Name: Ahetasham Shifat
E1 ID:
Wings: Intelligence & Perception, General, Systems & Application

Project 01: AI-Powered Loan Default Prediction System for Banking and Fintech

-Data Processing and Feature Engineering (week02)

Dataset: Loan Default Dataset [Kaggle]

1. Exploratory Data Analysis (EDA)

- The dataset contains **255,347 records** and **18 features**, including demographic, financial, and loan-related information.
- **Target variable:** Default (binary: 0 = no default, 1 = default).
- **Key observations:**
  - The dataset has a mix of **numerical** (Age, Income, LoanAmount, CreditScore, InterestRate, DTIRatio, etc.) and **categorical** features (Education, EmploymentType, MaritalStatus, LoanPurpose, etc.).
  - The target variable distribution shows **class imbalance**, with more non-default cases.
  - Numeric features are within expected ranges; correlation heatmap indicates relationships between CreditScore, RiskScore, InterestRate, and LoanAmount.

2. Missing Values and Duplicates

- **Null values:** No missing values were found across the dataset.
- **Duplicates:** Checked and **no duplicate rows** detected.

3. Dropped/Irrelevant Features

- LoanID was dropped as it is a unique identifier with no predictive value.

3.1 Feature Relevance Considerations

I noticed that some features such as **Age, MaritalStatus, and Education** may not seem directly predictive of loan default. However, I have decided to **keep them initially** to see how they impact the model's performance.

- **Age**- Could reflect financial maturity and repayment behavior.
- **MaritalStatus**- Might indicate household structure and shared income.
- **Education**- May serve as a proxy for earning potential or financial literacy.

Since I am not certain about their predictive power, I plan to analyze feature importance after training and decide later whether to retain or remove them.

ML Models that can be used in this case:

- **Tree-based models** such as Random Forest, XGBoost, or LightGBM, to capture nonlinear interactions and feature importance.
- Explainable ML using SHAP
- **Logistic Regression**, for a baseline and interpretable model.
- **Neural Networks**, to explore potential nonlinear relationships across features.

4. Feature Engineering

- **Binary columns** (HasMortgage, HasDependents, HasCoSigner) converted to numeric (0/1).
- **New features created:**
  - DTI_Level: Debt-to-Income ratio categorized into Low, Medium, High, Very High.
  - AgeGroup: Age categorized into Young, Adult, Middle Age, Senior.
  - RiskScore: Combined metric using CreditScore and InterestRate to indicate financial risk.

5. Preprocessing Summary

- Numeric and categorical features identified for scaling and encoding.
- Preprocessing pipeline created with:
  - **StandardScaler** for numeric features
  - **OneHotEncoder** for categorical features
- The dataset is now ready for **model training** (train-test split applied, target separated).

Conclusion:
The dataset has been successfully cleaned, preprocessed, and enhanced with engineered features, making it suitable for building predictive ML models for loan default prediction.

Google collab:

https://colab.research.google.com/drive/1HiG5O07c0dys3yGu00RAGRakBHqzWDWu?usp=sharing