Loan Data Analysis and Prediction Documentation

**Introduction:**

This document describes a Python code that analyzes loan data and predicts whether loans will be paid or defaulted. The code utilizes various libraries for data manipulation, visualization, and machine learning techniques.

**Data Loading and Exploration:**

- The code begins by loading loan data from a CSV file named `train\_loan\_data.csv` into a DataFrame.

- It displays the shape of the dataset and the first few rows to give an overview of the data.

- Additionally, it checks for missing values in the dataset.

**Data Preprocessing:**

- Missing values are handled by dropping rows with any null values.

- Categorical values in the 'loan\_status' column are converted into numerical values. 'Defaulted' is assigned 0, and 'Paid' is assigned 1.

**Target Encoding:**

- Categorical columns are encoded into numerical values using target encoding technique, where each category is replaced with the mean of the target variable.

**Data Visualization:**

- Various visualizations are generated to understand the distribution of loan status and numerical features.

- Bar plots and histograms are used to visualize the distribution of loan status and numerical features like annual income, FICO scores, etc.

**Model Building and Evaluation:**

- The dataset is split into training and testing sets.

- A Random Forest classifier is trained on the training data.

- The trained classifier is used to predict loan status on the testing data.

- The accuracy of the model is calculated and displayed.

**Model Saving:**

- The trained Random Forest model is saved using the joblib library for future use.

**Conclusion:**

This code provides a comprehensive analysis of loan data, including preprocessing, visualization, modeling, and prediction. The trained model can be used to make predictions on new loan dataeffectively.





