and suitable for analysis.
The correlation matrix revealed that individuals with higher income levels and higher average monthly credit card spending (CCAvg) were more inclined to accept personal loans. Additionally, those who had a CD (Certificate of Deposit) Account with the bank were also more likely to accept personal loans.

The dataset appeared to be male-dominated, with a majority of individuals being male. A significant proportion of respondents owned their own houses, indicating a high prevalence of home ownership.
Most individuals reported their income to be within the range of 10,000 to 80,000, while the majority of people tended to spend between 0 and 3 with their credit cards, suggesting relatively conservative credit card usage. Individuals with income levels exceeding 60,000 exhibited a higher likelihood of accepting personal loans, hinting at a potential correlation between higher income and loan acceptance.

3. Data Transformations :-

The dataset was divided into dependent (target) and independent (predictor) variables. A simple pipeline was then set up to handle missing values and prepare the data for modeling.
In this pipeline, missing values in the dataset were imputed using appropriate techniques, such as using the median for numerical features and the mode for categorical features. Categorical variables were one-hot encoded to convert them into a numerical format suitable for machine learning algorithms. Additionally, numerical features were standardized using the StandardScaler to ensure all variables had a similar scale, avoiding any bias during modeling.
As the dataset exhibited a highly imbalanced distribution of the target variable, a stratified train-test split was employed. This sampling technique ensures that the proportion of each class in the target variable is preserved in both the training and testing datasets. The final training dataset contained 4000 rows with 18 columns, and the testing dataset consisted of 1000 rows.

4. Training and Evaluation :-

The dataset was used to train multiple algorithms, including LogisticRegression, RandomForest, and Adaboost. During the evaluation process, various performance metrics such as accuracy, precision, and recall were considered. However, given the imbalanced nature of the data, the primary focus was on the f1 score, which provides a balanced assessment of both precision and recall.
After comparing the models' performances, Adaboost emerged as the top-performing algorithm, achieving the highest f1 score among all the tested models. As a result, Adaboost was selected as the final model to be used for making predictions and further analysis.

5. Future works :-

    a. Validation set :-

The model was developed without incorporating any validation sets. Introducing validation sets into the model development process could lead to improved model performance, increased robustness, and better predictions on new data.

b. Hyper parameter tuning :-

During the model development process, default hyperparameters were utilized for all algorithms without any fine-tuning. However, a few tweaks of these hyperparameters could potentially lead to better performance.

c. Feature engineering :-

Not much feature engineering was done in the dataset. With an in-depth analysis of the data, it may be possible to identify key patterns and relationships among the features. This could lead to the discovery of new features or transformations that could better represent the underlying patterns in the data.