Loan Status Prediction

Problem statement

The objective is to build a Machine Learning Model to predict the loan to be approved or to be rejected for an applicant. The project involves data preprocessing, feature engineering, model training, and evaluation to achieve optimal prediction accuracy.

Project Steps

Importing Libraries and Loading the Dataset

- Imported necessary libraries such as pandas, numpy, sklearn, and others required for data processing and model building.
- Loaded the dataset into a pandas DataFrame.

Handling Missing Values

- Checked for missing values in the dataset.
- Handled missing values using SimpleImputer:
    1. For numerical variables, replaced missing values with the mean of the column.
    2. For categorical variables, replaced missing values with the most frequent category.

Encoding Variables

- Encoded categorical variables using OneHotEncoder to convert them into numerical format.
- Encoded the target variable using LabelEncoder for binary classification.

Splitting the Dataset and Training Models

- Split the dataset into training and testing sets.
- Trained the model using different algorithms:
    1. Logistic Regression
    2. Decision Tree
    3. XGBoost
    4. Random Forest

Model Evaluation

- Evaluated the performance of each model:
- Logistic Regression: 78%
- Decision Tree: 68%
- XGBoost: 75%
- Random Forest: 77%

Further Model Training and Feature Engineering

- Created new features for the dataset to improve model performance.

Preprocessing Pipelines

- Developed preprocessing pipelines for numerical and categorical data.
- Combined preprocessing steps into a single pipeline.

Model Definition and Training

- Defined a model using the Random Forest Classifier.

Creating and Evaluating the Pipeline

- Created a pipeline that includes preprocessing and model training steps.
- Split the data into training and testing sets.

Hyperparameter Tuning

- Defined a parameter grid for GridSearchCV.
- Used GridSearchCV to find the best model parameters.
- Evaluated the best model on the test set, achieving an accuracy of 77% to 80%.

Making Predictions and Saving Results

- Loaded the dataset and handled missing values.
- Created new features for the dataset.
- Made predictions on the test data.
- Added the predictions to the test DataFrame.
- Saved the updated test DataFrame with predictions to a new CSV file.

Conclusion

This project successfully demonstrates the process of building a machine learning model to predict loan status. By following systematic steps, including data preprocessing, feature engineering, model training, evaluation, and hyperparameter tuning, we achieved a final model accuracy between 77% and 80%.