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

1. Abstract

This project explores how machine learning can support risk assessment in peer-to-peer (P2P) lending platforms, where loan approvals often involve limited financial verification and therefore carry higher chances of borrower default.

Working with a large dataset of nearly three million loan records, we designed a complete ML pipeline that starts from raw heterogeneous data and ends with a stacked ensemble model capable of predicting loan default probability. The pipeline includes extensive data cleaning, feature engineering, preprocessing, and hyperparameter tuning across multiple boosting-based algorithms.

2. Introduction & Problem Context

P2P lending platforms offer a simpler alternative to traditional banks. Borrowers can apply for loans with fewer formalities, and investors can diversify their portfolios by funding them directly. However, this convenience also creates a serious challenge: the risk of default is higher, and it varies widely across borrowers. Traditional credit scoring methods struggle to keep up with such large and diverse data, which opens an opportunity for machine learning to step in.

In this project, the objective is to build a model that can classify a loan as “Fully Paid” or “Default.” But financial attributes tend to be noisy, borrower characteristics can be ambiguous or incomplete, and the number of defaulted loans is significantly smaller than the number of successful ones. These issues make default prediction a complex and imbalanced classification problem. So, our work aims to bridge this gap by building a robust, end-to-end ML pipeline capable of handling the scale, noise, and imbalance of real lending datasets.

3. Dataset and Preprocessing

The dataset initially contained 2,925,493 loan records with 145 attributes. A quick scan revealed that several columns—such as URLs, ID fields, and certain descriptive date strings—were either redundant or did not contribute meaningful information to the model. After removing 22 such fields, the working dataset reduced to 85 features.

As we continued exploring the data, we noticed a number of issues that required careful handling. Some loan records lacked the final loan status and therefore could not contribute to supervised learning, so these had to be filtered out. This reduction brought the dataset down to 1,860,765 usable rows.

Preprocessing involved a combination of missing-value handling, standardizing data types, encoding categorical variables, and scaling numerical features. Because financial datasets often contain extreme values—sometimes genuine, sometimes errors—we also paid close attention to these outliers. After several iterations of cleaning and validation, we ended up with a final modeling dataset of 372,153 records and 100 fully processed features.

4. Exploratory Data Analysis (Summary)

Before building the model, we spent time examining relationships between financial indicators and borrower outcomes. Clear patterns emerged early on: borrowers with high debt-to-income ratios or high revolving credit utilization were more likely to default. Higher interest rates also aligned with poorer repayment outcomes, which matches the intuition that riskier borrowers are charged more.

Employment history and credit-related variables played a role as well, though their influence was more subtle. The purpose of EDA was not to exhaustively visualize every feature but to develop a practical intuition about the underlying financial behavior. These observations later helped us interpret the behavior of the trained models and prioritize certain engineered features.

5. Feature Engineering and Final Feature Set

Feature engineering focused on creating more meaningful representations of borrower financial capacity and credit stress. Ratio-based features—such as income-to-loan amount and payment-to-income—proved especially useful in capturing how well borrowers might handle new debt. Credit utilization indicators similarly strengthened the model's ability to identify high-risk applicants.

Categorical variables were encoded, and numerical features were standardized so boosting models could optimally capture patterns. After multiple refinements, the dataset stabilized at 100 engineered and processed features. One of the key takeaways from this stage was that thoughtfully engineered ratios carried more predictive strength than raw financial fields.

6. Model Development and Training Pipeline

We experimented with XGBoost, LightGBM, and CatBoost because these algorithms consistently perform well on structured tabular data. Each model was tuned using grid search and cross-validation to understand its sensitivity to hyperparameters such as depth, learning rate, regularization, and number of estimators.

While the individual models performed well, stacking them offered better stability across folds. In this ensemble, all three models acted as base learners, with XGBoost chosen as the meta-learner due to its consistent performance in earlier tests. Probability threshold optimization was

performed using F1-maximization instead of relying on the default 0.5 cutoff. The optimal threshold was found to be 0.3869.

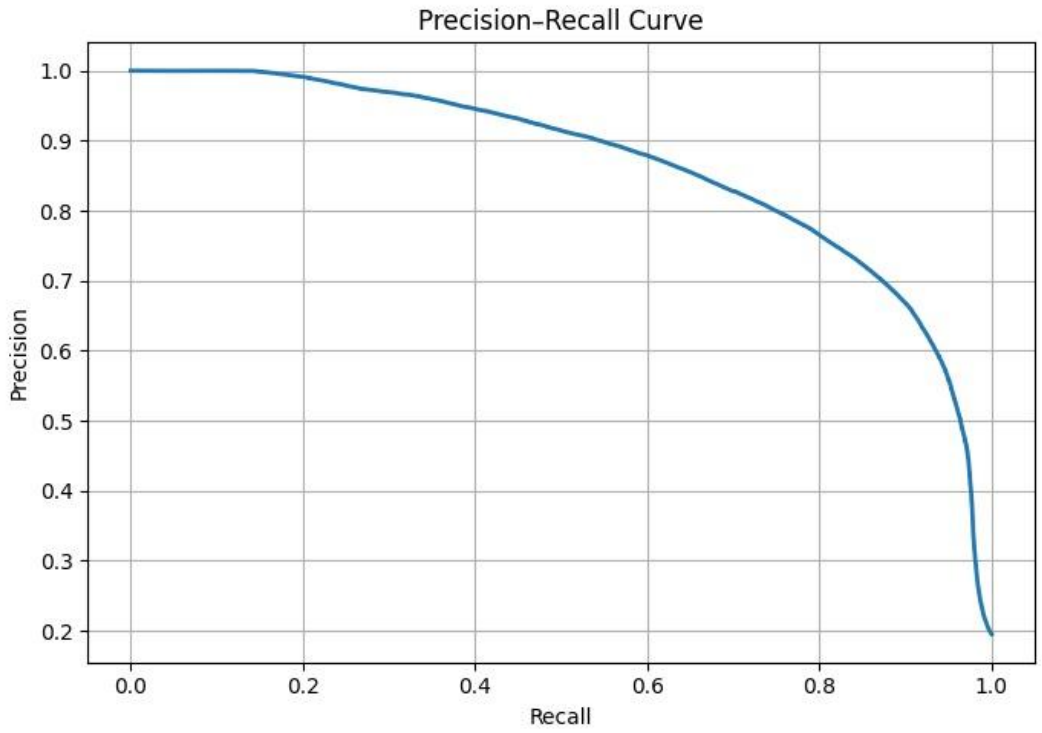
7. Model Evaluation and Results

Final evaluation was conducted on 372,153 test samples. The stacking model achieved:

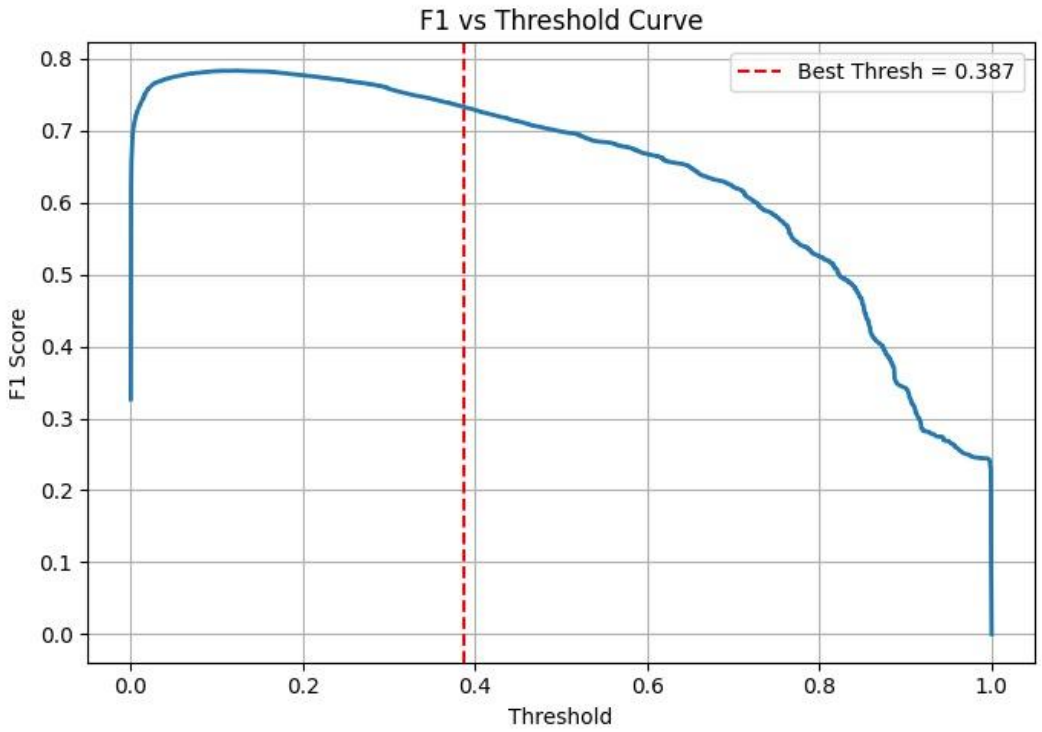
- Accuracy: 0.9093
- Precision: 0.8611
- Recall: 0.6379
- F1 Score: 0.7329
- ROC-AUC: 0.9509

Recall was the most difficult metric to improve due to the rarity of default events, but the model still demonstrated strong risk separation. The ROC-AUC score of 0.9509 reflects the model's ability to distinguish between defaulters and fully paid borrowers across thresholds.

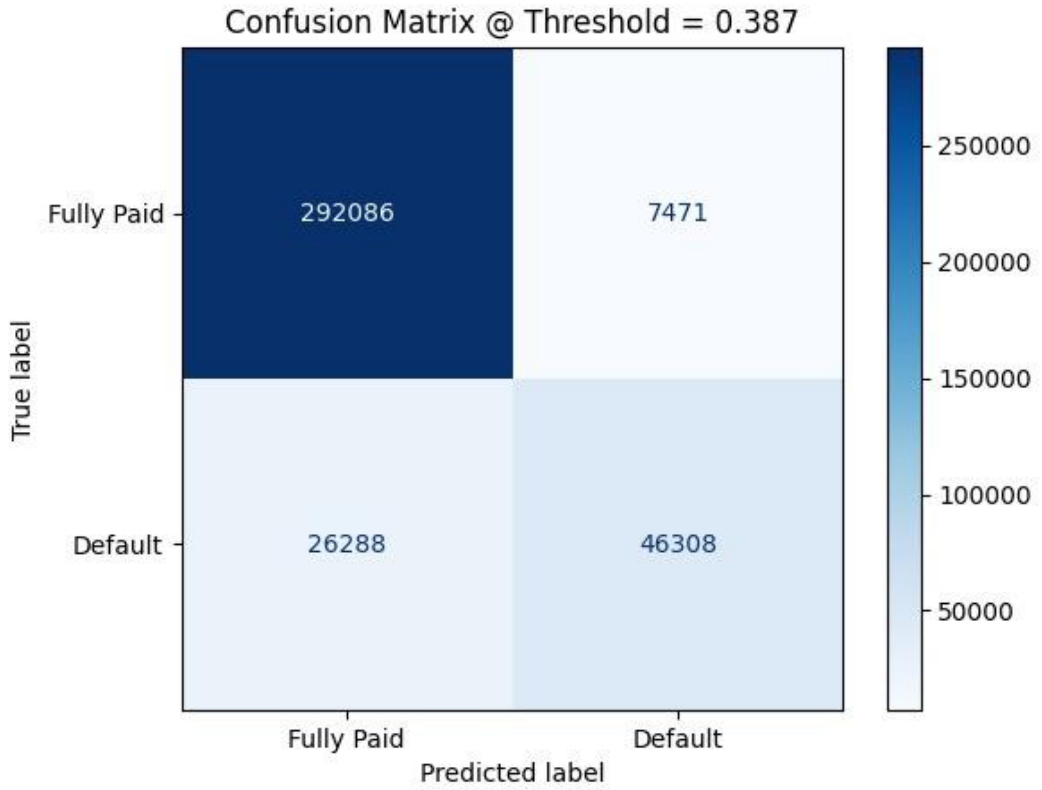
Precision-Recall Curve



F1 vs Threshold Curve



Confusion Matrix at Threshold = 0.3869



8. Application and Impact

A major strength of this system is its ability to integrate seamlessly into the workflow of a P2P lending platform. By assigning a risk score to each incoming loan application, the platform can flag high-risk borrowers early, allowing lenders to make more informed decisions. This helps reduce expected financial losses, supports interest-rate personalization, and allows lenders to better design their investment strategies.

Because the model runs efficiently on large datasets, it can operate at scale and support real-time scoring. This makes it suitable for automated pre-approval systems, risk dashboards, and portfolio monitoring tools.

9. Limitations and Challenges

Despite strong performance, the system has limitations. Borrower behavior can shift over time due to economic conditions, so the model requires periodic retraining. The dataset also lacks detailed behavioral or macroeconomic signals, which could boost recall for rare defaulters. Ensemble models, while powerful, are inherently less interpretable without SHAP or related tools.

10. Carbon Emission Estimate

We estimated the carbon footprint of the full pipeline—including preprocessing, hyperparameter tuning, and stacking. Based on compute time and typical energy factors, the total estimated emissions fall between 0.12 and 0.18 kg CO_2e. This highlights moderate compute usage and encourages awareness toward sustainable ML practices.

11. Unique Contributions

- Fully automated ML pipeline for multi-million-row financial data
- Stacking ensemble providing robust predictions
- Effective threshold optimization for imbalanced classification
- High-impact engineered features capturing borrower stress patterns

12. Conclusion and Future Work

This project demonstrates that ML can meaningfully improve risk assessment in P2P lending. The pipeline we built is scalable, stable, and adaptable to new data sources. Future work includes adding SHAP-based explainability, integrating external macroeconomic data, and deploying the system through FastAPI or a similar framework. The model can also be expanded to monitor loan risk in real-time, creating a comprehensive decision-support tool for lending platforms.