Introduction

As part of the machine learning internship selection process for treeleaf, I undertook the task of developing a predictive model to classify whether a personal loan was accepted based on provided customer information. The dataset comprises 16 columns, consisting demographic details, financial information, and historical interactions with the bank. This report outlines the systematic approach taken to preprocess the data, perform exploratory data analysis (EDA), feature engineering, and ultimately construct a machine-learning model.

Data Loading and Initial Exploration

The data consisted of 5000 rows and 16 columns. Among the 16 columns given ‘Personal Loan’ was our target feature rest of them are independent feature. It was clearly a binary classification problem statement to check whether a person is eligible for taking personal loan from a bank or not.

Data Preprocessing

Upon initial exploration of the dataset, a significant class imbalance was identified, with a ratio close to 1:10 between the majority and minority classes. The highly imbalanced nature of the dataset could potentially lead to model biases. The RandomUnderSampler module from imblearn library was employed to systematically reduce the number of instances in the majority class, thereby addressing the disproportionate class distribution.

Before analysis we need to handle missing values to ensure analysis and modeling are based on a complete and accurate dataset. Some independent features had missing values in them. Here is the table showing the percentage of missing values in the dataset.

Independent Features	% of missing values
Gender	31.92
Income	1.34
Home Ownership	23.78
Online	0.80

Gender and Home Ownership have very high number of missing values whereas Income and Online has very less amount of missing values. Therefore, the approach to dealing with them will also be different.

Since Gender and Home Ownership have a high percentage of missing values, dropping all of them would not be suitable. Instead of dropping the Nan values I created another category for both by replacing the Null values with some other values. This will help the ml model in understanding the data.

Exploratory Data Analysis (EDA)

EDA was conducted to gain insights into the dataset's characteristics, relationships between variables, and potential patterns. Visualization techniques such as histograms, scatter plots, and correlation matrices were employed to uncover trends and outliers.

Key insights drawn from the analysis of numerical features:

- Some features, including “Family”, “education”, “credit card”, “cd account”, “online” and “securities account” contain discrete data.
- “Experience”, “Income”, “Age”, “CCAvg” and “Mortgage” exhibit continuous data.
- “Experience” has a minimum value of -3, indicating potential outliers as work experience cannot be negative.
- “Age” has unrealistic values such as 978, suggesting outliers in the age column.
- “Income”, “CCAvg” and “CD account” are identified as potentially more relevant to the target feature compared to other independent features.
- “Income”, “CCAvg” and “CD account” are highlighted as having a stronger association with the target feature compared to other independent features.

Similarly, after performing a detailed analysis of categorical features following insights were drawn:

- “Gender” column has 2 unknown classes other than M, F and O
- No. of males getting personal loan is higher in compared to others.
- People having Home Mortgage have more chance in getting personal loan than Home Owner or Rent.
- People having 3-4 family members are more likely to get personal loans, similarly people having bachelor’s or master’s degree are more likely to get loan and so on.

Feature Engineering

Feature engineering aimed to enhance the model's predictive performance by creating new features or transforming existing ones. ID and ZIP Code were removed from the dataset. Scaling of continuous numerical features (Age, Education, Income, CCAvg, Mortgage) and encoding of categorical features(Gender, Home Ownership) were performed in this step.

Model Selection and Training

I experimented with various machine learning model such as Random Forest, Logistic Regression, Support Vector Classifier, Naive Bayes and found out that Random forest was performing much better than other algorithms. Since the no. of data points in training dataset and random forest is quite robust to overfitting. This could be reason of random forest outperforming other algorithms with an accuracy of 97 percent.