**Loan Default Prediction Project Documentation**

1. **Introduction**

This project aims to predict loan default risk using a dataset containing demographic and financial information about borrowers. The objective is to provide insights into which factors contribute to loan defaults and develop a predictive model that accurately assesses the risk associated with loan applications.

1. **Methodology**
   1. **Data Loading and Basic Exploration**

**Data Loading:** The dataset was loaded into the Python environment using Pandas, with the DataFrame created under the name `loan\_df`.

**Initial Analysis:** Pandas functions like `info()` and `describe()` were used to analyze the dataset, detect null values, and understand basic features such as data types and statistical summaries.

* 1. **Data Cleaning**

**Handling Missing Data:**

**Gender:** Missing values in the `Gender` column were filled alternatively since the gender distribution was almost equal.

**Employment Status:** Missing values in the `Employment Status` column were filled based on the `Income` level, assuming a correlation between income and employment status.

**Feature Engineering:** New columns were created to enhance model predictions:

Age Group

DTI (Debt-to-Income) Group

Credit Score Group

Loan-to-Income Ratio

Interest-to-Income Ratio

EMI (Equated Monthly Installment)

**Outliers:** 10 outliers were detected in the `Income` column. However, removing them reduced model accuracy, so the outliers were retained.

* 1. **Data Encoding and Normalization**

**Label Encoding:** All non-numeric columns were converted to numerical format using Label Encoding.

**Normalization:** The dataset was normalized using One-Hot Encoding for categorical variables and MinMax Scaler for numerical variables.

* 1. **Feature Selection**

**Feature Importance:** After assessing feature importance, less important features were initially removed, but this led to a significant reduction in model accuracy. As a result, all features were retained.

* 1. **Data Splitting and Class Imbalance**

The dataset was split into training and testing sets: `X\_train`, `X\_test`, `y\_train`, and `y\_test`.

Class Imbalance: The target variable showed an imbalance, which was addressed using SMOTE (Synthetic Minority Oversampling Technique) to balance the dataset.

* 1. **Model Training and Evaluation**

**Models trained:**

Logistic Regression

Decision Tree

Random Forest

Support Vector Machine (SVM)

Using default parameters, Random Forest provided the best accuracy and ROC-AUC score.

**Hyperparameter Tuning:** Optimal parameters were identified through tuning, and the models were retrained. Random Forest remained the top-performing model.

* 1. **Model Saving**

The final Random Forest model was saved for future predictions based on user input, ensuring the most accurate model is stored for use.

1. **Insights and Recommendations**
   1. **Demographic Analysis**

Middle-aged Borrowers (30-50 years old) have a relatively low default rate (18%) and can be targeted for loan approvals.

Seniors (50-65 years old) show a higher default risk (21%), suggesting the need for caution.

Elderly (65+ years) are not accessing loans, so they can be excluded from loan advertisement targeting.

* 1. **Credit Score Analysis**

Borrowers with Poor Credit Scores (125-580) have a low default rate (16%) and are suitable for loan consideration.

Fair and Excellent Credit Scores present higher default risks (24% and 23%, respectively), requiring additional scrutiny.

* 1. **Debt-to-Income (DTI) Analysis**

Very Low DTI Scores (0.0-0.2) have a high default rate (22%), necessitating caution.

Moderate DTI Scores (0.4-0.6) also show a high default risk (21%).

* 1. **Gender Considerations**

Female borrowers have a higher default rate (21%) compared to males (17%), indicating the need for careful assessment.

* 1. **Employment Status**

Employed individuals exhibit a slightly higher default rate (20%) compared to unemployed individuals (18%).

* 1. **Loan Amounts**

Loan defaults are distributed nearly equally across all loan amounts, with more loans concentrated in the INR 25,000 - INR 45,000 range. The bank should focus on this loan range while maintaining vigilance due to the default distribution.

1. **Conclusion**

The project successfully developed a predictive model using Random Forest, which showed the best performance after hyperparameter tuning. The model provides actionable insights to help the bank make informed lending decisions by targeting lower-risk borrowers and identifying potential high-risk profiles.

Thank you for reviewing this documentation. Your engagement is highly appreciated!