Proposal: Loan Default Prediction Model

Introduction

Loan default prediction is a key challenge in the financial sector. Lenders must carefully balance two competing risks: financial loss from defaults if loans are approved for high-risk applicants, and missed opportunities if creditworthy applicants are unfairly rejected. The objective of this project is to develop a reliable AI-based predictive model to help identify high-risk applicants while ensuring fair access for genuine borrowers.

Objectives

1. **Data Preparation & Cleaning** – Handle missing values, duplicates, and outliers to ensure high data quality.
2. **Exploratory Data Analysis (EDA)** – Understand distributions, correlations, and relationships in the data to identify trends distinguishing defaulters from non-defaulters.
3. **Feature Engineering** – Create meaningful derived features such as debt-to-income ratio, credit utilization, and credit history aggregates.
4. **Predictive Modeling** – Build, train, and evaluate machine learning models to classify loan applicants as high or low risk.
5. **Risk Analytics & Insights** – Identify the main factors influencing loan defaults and provide business insights for risk management.
6. **Decision Support** – Generate probability-based risk scores to guide loan approval decisions and maintain a balance between minimizing defaults and not rejecting creditworthy applicants.

Project Approach

1. Data Collection & Understanding

The dataset includes applicant demographic and financial details, previous loan applications, and loan outcomes. This data will serve as the foundation for modeling borrower behavior and risk.

2. Data Cleaning & Preparation

• Identify and handle missing or inconsistent values.
• Detect and treat outliers.
• Encode categorical variables and normalize numerical data.
• Split data into training, validation, and test sets.

3. Exploratory Data Analysis (EDA)

• Visualize key variables to understand relationships with default probability.
• Examine correlations and variable distributions.
• Identify significant predictors of default.

4. Feature Engineering

- Create domain-specific variables such as:
- Debt-to-income ratio.
- Number of previous applications.
- Ratio of approved to rejected past loans.
- Credit utilization percentage.
- Apply dimensionality reduction or feature selection techniques where needed.

5. Model Development

- Test multiple machine learning models.
- Evaluate models using metrics such as Accuracy, Precision, Recall, F1 Score, and ROC-AUC.
- Perform hyperparameter tuning for optimal performance.
- Calibrate probabilities to ensure reliable risk scoring.

6. Model Interpretation & Insights

- Use feature importance to interpret model predictions.
- Provide actionable insights into which factors most influence defaults.

7. Model Deployment & Decision Support

- Build a scoring system that classifies applicants as **High Risk** or **Low Risk**.
- Deliver probability-based risk scores to guide approval or rejection decisions.
- Document model performance, data pipelines, and risk thresholds.

Deliverables

- Cleaned and preprocessed dataset.
- EDA report with key insights and visualizations.
- Trained and validated machine learning model.
- Risk scoring system with threshold recommendations.
- Technical documentation and model explanation report.

Tools & Technologies

- **Programming:** Python
- **Libraries:** pandas, scikit-learn, XGBoost, LightGBM, matplotlib, SHAP
- **Environment:** Jupyter Notebooks
- **Version Control:** Git / GitHub

Timeline (Estimated 6–8 Weeks)

1. **Week 1:** Data understanding and cleaning.
2. **Week 2:** Exploratory Data Analysis.
3. **Weeks 3–4:** Feature engineering and model training.
4. **Weeks 5–6:** Model validation and optimization.

5. **Week 7:** Risk analytics and interpretation.
6. **Week 8:** Deployment preparation and documentation.

Expected Outcomes

- Improved accuracy in identifying high-risk loan applicants.
- Reduction in financial losses due to defaults.
- Better decision support for loan approval processes.
- Actionable insights into key risk drivers.

Conclusion

By implementing this loan default prediction model, the company will gain a reliable AI-driven tool to evaluate applicants, minimize default risks, and improve overall lending decisions. The combination of advanced analytics, domain-specific features, and explainable AI ensures that the model aligns with both financial and ethical business objectives.