Report on Bank Loan Classification Task

Introduction

For this analysis, we have a dataset containing comprehensive information about bank loan applicants. The ultimate goal is to construct a machine learning model capable of accurately classifying whether a personal loan application was accepted or not.

Data Preprocessing

The given dataset consists of 15 columns and a corresponding target column. Data Preprocessing is the most crucial step while constructing a model and consumes significant time. During this process, the following steps were performed:

- Checking missing values: After checking for missing values, 4 columns were found to have missing values.
- Handling missing values: The missing values were handled by dropping the rows for ‘Online’ and ‘Income’ column, by dropping the ‘Gender’ column as it had more than 30% missing values, and by filling missing values with most frequently occuring value i.e. mode for ‘Home Ownership’column.
- Handling outliers: While exploring the data outliers were discovered in age column which was filtered.
- Datatype Conversion: Categorical columns such as ‘Online’, ‘Income’ etc were converted from object or float data type to int.
- Multicollinearity: The columns ‘Age’ and ‘Experience’ was hight correlated to each other due to which ‘Experirnce’ column was dropped.

Exploratory Data Analysis

EDA was performed to gain insights from the data. The insights gained are:

- The majority of customers in the dataset possess a form of home ownership categorized as 'Home Mortgage.
- The education level of customers is mostly Bachelor level but there is significant amount of customers having Master as well as Advanced degree.

- The customers in this dataset do not heavily rely on credit cards.
- The majority of customers do not have Security account.
- The majority of customers do not have CD account.
- There is significant amount of customer who uses internet banking facilities along with customers who do not.
- Mostly this dataset consists of significant amount of customers whose personal loan is not accepted.
- The age of customers in this dataset widely range between 20-70, with most people falling between 40-60.

Feature Scaling

Some features such as Age, Income were scaled by using Standard Scaler, so that the machine learning models could interpret these features on the same scale.

Model Training and Evaluation:

For personal loan dataset, different models were considered such a Logistic Regression, Decision Tree, Random Forest Classifier and SVM. After hypertuning the parameter, it was concluded that Decision Tree performs well among other models with accuracy of about 98.6 %.

Hence, the model was trained using Decision Tree Classifier and saved so that it could be used for prediction.

Key Finding:

The factor that influenced the target variable i.e. Personal Loan was the Income variable. Another factor that influenced the target variable to some extent is the Home Ownership variable.

Conclusion:

Conclusively, in the process of constructing this model, the 'Gender' column was excluded due to its substantial number of missing values and the presence of uninterpretable entries. Given additional time and comprehensive data, a deeper investigation could be undertaken to ascertain whether this feature bears any noteworthy influence on the target variable.