Dhirubhai Ambani University
Technology

Formerly DA-IICT

DS605 - Fundamentals of Machine Learning

PROJECT REPORT
(Autumn 2025)

Name	Student ID
Dhruv Parmar	202518030
Mahak Khurdia	202518039
Falak Parmar	202518053
Aditya Jana	202518035

Peer-to-Peer Lending Risk Management – Final Project Report

Declaration

We hereby declare that this project report titled "Peer-to-Peer Lending Risk Management" is an original work completed as part of DS601 (Fundamentals of Machine Learning).

Acknowledgements

We thank our instructor and teaching assistants for their guidance throughout the project.

Table of Contents

1. Abstract

This report presents a complete end-to-end machine learning pipeline designed to predict default risk in peer-to-peer lending platforms. Using 2.9 million raw loan records, the system performs large-scale preprocessing, feature engineering, and model training using XGBoost, LightGBM, and CatBoost. A stacked ensemble with XGBoost as the meta-model achieves strong performance with an F1 score of 0.75 at an optimized threshold of 0.5278 and ROC-AUC of 0.9538 on the final test set of 372,153 records.

2. Introduction

Peer-to-peer lending enables direct interaction between borrowers and investors but comes with significant credit risk. Machine learning provides a scalable approach for automated borrower risk assessment. This project develops a fully modular pipeline capable of handling multi-million-row datasets, performing robust data cleaning, generating predictive features, and optimizing advanced gradient-boosting models.

3. Problem Statement

The goal is to classify loan outcomes as Fully Paid or Default using borrower-level, financial, and loan-level variables. Major challenges include extreme class imbalance, high-dimensional data, noisy financial fields, and distributional distortions caused by outliers and missing values.

4. Dataset Description

Raw dataset: 2,925,493 rows × 145 columns.
After dropping 22 irrelevant/redundant columns, the dataset reduced to 85 features.
After filtering invalid targets, the modeling dataset contained 1,860,765 rows.
Final train/test data after preprocessing: 372,153 rows × 100 features.

5. Data Preprocessing

- Loaded dataset with 2.9M rows and 145 columns.
- Removed 22 columns including IDs, URLs, date fields, and redundant payment summaries.
- No duplicates found.
- Standardized data types and parsed categorical features.
- Created the target variable 'is_default'.
- Applied missing-value handling and numeric preprocessing.
- Final modeling shape: (372,153 rows, 100 columns).

6. Exploratory Data Analysis (Summary)

EDA confirmed strong correlations between default likelihood and indicators such as high debt-to-income ratio, high revolving utilization, and elevated interest rates. Loan term, grade, sub-grade, and delinquencies show strong separation between defaulters and fully paid borrowers.

7. Feature Engineering

- Engineered ratios including payment-to-income and credit-utilization features.
- Combined categorical encodings and scaled numeric attributes.
- Retained 100 final engineered and preprocessed features.

8. Model Training

The pipeline tuned XGBoost, LightGBM, and CatBoost using cross-validated grid search. The best parameters obtained:

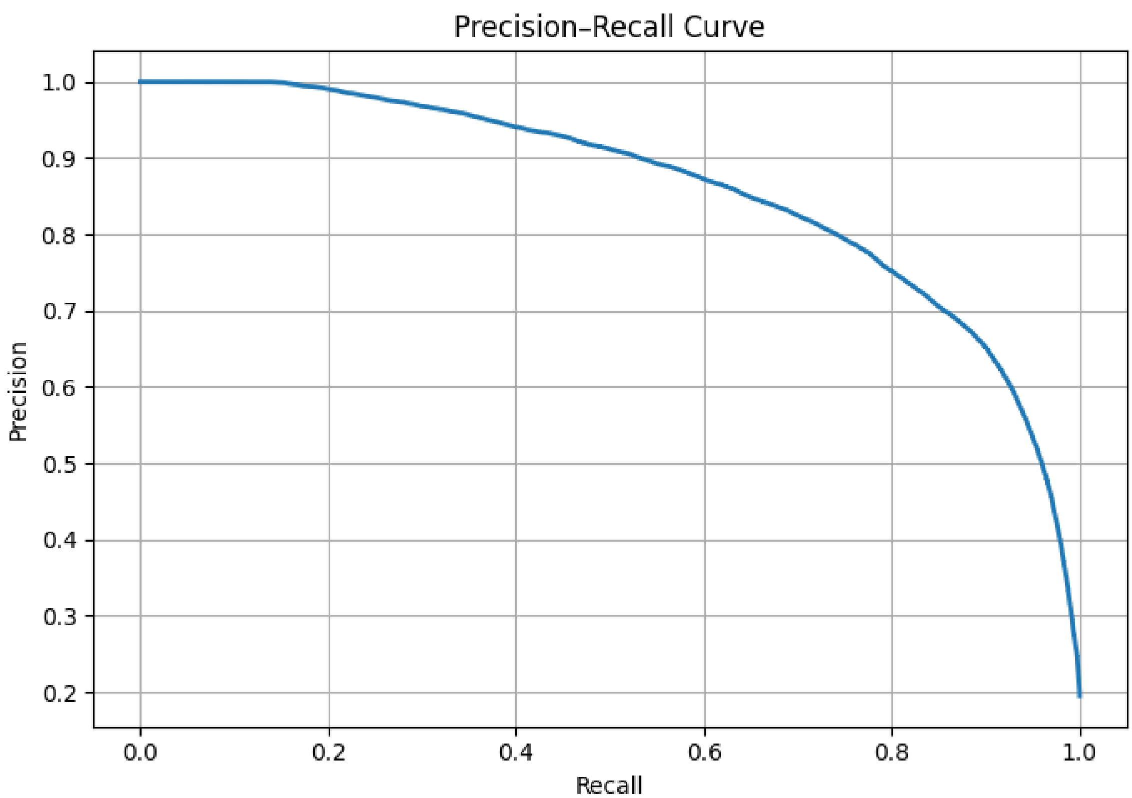
XGBoost: max_depth=6, learning_rate=0.03, n_estimators=500, gamma=0.2, colsample_bytree=0.6
LightGBM: num_leaves=64, learning_rate=0.05, n_estimators=800, subsample=0.6
CatBoost: depth=4, iterations=800, learning_rate=0.03

A stacking ensemble with XGBoost as the meta-model was used for final predictions.
Optimal threshold determined by F1-maximization: 0.3869.

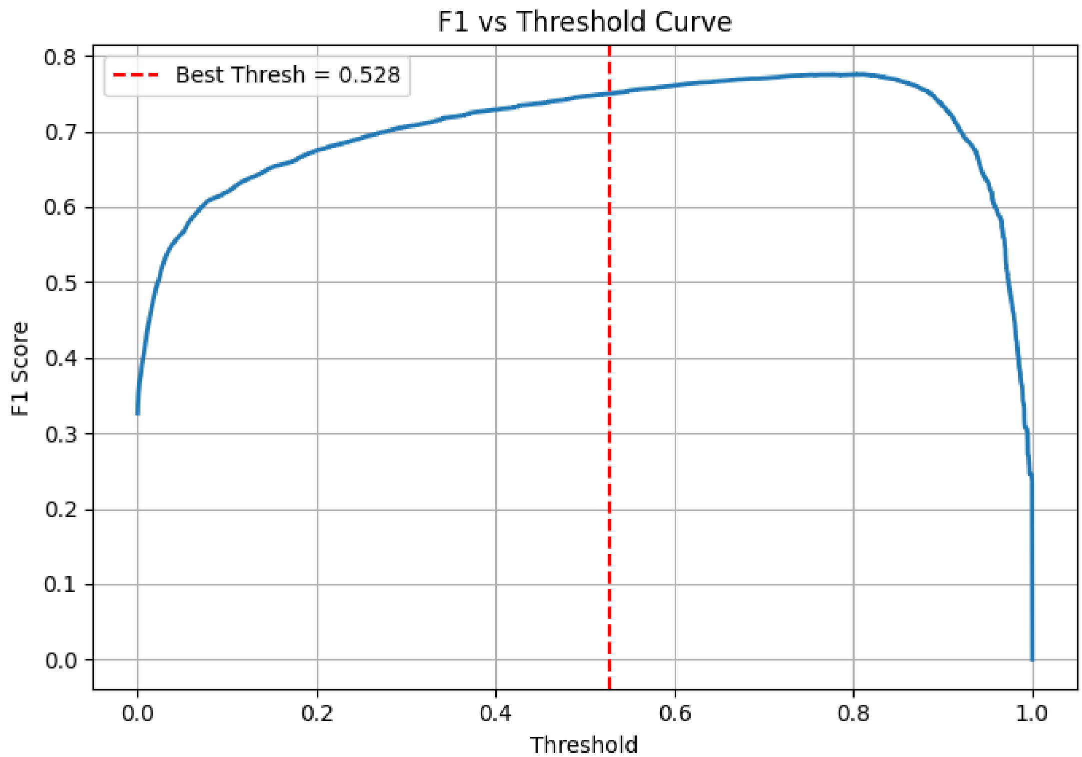
9. Model Evaluation

Final Test Metrics:
Accuracy: 0.8821
Precision: 0.6395
Recall: 0.9067
F1 Score: 0.75
ROC-AUC: 0.9538
Test size: 372,153

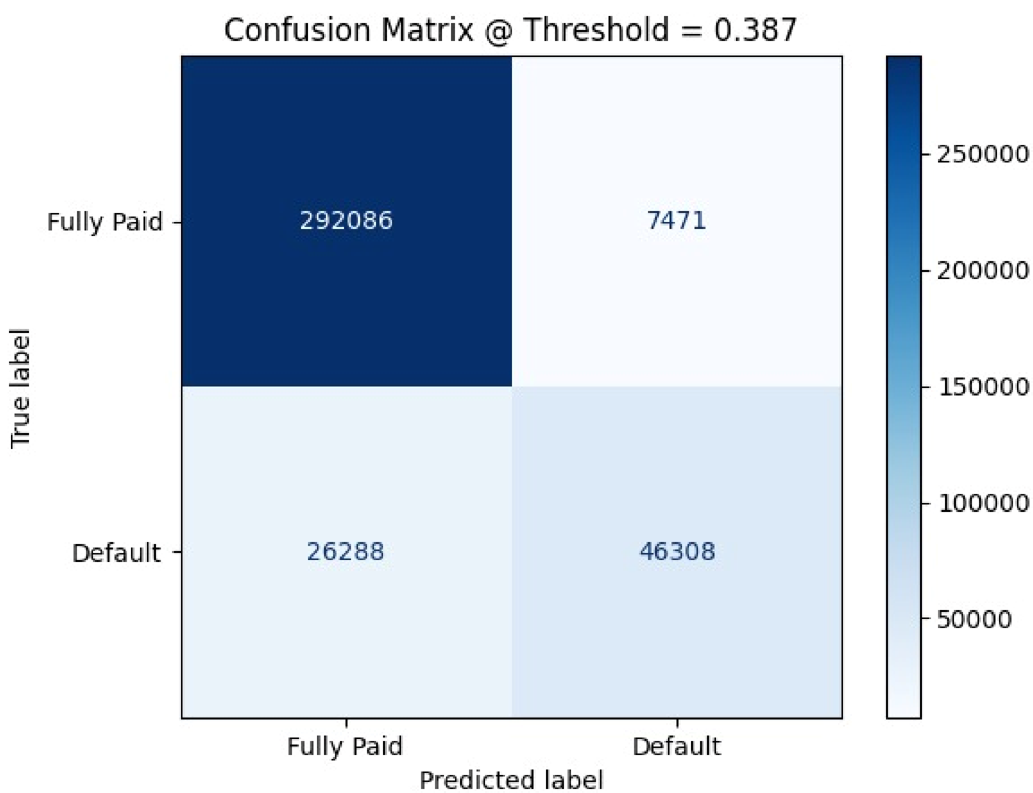
Precision-Recall Curve



F1 vs Threshold Curve



Confusion Matrix at Threshold = 0.3869



10. Unique Contributions

• Complete automated ML pipeline with logging and modular stages.
• Stacking ensemble achieving strong balanced performance.
• Large-scale data preprocessing for multi-million-row datasets.
• Threshold optimization for imbalanced classification.

11. Conclusion

The developed ML pipeline successfully predicts loan default risk in large-scale P2P lending data. The stacking ensemble demonstrates strong performance with a high ROC-AUC and stable F1 score at the optimized threshold. This system can be extended for deployment in real-world P2P lending platforms.

12. Future Work

• Add SHAP-based interpretability.
• Deploy using FastAPI or Streamlit.
• Integrate credit-bureau and macroeconomic variables.
• Add real-time drift monitoring.

13. References

• Scikit-learn Documentation
• XGBoost Documentation
• LightGBM Documentation
• CatBoost Documentation
• Lending Club Dataset Documentation