
INDIAN INSTITUTE OF TECHNOLOGY BOMBAY
भारतीय प्रौद्योगिकी संस्थान मुंबई
ज्ञानम् परमम् ध्येयम्

Project Report

Financial Default Prediction with Ensemble Learning

Abstract

This project focuses on predicting the likelihood of credit card payment defaults using a dataset of customer demographic, account, and payment behavior features. The goal is to develop a robust machine learning model capable of identifying potential defaulters in advance, enabling financial institutions to take proactive measures in risk mitigation and credit policy design. The approach combines exploratory data analysis, feature engineering, model benchmarking, ensemble learning, threshold optimization, and model explainability techniques.

---

1. Introduction

Credit card default risk is a significant concern for banks and lending institutions. An accurate prediction model can help reduce financial losses and improve decision-making for credit approval and collection strategies. This project uses the UCI Credit Card Default dataset to create an end-to-end machine learning pipeline incorporating both traditional and advanced techniques.

---

2. About the Data

- Source: UCI Machine Learning Repository — Default of Credit Card Clients Dataset
- Rows: ~30,000 customers
- Features:
    - Demographics: Age, Sex, Education, Marriage
    - Financial Attributes: Credit limit, bill statements, payment amounts
    - Behavioral History: Payment status over the last six months
- Target Variable: default_payment_next_month (1 = Default, 0 = No Default)
- Class Balance: Dataset is imbalanced, with non-defaulters forming the majority.

---

3. Methodology

3.1 Exploratory Data Analysis (EDA)

- Inspected missing values, data types, and statistical summaries
- Analyzed class imbalance — defaults ~22%, non-defaults ~78%
- Identified correlations between payment delays, bill amounts, and default likelihood

- Visualized distributions and payment history trends using histograms and boxplots

3.2 Data Preprocessing

- Missing Value Handling: Imputed missing values for applicable features
- Encoding: Applied One-Hot Encoding for categorical features, ordinal encoding for ordered variables
- Scaling: RobustScaler used to handle skewness and outliers in numeric data
- Class Imbalance Handling: Used SMOTETomek to balance classes before model training

3.3 Feature Engineering

- Derived features such as total bill amount, average payment ratio, and payment delay count
- Checked multicollinearity using Variance Inflation Factor (VIF)
- Dropped highly correlated redundant features to improve model stability

3.4 Model Development & Benchmarking

- Base Models:
    - Logistic Regression
    - Random Forest
    - HistGradientBoostingClassifier
    - XGBoost, LightGBM, CatBoost (if available)
- Ensemble Approaches:
    - Voting Classifier (hard & soft voting)
    - Stacking Classifier combining multiple algorithms

3.5 Hyperparameter Tuning

- Used RandomizedSearchCV with StratifiedKFold cross-validation to optimize parameters while controlling runtime
- Tuned parameters like tree depth, learning rate, regularization terms

3.6 Evaluation Metrics

- Accuracy, Precision, Recall, F1-score
- ROC AUC Score (primary metric due to imbalance)
- Confusion Matrix to evaluate true/false positive rates
- Precision-Recall curves for assessing performance in imbalanced settings

3.7 Calibration & Threshold Optimization

- Plotted calibration curves to ensure probability predictions are well-calibrated
- Adjusted decision thresholds to balance sensitivity and specificity depending on business need

3.8 Model Explainability

- Used SHAP values to interpret feature importance at both global and local levels
- Identified that payment history and recent bill amounts were the most influential features in predicting default

---

4. Key Findings

- Payment history features (PAY_1, PAY_2, etc.) were the strongest predictors of default
- Credit limit and recent bill amounts had moderate predictive power
- Class imbalance handling improved recall for default cases by ~15% without a significant drop in precision
- Ensemble methods outperformed single models, with Stacking Classifier achieving the best ROC AUC score
- Threshold tuning enabled different operating points depending on risk tolerance

---

5. Business Implications

- The model can help banks preemptively flag high-risk customers
- Allows dynamic credit limit adjustments and targeted payment reminders
- Improves overall portfolio risk management and reduces NPA (Non-Performing Asset) growth

---

6. Conclusion

This project successfully built a complete, industry-grade credit default prediction system. The pipeline included EDA, preprocessing, class imbalance correction, model benchmarking, ensemble learning, and explainability. The final model achieved high discriminatory power and interpretability, making it suitable for real-world deployment in financial institutions.