Project Action Plan

Project Title:

Loan Default Risk Prediction

Credit Risk Analytics using Machine Learning

Prepared By:
Omkar Bhalekar
Email: omkarbhalekar2003@gmail.com

Project Goals and Success Criteria

The primary goal of this project is to develop a supervised machine learning model to predict loan default risk using historical lending data. The project aims to combine credit risk analytics with statistical validation and machine learning modeling.

Success Criteria:

- Build a clean and leakage-free dataset for modeling.
- Achieve meaningful predictive performance (ROC-AUC ≥ 0.70).
- Demonstrate improvement in default detection through threshold optimization.
- Provide business-oriented interpretation of model outputs.

Dataset Understanding and Feature Categories

The dataset consists of borrower-level and loan-level attributes. Features are categorized as follows:

Numerical Features:

- Loan Amount
- Interest Rate
- Annual Income
- FICO Score (fico_range_low)
- Debt-to-Income Ratio (DTI)
- Installment Amount

Categorical Features:

- Loan Grade
- Loan Term
- Home Ownership Status

Tools, Libraries, and Environment Setup

Programming Language: Python 3.11

Development Environment: Jupyter Notebook

Libraries Used:

- pandas, numpy (data manipulation)
- matplotlib, seaborn (visualization)
- scikit-learn (modeling and evaluation)
- joblib (model persistence)

Exploratory Data Analysis Strategy

The EDA phase focuses on understanding default patterns and validating financial risk drivers.

Key Analytical Steps:

- Compute default rate distribution.
- Analyze default rates across loan grades and interest rates.

• Examine relationship between FICO score and default probability.
• Perform hypothesis testing to validate statistical significance.

Data Preprocessing and Feature Engineering Plan

• Remove active loans to prevent data leakage.
• Handle missing values using median imputation.
• Encode categorical variables using One-Hot Encoding.
• Standardize numerical features using StandardScaler.
• Perform stratified train-test split (80/20).

Model Selection and Evaluation Approach

Logistic Regression will serve as the baseline model due to its interpretability and suitability for binary classification. Model performance will be evaluated using confusion matrix, precision, recall, F1-score, ROC-AUC, and Precision-Recall analysis.

Threshold optimization will be applied to balance risk detection and approval rates from a credit risk management perspective.

Project Timeline and Milestones

Phase	Description
Phase 1	Dataset acquisition and understanding
Phase 2	Data cleaning and preprocessing
Phase 3	Exploratory data analysis and hypothesis testing
Phase 4	Model building and evaluation
Phase 5	Threshold optimization and business interpretation