Loan Default Prediction using Machine Learning

Problem Statement / Requirement Specification Document

Prepared by: Habib Shaikh (Agentic AI Expert)

Date: October 2025

Version: 2.0

1. Business Context

Financial institutions face significant challenges due to customer loan defaults, which impact profitability and risk exposure. Predicting whether a borrower is likely to default on a loan is essential for risk management, credit scoring, and decision-making.

This project focuses on building an automated loan default prediction system using machine learning techniques to assess the probability of default for each applicant, enabling proactive measures for risk mitigation.

2. Objective

To design and deploy a predictive analytics system that:

1. Uses historical loan and borrower data to predict loan default risk.
2. Implements a supervised machine learning model (Logistic Regression / Random Forest).
3. Tracks experiments using MLflow for reproducibility.
4. Provides a FastAPI-based REST service for real-time scoring.
5. Supports containerized deployment via Docker.
6. (Optional) CI/CD validation through GitHub Actions.

3. Scope of Work

The solution includes data preprocessing, feature engineering, model training and evaluation, model deployment through FastAPI, Docker containerization, and (optional)a CI/CD pipeline for automated testing.

4. Data Description

Input Dataset: loan_default_sample.csv

Columns:

- loan_id: Unique loan identifier
- age: Age of the applicant
- annual_income: Annual income of the applicant
- employment_length: Employment duration in years
- home_ownership: Type of home ownership (OWN, RENT, MORTGAGE)
- purpose: Purpose of the loan (debt_consolidation, credit_card, etc.)
- loan_amount: Total loan amount

- term_months: Repayment term in months
- interest_rate: Interest rate for the loan
- dti: Debt-to-income ratio
- credit_score: Applicant's credit score
- delinquency_2yrs: Number of delinquencies in past 2 years
- num_open_acc: Number of open accounts
- target_default: Binary flag (1 = Default, 0 = Non-default)

5. Feature Engineering

Derived features include:
- income_to_loan_ratio = annual_income / loan_amount
- employment_risk = 1 if employment_length < 2 years else 0
- credit_score_binned = categorical bands based on credit_score

Features are standardized using StandardScaler and categorical variables encoded using OneHotEncoder.

6. Modeling Approach

Algorithm: Logistic Regression (baseline) or Random Forest (advanced)
- Data Split: 80% training, 20% testing
- Evaluation Metrics: Accuracy, Precision, Recall, F1-Score, ROC-AUC
- MLflow used for experiment tracking and artifact management.

7. Deliverables

Code:
- src/, predict_api/, exported_model/

Artifacts:
- model.pkl, metrics.json

Supporting:
- notebooks/, Dockerfile,
- README.md
- (Optional) github/workflows/docker-ci.yml, tests/

8. API Design

Endpoints:
- GET /health — Returns model load status.
- POST /predict — Accepts applicant details and returns predicted default probability.

Sample request:

```
{
  "age": 32,
  "annual_income": 60000,
  "employment_length": 3,
  "home_ownership": "RENT",
```

```
  "purpose": "credit_card",
  "loan_amount": 15000,
  "term_months": 36,
  "interest_rate": 12.5,
  "dti": 20.3,
  "credit_score": 720,
  "delinquency_2yrs": 0,
  "num_open_acc": 6
}
```


9. Deployment

Docker-based multi-stage build with non-root user and lightweight runtime image. Port 9000 exposed for API service.
(Optional) CI/CD workflow automatically builds the image, runs tests inside the container, and verifies service health.

10. Expected Outcomes

• Accurate binary classification model predicting loan default risk.
• End-to-end deployable FastAPI microservice for real-time inference.
• Automated CI/CD validation pipeline for reliability and scalability.
• MLflow experiment tracking for transparency and versioning.

11. Success Criteria

AUC > 0.85, Precision > 0.8, CI/CD pipeline 100% pass rate, model response latency < 300ms.

12. Future Enhancements

Add SHAP-based explainability to interpret model predictions, integrate additional features (loan-to-value ratio, region, employment type), and deploy in a cloud environment (AWS Sagemaker / Azure ML) with real-time monitoring dashboards.