AI-Powered Default Risk Predictor for FinTech

(Financial Services) Capstone 2 Project Proposal (By - Altamash Ansari)

1. Executive Summary

In the dynamic landscape of financial services, predicting loan default risk is pivotal for sustainable growth and risk management. This project aims to develop an interactive dashboard using Streamlit, powered by machine learning models, to predict whether a loan applicant is likely to default. By leveraging applicant-level insights, feature engineering, and explainable AI (e.g., SHAP), the dashboard will enable credit risk analysts to forecast default probabilities, visualize key risk factors, and optimize loan approval processes. This solution bridges traditional credit assessment with data-driven intelligence, enhancing decision-making and reducing financial exposure.

2. Problem Statement

Financial institutions manage vast portfolios of loan applications, yet often lack a unified tool to assess default risk accurately in real-time. Without integrated visibility into applicant behavior, credit history, and economic factors, misjudgments lead to loan defaults, impacting profitability and regulatory compliance. This project addresses the critical need for a predictive, visual decision-support tool that empowers stakeholders to track risk patterns, evaluate applicant profiles, and mitigate potential losses effectively.

3. Data Sources

Two primary datasets are utilized in this project:

- **Loan Status Prediction.csv ➤** Contains detailed applicant data including gender, marital status, dependents, education, income, loan amount, credit history, property area, and loan status (default indicator).
- **Processed Data (Derived) ➤** Enriched version with engineered features: debt-to-income ratio (DTI), loan-to-income ratio, log-transformed loan amounts, and binary flags for default risk assessment. These datasets are merged and transformed using applicant identifiers (e.g., Loan_ID) as the primary key, with additional preprocessing to handle missing values and encode categorical variables.

4. Methodology

❖ **Data Integration & Feature Engineering:** The raw dataset is cleaned, merged, and enhanced with derived metrics such as DTI, loan-to-income ratio, and log-transformed features to enable granular risk analysis. Class imbalance is addressed using SMOTE to ensure robust model training.

❖ **Interactivity & Filtering:** Users can filter insights by applicant demographics (e.g., gender, education), income brackets, property area, loan amount ranges, or credit history status. Drill-down capabilities allow exploration of individual applicant risk profiles.

❖ **Visualization Strategy:** The dashboard will feature:

- Default risk distribution across applicant segments
- Probability heatmap linking income, loan amount, and default outcomes
- SHAP-based explanation of key risk factors (e.g., credit history, DTI)

- Time-series trends of default rates by property area or income level
- Impact of feature interactions on default probability

5. Expected Outcomes

1. A comprehensive view of default risk across all applicant profiles
2. Accurate default probability forecasting using segment-level trends
3. Identification of high-risk applicant segments and bottlenecks
4. Strategic segmentation of loans by income, property area, and credit history
5. Enhanced ability to prioritize low-risk loan applications
6. Data-backed decision support for credit risk management and policy formulation

6. Tools and Technologies

- **Python** – For data preprocessing, feature engineering, and model development
- **Scikit-learn, XGBoost, LightGBM** – For building and training ML models
- **SHAP** – For explainable AI and feature importance analysis
- **Streamlit** – For interactive dashboard design and deployment
- **Pandas, NumPy** – For data handling and manipulation
- **Matplotlib, Seaborn** – For visualization of EDA and insights
- **Joblib** – For saving and loading trained models

7. Risks and Challenges

- **Imbalanced Default/Non-Default Labels:** Uneven class distribution may skew predictions if not normalized properly (mitigated with SMOTE).
- **Missing Data:** Some applicants may lack consistent income or credit history data, affecting model reliability.
- **Feature Sparsity:** Certain combinations (e.g., property area + income level) may have low frequency, impacting the robustness of segment-level insights.

8. Conclusion

The **AI-Powered Default Risk Predictor for FinTech** bridges the gap between traditional credit risk assessment and modern data-driven decision-making. It enables financial institutions to proactively monitor loan portfolios, predict defaults with greater accuracy, and streamline approval processes. For a data science professional with a commerce background, this project demonstrates how analytical intelligence can directly translate into measurable business impact, enhancing risk management and profitability in the financial sector.