Beyond the Banks: A Machine Learning Odyssey in Loan Default Prediction

Dhanashri Deshpande, Abhigna Balusani, Shuyang Ren
Advisor: Prof. Yifan Hu

Northeastern University

Problem Statement and Motivation

Problem:
Loan defaults pose a significant financial risk to lending institutions, yet accurate prediction remains challenging due to extreme class imbalance. Traditional models often favor the majority class, resulting in undetected high-risk borrowers and increased financial loss.

Goal:
Develop a robust machine learning pipeline that:

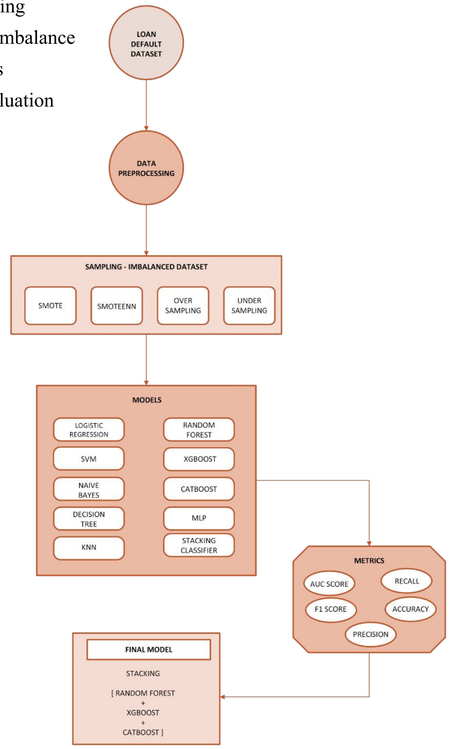
- Effectively addresses class imbalance in the dataset.
- Trains and evaluates different models ranging from simple baselines to advanced ensemble methods.
- Maximizes recall while maintaining strong F1 score and AUC for balanced performance and reliable fraud detection.

Purpose:
To enhance the early identification of high-risk loan applicants and minimize financial losses by deploying a predictive model that is both accurate and sensitive to rare default cases.

Project Overview

Our project pipeline consists of the following:

1. Data Preprocessing
2. Handling Data Imbalance
3. Training Models
4. Testing and Evaluation



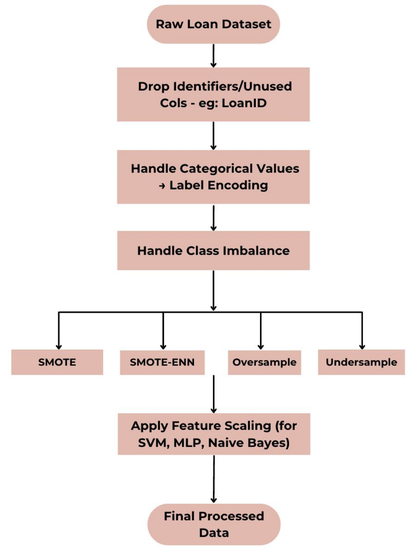
Dataset & Preprocessing

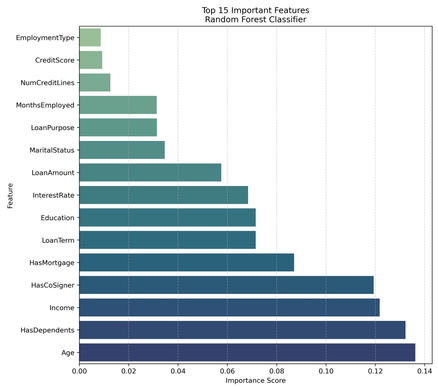
Dataset: from *Coursera's Loan Default Prediction Challenge*. Contains 255,347 rows of data and 18 columns in total.

Insights:

- Many non-binary categorical variables
- Target variable is imbalanced with defaults being rare
- Preliminary analysis yields no strong relations between variables
- No missing values or outliers

Preprocess: We prepared and refined the data, addressing issues to best train models effectively:





Classification and Models

Baseline Models:

- Logistic Regression
- Support Vector Machine (Linear)
- Naive Bayes (with priors & SMOTE)
- Decision Tree (Default & Balanced)
- K-Nearest Neighbors (KNN with SMOTE, Over, UnderSampling)

Advanced Models:

- Random Forest (Tuned with class weights)
- XGBoost: Gradient boosting with high precision
- CatBoost: Optimized for categorical features, best single-model performer
- Multilayer Perceptron (MLP): Neural net capturing non-linear relationships

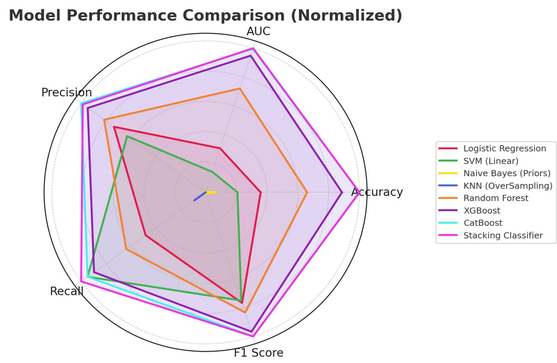
Ensemble Approach:

- Stacking Classifier:
  - Combines Random Forest, XGBoost, CatBoost
  - Achieved highest Accuracy, F1, AUC

Evaluation Setup:

- ROC-AUC, Precision, Recall, Accuracy, F1-score
- Threshold tuning (esp. for Naive Bayes)

Model	Precision	Recall	AUC	Accuracy
Logistic Regression	0.77	0.8	0.82	0.75
SVM(Linear)	0.69	0.89	0.78	0.71
Random Forest	0.83	0.83	0.91	0.83
XGBoost	0.93	0.88	0.96	0.89
CatBoost	**0.97**	0.89	**0.97**	**0.92**
Stacking Classifier	0.96	**0.90**	**0.97**	**0.92**



Recommendation

- Use **CatBoost** or the **Stacking Ensemble** in production for high-stakes loan approval — they achieved the best AUC (0.9689) and F1-Score (0.93).
- Prefer **class priors** over SMOTE when using **Naive Bayes** — it yields better recall and F1.
- Use **XGBoost** when model explainability or speed is critical.

Conclusion and Impact

Conclusions:

- Our ML pipeline significantly improves the identification of potential defaulters, even when default cases are rare (class imbalance).
- It enhances default prediction accuracy by:
  - Using resampling techniques (SMOTE, SMOTE-ENN, priors)
  - Comparing 13 models from simple to ensemble-based
  - Prioritizing recall, to reduce false negatives
- **Best Performing Models**:
  - **CatBoost**: Robust with categorical features; strong performance across all metrics
  - **Stacking Ensemble** (CatBoost + XGBoost + RF): Highest AUC (0.9689) and F1-score (0.93)

Impact:

- Our machine learning pipeline contributes to assessing the **Probability of Default (PD)** — a key pillar of credit risk evaluation used by banks and lenders.
- By accurately identifying borrowers likely to default, the model:
  - Strengthens early risk detection
  - Supports credit approval and underwriting decisions
  - Helps reduce non-performing loans (NPLs)
  - Enables more responsible and data-driven lending
- **Approach is adaptable for use in:**
  - 🏦 Predicting Creditworthiness
  - 🕵️ Fraud detection
  - 📊 Risk analytics
- **Future work:** cost analysis, real-world simulation.

More details

Scan to access:

- 📁 Full GitHub repository with code, notebooks, and documentation
- 📊 Results - metrics, confusion matrices
- 📌 Model training and evaluation logs
- 📎 ReadMe with setup instructions