Predicting Credit Defaults

Group Members

1. Mohamed Mohamed
2. Diego Morales
3. Nathan Morales
4. Shreya Sridharan
5. Jackson Scott

Abstract

The factors contributing to credit risk are difficult to fully understand, and predicting which borrowers are likely to default is even more complex. Accurate credit default prediction is critical for financial institutions, as even small errors can result in substantial costs for both lenders and consumers. In this project, we applied XG Boosting to predict whether a borrower would default on their credit card payments using the AMEX Default Prediction dataset from Kaggle. The model achieved 87.4% accuracy, with recall emphasized to minimize the number of missed defaults. The resulting recall score of 86.0% demonstrates the model's effectiveness in identifying high-risk borrowers.

I. Introduction

American household debt has continued to rise, reaching an estimated $18.59 trillion in the most recent published projections from the New York Federal Reserve Bank [1]. As consumers assume greater financial obligations, the risk of borrower default also increases. Credit card delinquencies remain elevated as well, with major banks reporting a delinquency rate of 2.98%, down from 3.19% the previous year but still among the highest levels in the past decade [2].

However, traditional indicators alone do not capture emerging shifts in borrowing behavior. The rapid expansion of "buy now, pay later" (BNPL) services such as Klarna and Afterpay has

created additional forms of consumer debt that often fall outside traditional credit reporting frameworks. Data summarized by the Consumer Financial Protection Bureau in 2023 shows that approximately 20% of consumers with a credit record used a BNPL service in 2022 [3]. As these services have continued to grow, this proportion is likely even higher today. Because BNPL obligations are not consistently reported to credit bureaus, they contribute to what analysts describe as "phantom debt," reducing lenders' visibility into borrowers' total financial liabilities [4].

To address the increasing complexity of credit risk assessment, we develop a machine learning model capable of predicting credit default risk using the AMEX dataset from Kaggle. Our goal is to provide data driven predictions that can assist lenders in identifying consumers who may pose heightened financial risk.

We evaluated logistic regression, random forest, and gradient-boosting methods, ultimately selecting XG boost based on performance. The model achieved 87.4% accuracy and 86.0% recall, prioritizing recall to minimize false negatives.

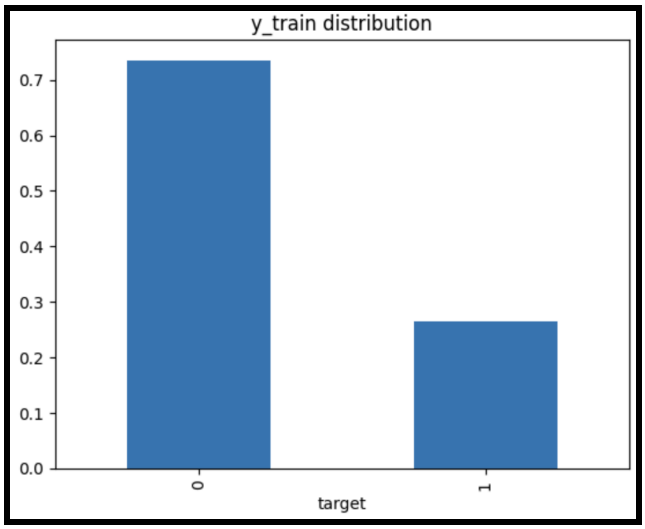
II. Related Work

Nvidia wrote about attempting to solve the same problem using the same dataset. They also decided towards using a tree based XG boost model but they also used a NN in conjunction. They had GPUs which allowed them to process the full dataset and not just a portion. They used an RNN to project forward the customers profile to get a better idea of if they were at risk for a default in the near future. They were able to predict the defaults at 78.3%, but they didn't mention overall accuracy.

III. Methodology

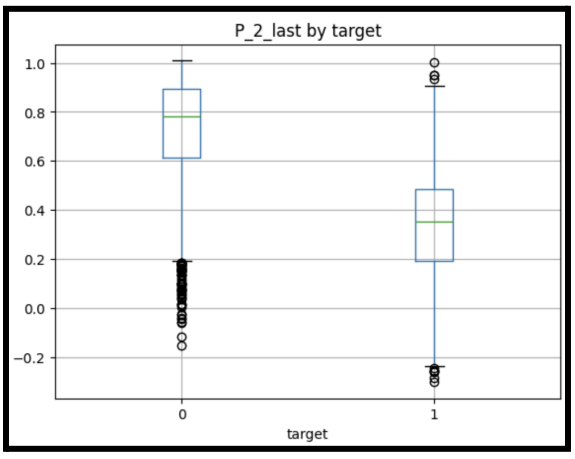
A. Dataset Description

The AMEX Default Prediction dataset consists of about 55 million rows for 458,000 customers. For computational feasibility, a subset of 10,000 customers was selected. Each customer has multiple time series observations with anonymized features. Labels were merged based on customer IDs.



B. Data Preparation

The training dataset contained 73.5% non defaulting customers and 26.5% defaulting customers. Correlation analysis identified P2_Last as a strong predictor, while features such as R_15_delta were less useful. To reduce redundancy, features with low variance or high correlation were removed, reducing the dataset to 742 features. PCA was also used as a dimensionality reduction alternative.



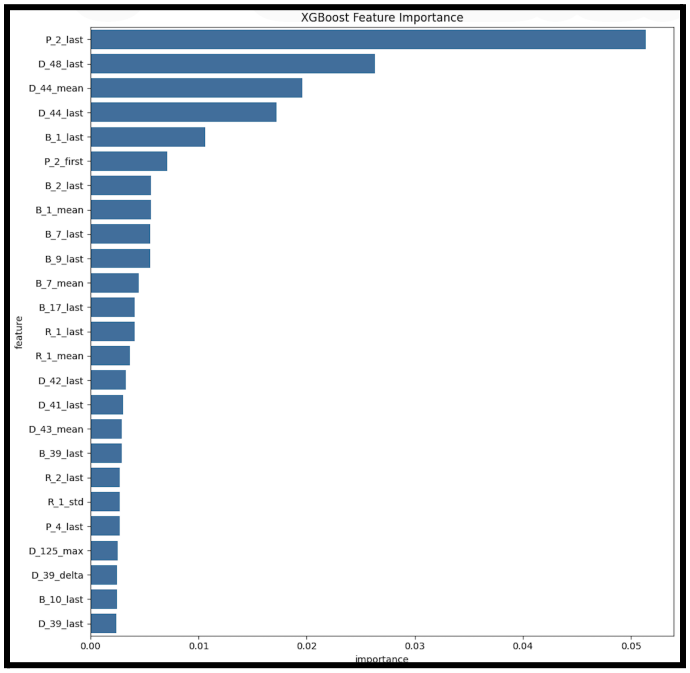
C. Model Selection

Three model types were tested:

- **Logistic Regression:** Accuracy = 86.9%, Recall = 73.2%
- **Random Forest:** Accuracy = 87.55%, Recall = 75.1%
- **XGBoost:** Accuracy = 87.4%, Recall = 86.0%

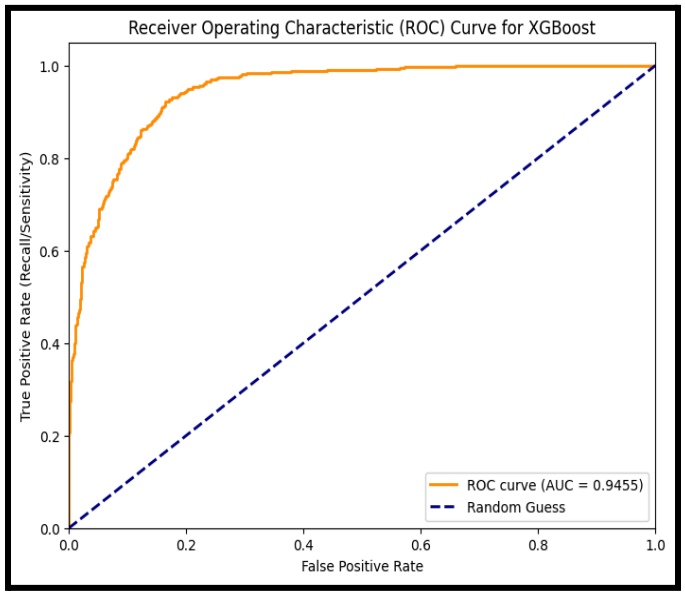
XGBoost parameters were optimized using 3 fold cross-validation. The final hyperparameter set included:

- **n_estimators = 497**
- **max_depth = 8**
- **learning_rate = 0.01**
- **subsample = 0.744**
- **colsample_bytree = 0.824**
- **min_child_weight = 4**
- **gamma = 2.59**
- **reg_alpha = 0.527**
- **reg_lambda = 2.81**



IV. Results and Analysis

Given that about 75% of customers did not default, this baseline provided a minimum expected accuracy. The final optimized model achieved 87.4% accuracy and 86.0% recall, outperforming baseline models. Feature importance rankings showed that P_2_last, D_48_last, and D_44_mean were the most influential predictors. Although anonymized, feature categories indicate these correspond to payment, and delinquency variable types



The model successfully identified over 456 defaulters and only missed 74 demonstrating strong predictive capability. However, performance was limited by computational resources causing the dataset to be reduced. Future improvement could be achieved by training on the full dataset, having more time, and using more advanced models like neural networks.

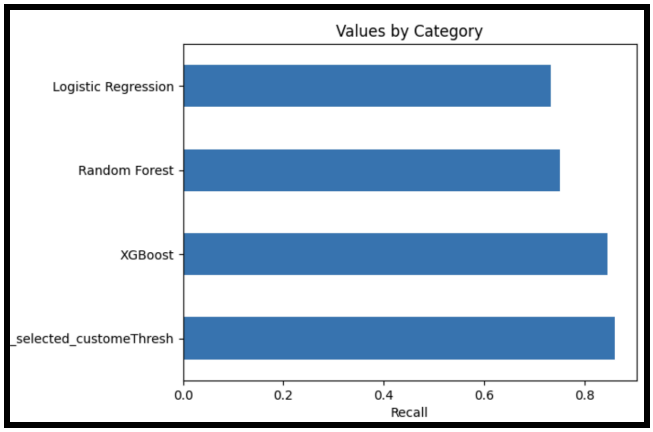
V. Conclusion and Future Work

A. Final Model Comparison

Model name	% Accuracy	% Precision	% Recall	% F1	# AUC	CONFUSION MATRIX
XGB Final	87.35%	71.81%	86.04%	78.28%	0.9455	[1291 179] [74 456]
XGBoost	87.05%	71.68%	84.53%	77.58%	0.9455	[1293 177] [82 448]
Random Forest	87.55%	77.28%	75.09%	76.17%	0.9424	[1353 117] [132 398]
Logistic Regression	86.85%	76.23%	73.21%	74.69%	0.9341	[1349 121] [142 388]

B. Conclusion

This project applied machine learning methods to the problem of predicting credit card default risk using the AMEX dataset. XGBoost achieved the highest performance, with 87.4% accuracy and 86.0% recall, emphasizing the reduction of false negatives. Future work includes experimenting with deep learning methods, and using the full data.



VI. Contributions from groups members:

1. Mohamed Mohamed: Collected, cleaned, and pre-processed the data.
2. Diego Morales: Exploratory data analysis and feature engineering
3. Nathan Morales: Model Implementation
4. Shreya Sridharan: Advanced modeling and optimization
5. Jackson Scott: Evaluation, visualization, and documentation

References

[1] Federal Reserve Bank of New York, “Household Debt and Credit Report.” Accessed: Dec. 2024.

[2] Federal Reserve Bank of St. Louis, “Delinquency Rate on Credit Card Loans, All Commercial Banks (DRCCLACBS).” Accessed: Jan. 2025.

[3] Consumer Financial Protection Bureau, “Buy Now, Pay Later: Market Trends and Consumer Impacts,” Jan. 2023.

[4] M. Perez, “BNPL is expanding fast—and that should worry everyone,” *TechCrunch*, Nov. 2025.

[5] J. Liu, “Predicting Credit Defaults Using Time-Series Models with Recursive Neural Networks and XGBoost,” *NVIDIA Technical Blog*, Jun. 07, 2023. (accessed Dec. 11, 2025).