TREELEAF TECHNOLOGIES PVT. LTD

Task 1 Bank Loan Classification Report

Introduction:

The code performs a classification task on a bank loan dataset using various machine learning models. The code loads a dataset from an Excel file called "NewBank_loan_data.xlsx" into a pandas DataFrame df.

Data Preprocessing:

- More loans are rejected than accepted. Data is imbalanced.
- Missing values are handled for the 'Income', 'Home Ownership', and 'Online' columns using the mode value for each column. The row of missing value for Personal Loan was dropped.

Exploratory Data Analysis (EDA):

- Various exploratory data analysis (EDA) operations are performed, such as checking the shape of the DataFrame (df.shape) which has output of (5000, 16) i.e. 5000 rows and 16 columns, data types of columns (df.dtypes), checking for missing values (df.isna().sum()), and displaying summary statistics (df.describe()).

Here,

the correlation between Income and CCAvg is 0.62,
the correlation between Income and Mortgage is 0.19,
the correlation between Income and CD Account is 0.13,
the correlation between CCAvg and Mortgage is 0.11,
the correlation between CCAvg and CD Account is 0.14,
the correlation between Securities Account and CD Account is 0.32,
the correlation between CreditCard and CD Account is 0.28.

Feature Engineering:

The target variable 'Personal Loan' is encoded into numerical values using LabelEncoder(), while categorical columns are one-hot encoded using pd.get_dummies().

Model Selection and Training:

- The dataset is split into training and testing sets using train_test_split().
- Four machine learning models are trained and evaluated: Logistic Regression, Support Vector Machine (SVM), Naive Bayes and Decision Tree Classifier.
- Evaluation metrics such as precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve(AUC-ROC) are calculated.

- The Decision Tree Classifier is selected as the final model based on the highest F1-score and AUC-ROC.
- The selected model is saved using joblib.dump() to a file named 'bank_loan_classification.pkl'.

Key Findings:

- The ID column holds a unique identifier for each customer and does not provide any useful information for predicting whether a personal loan was accepted.
- The Experience column, when plotted in boxplot, shows that the customers have 10-30 yrs of experience and the customers of those age groups equally get and not get loans.
- There is no significant difference if the customers are Male or Female or others get the loan.
- ZIP Code is not useful for prediction.
- The correlation between Securities Account and CD Account is 0.32 and the correlation between CreditCard and CD Account is 0.28. Hence, keeping the CD Account column and dropping others.
- The Income column is highly relevant as it directly affects the likelihood of loan acceptance. And other columns are also relevant to predict loan acceptance.

	Precision score	Recall score	f1 score	auc_roc
Logistic Regression	0.7592	0.4712	0.5815	0.7285
SVC	0.65	0.1494	0.2429	0.5708
Naive Bayes	0.4	0.5057	0.4467	0.7167
Decision Trees	0.9493	0.8620	0.9036	0.9288

Decision Trees
Precision Score:
Of all the customers predicted to accept the loan, 95% actually did. This high precision indicates that the model is very good at not misclassifying customers who did not accept the loan as having accepted it.

Recall Score:

The model correctly identified 86% of the customers who actually accepted the loan. This indicates that the model is good at capturing most of the customers who accepted the loan.

F1 Score:
The F1 score of 90% indicates a good balance between precision and recall. This suggests that the model performs well in identifying customers who accepted the loan while maintaining a high level of precision.

AUC-ROC Score:
The AUC-ROC score of 93% indicates excellent model performance in distinguishing between customers who accepted and did not accept the loan.

Conclusion:
In conclusion, Decision trees model was chosen for its high f1 score and AUC-ROC score.