Model Research: Development of AI-Powered Loan Eligibility Advisory System

This document outlines the research and technical choices for the AI-Powered Loan Eligibility Advisor, focusing on open-source tools to build a transparent and scalable system.

1.Problem Understanding:

Loan eligibility is a binary classification task to predict whether a loan request should be approved or rejected. Key challenges include imbalanced data (fewer rejections than approvals) and the regulatory need for explainable AI. The system must also be scalable for real-time inference.

2. Dataset Research:

All datasets considered are open-source, including the German Credit Data, Loan Prediction Dataset, and Credit Approval Dataset. The primary features used for analysis are income, employment status, credit score, loan amount, and loan history.

3. Model Candidates:

All model candidates were chosen for their balance of performance and interpretability

- **Logistic Regression (scikit-learn):** Simple and interpretable, but may struggle with non-linear patterns.
- **Random Forest (scikit-learn):** A robust ensemble model with good accuracy, less prone to overfitting than a single decision tree.
- **LightGBM:** Optimized for speed and memory efficiency, particularly effective with large datasets and high-dimensional features; like XGBoost, it requires careful parameter tuning.
- **XGBoost:** Known for high accuracy, efficiency, and handling missing values, though it requires careful tuning.

4. Model Evaluation & Explainability:

Due to the high cost of incorrect predictions, evaluation goes beyond simple accuracy. We can use a comprehensive set of metrics including Accuracy, Precision, Recall, F1-score, and ROC-AUC to measure model performance.

- **Explainability:** To ensure transparency, SHAP and LIME libraries can be used for global and local explanations of model predictions. Visualizations can be created using Matplotlib and Seaborn.

5. Chosen Model & Integration:

XGBoost is selected as the primary model due to its superior performance. Logistic Regression will serve as a baseline, and Random Forest as a backup.The application can be built entirely with open-source tools for each stage of the pipeline:

- **Data Processing:** Pandas and NumPy
- **ML Training & Inference:** Scikit-learn and XGBoost
- **Explainability & Visualization:** SHAP, LIME, Matplotlib, and Seaborn

- **Report Generation:** FPDF
- **Chatbot:** Hugging Face Transformers
- **Web App & Deployment:** Streamlit

6. Expected Outcomes:

The final outcome is an AI-powered web application that automates loan eligibility analysis. This system will predict approval likelihood based on user-provided financial information and generate transparent, PDF-based decision summaries. By integrating a financial guidance chatbot, the platform will assist users in understanding credit factors, thereby improving transparency and access. The solution will empower both applicants and lenders with real-time, explainable insights**,** while also incorporating fairness checks to detect and mitigate potential biases, ensuring equitable and responsible decision-making.