Credit Risk Prediction: Unlocking Customer Behavior Patterns

Jeena Mole
M.S Computer Science
G01460309
jmole@gmu.edu

Parshwa Gandhi
M.S Computer Science
G01511122
pgandhi6@gmu.edu

Krish Sanghvi
M.S Computer Science
G01521041
ksanghvi@gmu.edu

***Abstract*—This project aims to predict customer credit defaults for American Express using advanced machine learning models, including XGBoost, CatBoost, and LightGBM. The data preprocessing involves extracting the latest customer records, handling missing values, and encoding categorical features. Key features were grouped into domains such as Delinquency, Spend, Payment, Balance, and Risk, with exploratory data analysis (EDA) revealing their distributions and correlations. Extensive data preprocessing techniques were used to address the challenges faced such as class imbalance. Model performance evaluation metrics such as F1-score, precision, recall, accuracy, AUC-ROC, log-loss and confusion matrix were used to evaluate the models. Visualization methods such as correlation heatmaps, KDE plots complemented the analysis. The implementation demonstrates a robust workflow for solving imbalanced prediction problems while providing actionable insights into feature importance and selection using customer behavior.**

***Index Terms*—Credit Risk Prediction, Machine Learning, CatBoost, Feature Engineering, Financial Systems.**

1. Introduction

Credit card default prediction is a crucial aspect of modern financial systems, enabling institutions to mitigate risks, optimize credit offerings, and maintain economic stability. This project focuses on predicting the probability of default using customer behavior data provided by American Express. The anonymized dataset includes aggregated features categorized into delinquency, spending, payment, balance, and risk variables. The study leverages the CatBoost algorithm, known for its efficiency with categorical data and robustness against overfitting, to develop an accurate predictive model. To ensure a comprehensive analysis, we employed advanced preprocessing techniques, including grouping data by customers, handling missing values, and encoding categorical features. Extensive exploratory data analysis (EDA) was conducted to uncover patterns, using visualizations such as kernel density plots and correlation heatmaps. Model performance was rigorously evaluated using multiple metrics, including F1 score, precision, recall, accuracy, AUC-ROC, and log loss. Furthermore, the Receiver Operating Characteristic (ROC) curve was plotted to visualize the model's discrimination power. Using, different machine learning algorithms helped us compared their performances against each other and identify the best suited algorithm for tackling such problems.

2. Dataset Overview

The dataset originates from an anonymized credit risk prediction competition hosted by American Express, with the aim of predicting the likelihood of credit card customers defaulting on their balances. It contains aggregated customer behavioral data, covering variables related to delinquency, spending, payments, balances, and risk metrics.

The original training dataset consists of 458,913 rows and 190 columns, representing one record per customer after grouping by `customer_ID`. The train data covers the period from March 1, 2018, to March 31, 2018, and includes a binary target variable, `target`, where 1 indicates a customer defaulted and 0 indicates no default. The original test dataset includes 924,621 rows and 189 columns, excluding the target variable, spanning the period from April 1, 2019, to October 31, 2019. But, due to the large dataset size, we used the compressed version of the train and test sets AMEX-Feather-Dataset and they took the last statement for each customer. So, training and testing data are loaded from feather files.

The features in the dataset are categorized into five groups. Delinquency features (`D_`$\star$) indicate late payments or overdue balances and are strong predictors of future defaults. Spending features (`S_`$\star$) describe monthly expenditures and spending patterns, reflecting potential financial strain. Payment features (`P_`$\star$) describe payment behaviors such as frequency and payment-to-balance ratios, which correlate with default risks. Balance features (`B_`$\star$) represent credit utilization trends and available balances, while risk features (`R_`$\star$) quantify default likelihood using inferred metrics.

The dataset is anonymized to protect customer privacy, with features normalized to ensure consistent scales. However, it contains missing values, which were visualized using heatmaps and addressed through imputation strategies during preprocessing. Another challenge is the class imbalance in the target variable, with significantly fewer cases of defaults compared to non-defaults, necessitating tailored evaluation metrics to ensure balanced predictions.

3. Methodology

3.1. Data Preprocessing

Effective preprocessing is crucial for building an accurate and reliable prediction model. The dataset used in this project, which includes anonymized and normalized features, underwent several preprocessing steps to ensure consistency and prepare it for machine learning algorithms.

Initially, the data was grouped by the customer_ID column to retain only the most recent statement for each customer. This step ensured the relevance of the data by focusing on the latest available customer behavior. Temporal features, such as the S_2 column representing dates, were converted to a datetime format to facilitate temporal analysis and indexing.

The dataset contained missing values across several features. These missing values were visualized using heatmaps to understand their patterns and distributions. Imputation strategies were then applied to fill in the gaps, ensuring that the data remained usable without introducing bias.

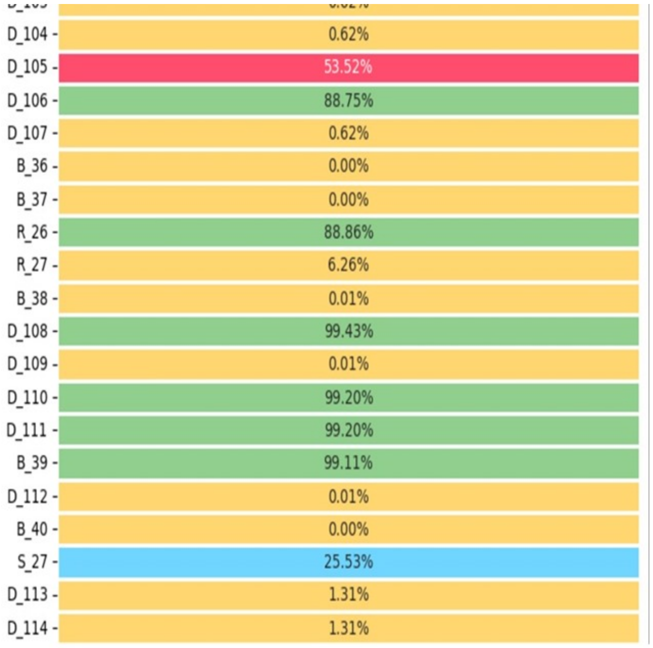

Figure 1. Heatmap Highlighting Dataset Incompleteness.

To prepare the data for training, the dataset was split into training and test subsets. An 80-20 split was used, with 80% of the data allocated for model training and 20% reserved for evaluation.

Finally, the data was standardized to ensure consistent scaling across all features, enhancing model performance and stability. These preprocessing steps established a solid foundation for feature exploration and model training, enabling the development of a robust prediction framework.

3.2. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a critical step in the data analysis pipeline, aimed at understanding the dataset's structure, summarizing its main characteristics, and identifying patterns and relationships between features. In this study, EDA was employed to gain insights into customer behavior and its impact on credit default risk. The outputs from EDA played a pivotal role in guiding preprocessing, feature selection, and model development.

The dataset's features were categorized into five groups: delinquency (D_⋆), spending (S_⋆), payment (P_⋆), balance (B_⋆), and risk (R_⋆). For each feature group, correlation heatmaps were generated to visualize relationships within the group and their correlation with the target variable. These visualizations identified features with high predictive potential, which were subsequently refined and used in the machine learning model.

In addition to heatmaps, Kernel Density Estimation (KDE) plots were utilized to analyze the distribution of features across the target classes (default vs. non-default). This allowed for the identification of key variables with distinct distributions, enhancing the model's ability to differentiate between the classes. For instance, delinquency features showed significant divergence in KDE plots between defaulting and non-defaulting customers, highlighting their importance in predicting default risk.

A comprehensive correlation analysis was performed to understand the relationships between individual variables and the target. Features with higher correlations to the target variable were prioritized for feature selection and engineering. These correlations informed preprocessing decisions, such as encoding and scaling, and helped streamline the modeling process by focusing on the most relevant predictors.

Overall, EDA outputs guided the selection of critical features and transformations that improved model performance, ensuring a robust and data-driven approach to credit default prediction.

3.2.1. Delinquency Features (D_⋆).

3.2.1.1. . Correlation heatmap of Delinquency Features:

The correlation heatmap for delinquency features (D_⋆) visualizes the relationships among variables related to late payments and overdue balances, providing insights into their role in predicting credit default risk. The heatmap reveals clusters of highly correlated features, indicating that these variables measure related aspects of customer payment behavior. Strong positive correlations among features like D_39, D_44, and D_48 highlight their importance, while weaker or negative correlations suggest complementary perspectives. These observations guided the selection of key delinquency features for the predictive model, prioritizing variables with strong correlations to the target and handling redundancies to reduce multicollinearity. This analysis underscores the importance of delinquency features in distinguishing defaulting customers from non-defaulting ones and informed feature selection and transformation steps in the modeling process.

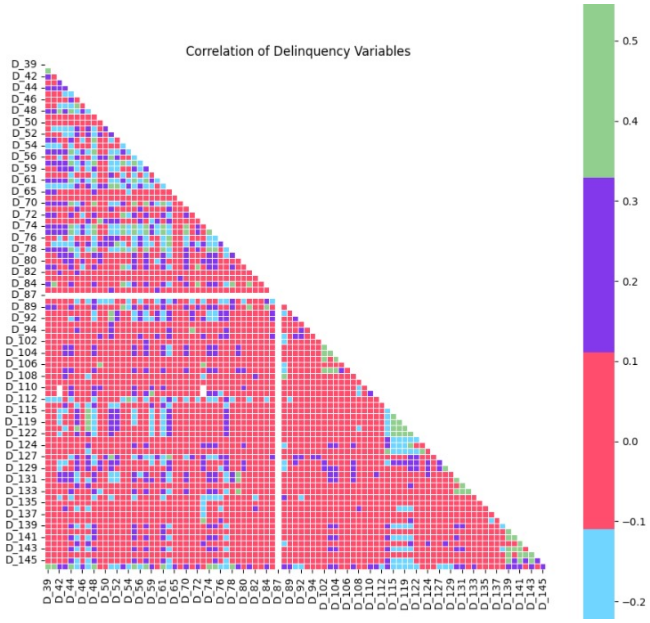

Figure 2. Correlation Heatmap of Delinquency Features.

3.2.1.2. . Kernel Density Estimation (KDE) of Delinquency Features :

The Kernel Density Estimation (KDE) plots for delinquency features (D_⋆) provide insights into the distribution of variables related to late payments and overdue balances across defaulting (target = 1) and non-defaulting (target = 0) customers. These plots reveal key patterns in customer behavior, with several features showing significant differences in density between the two target classes. For example, certain delinquency variables exhibit higher densities for defaulting customers, indicating that overdue balances or frequent late payments are strong indicators of credit default risk. Conversely, some features demonstrate overlapping distributions, which provide limited discriminatory power but may offer complementary insights when combined with other variables. While the displayed KDE plots are examples for a subset of delinquency features, similar analysis was performed across all features in this category, ensuring a thorough understanding of their impact on credit risk prediction. The insights from these plots were critical for selecting the most relevant delinquency features, refining the model's ability to distinguish high-risk customers.

3.2.2. Spending Features (S_⋆).

3.2.2.1. . correlation heatmap of Spending Features :

The correlation heatmap for spending features (S_⋆) illustrates the relationships among variables representing customer expenditure patterns and their relevance to credit default prediction. Several features, such as S_2, S_5, and S_12, exhibit moderate positive correlations, indicating shared spending behaviors that may predict financial reliability. In contrast, features like S_18 and S_26 show weak or negative correlations with other variables, highlighting diverse financial behaviors that offer complementary insights. The heatmap also reveals clusters of strongly correlated features, suggesting patterns of related spending trends, which informed feature grouping and dimensionality reduction during preprocessing. Features with strong correlations to the target variable were prioritized for the predictive model, while redundant features were handled to reduce multicollinearity. This analysis underscores the significance of spending behaviors in distinguishing defaulting customers from non-defaulting ones and refining the model's accuracy.

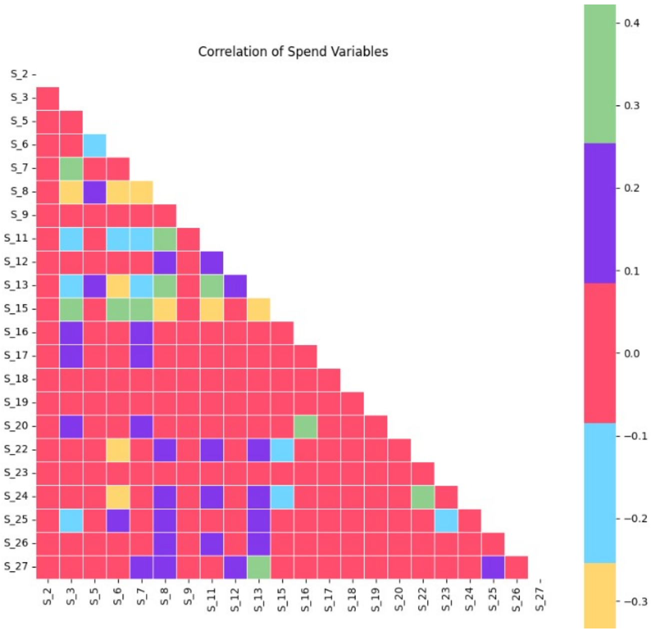

Figure 3. Correlation Heatmap of Spending Features.

3.2.2.2. . Kernel Density Estimation (KDE) of Spending Features :

The Kernel Density Estimation (KDE) plots for spending features (S_⋆) illustrate the distribution of these variables for both defaulting (target = 1) and non-defaulting (target = 0) customers. These plots provide critical insights into how spending behaviors vary across the two target classes. Many features, such as S_3, S_11, and S_17, exhibit distinct distribution patterns between the classes, suggesting their strong predictive potential in distinguishing defaulting customers from non-defaulting ones. For instance, some spending features show higher densities for defaulting customers in specific value ranges, indicating erratic or excessive spending behavior as a potential indicator of default risk. Conversely, features with overlapping distributions, such as S_18, provide less discriminatory power but may still offer complementary insights when combined with other variables. The KDE plots guided feature selection by identifying variables with the greatest class separation, ensuring the model captures spending behaviors relevant to credit risk prediction.

3.2.3. Payment Features (P_*).

3.2.3.1. . correlation heatmap of Payment Features :

The correlation heatmap for payment features (P_*) offers insights into customer payment behavior and its relationship to credit default risk. Moderate positive correlations are observed between certain pairs, suggesting interdependence among payment amounts and frequencies. The presence of weaker correlations indicates diverse payment behaviors providing unique insights into financial reliability. Critical features were identified for predicting default risks, ensuring the model effectively captures customer payment trends. These findings informed feature selection and handling of potential redundancies to minimize multicollinearity in the predictive model.

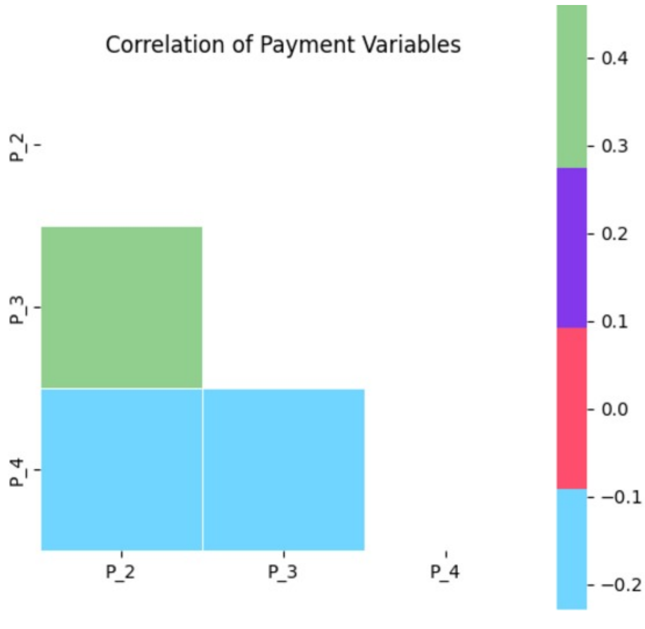

Figure 4. Correlation Heatmap of Payment Features.

3.2.3.2. . Kernel Density Estimation (KDE) of Payment Features :

The Kernel Density Estimation (KDE) plots for payment features (P_*) reveal distinct distribution patterns between defaulting and non-defaulting customers. Features like P_2, P_3, and P_4 show distinct patterns, indicating their significance in predicting default risk. For instance, a higher density for defaulting customers in lower value ranges indicates insufficient payment amounts. Conversely, distinct peaks for non-defaulting customers indicate consistent or higher payment behaviors. However, overlapping distributions suggest limited discriminatory power for P_4. These insights are crucial for accurately capturing payment-related patterns that distinguish high-risk customers.

3.2.4. Balance Features (B_*).

3.2.4.1. . correlation heatmap of Balance Features :

The correlation heatmap for balance features (B_*) reveals relationships among variables affecting credit utilization and account balances, providing insights into customer financial behavior and its impact on credit default risk. Several features, such as B_1, B_9, and B_14, exhibit strong positive correlations, forming distinct clusters that represent related aspects of balance management. Strong positive correlations form clusters representing related aspects of balance management, suggesting consistent financial behavior, such as high credit utilization, which can signal higher default risks. Conversely, weak or negative correlations highlight complementary perspectives, providing additional nuances to customer profiling. The heatmap informs feature selection and reduces redundancy to minimize multicollinearity. Prioritizing features with stronger correlations enhances the model's predictive accuracy by capturing default risk trends related to balance utilization and account management.

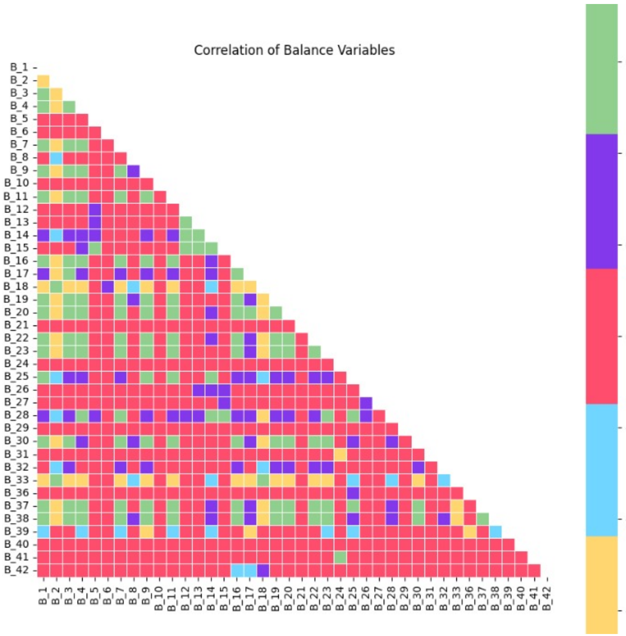

Figure 5. Correlation Heatmap of Balance Features.

3.2.4.2. . Kernel Density Estimation (KDE) of Balance Features :

The Kernel Density Estimation (KDE) plots for balance features (B_⋆) visualize the distribution of variables related to credit utilization and account balances for both defaulting (target = 1) and non-defaulting (target = 0) customers. These plots reveal distinct trends in balance-related behaviors between the two classes, with several features, such as B_3, B_7, and B_11, showing significant density peaks for specific value ranges that distinguish defaulting customers from non-defaulting ones. Features with overlapping distributions, such as B_18 and B_25, provide less direct separation but may still contribute to the overall predictive power when combined with other features. The KDE analysis highlights the importance of balance features in capturing patterns of credit utilization, which are critical indicators of financial risk. These insights were instrumental in selecting key features for model training, ensuring that the predictive model effectively identifies customers with a higher likelihood of defaulting.

3.2.5. Risk Features (R_⋆).

3.2.5.1. . correlation heatmap of Risk Features :

The correlation heatmap for risk features (R_⋆) visualizes the relationships among variables that quantify the likelihood of default, providing critical insights into customer risk assessment. Several features, such as R_2, R_5, and R_13, exhibit moderate to strong positive correlations, forming clusters that highlight shared attributes related to risk evaluation. These clusters indicate patterns of interrelated risk indicators, which are crucial for understanding customer default behavior. Conversely, weaker or negative correlations observed between some features suggest distinct risk dimensions, offering complementary perspectives that enrich the model's understanding of customer profiles. The heatmap also emphasizes the importance of managing multicollinearity by identifying redundant features. Features with higher correlations to the target variable were prioritized during feature selection to improve model accuracy, ensuring that the predictive model effectively captures variations in default risk. This analysis reinforces the significance of risk features in modeling credit default prediction.

3.2.5.2. . Kernel Density Estimation (KDE) of Risk Features :

The Kernel Density Estimation (KDE) plots for risk features (R_⋆) visualize the distribution of these variables across defaulting (target = 1) and non-defaulting (target = 0) customers. These plots highlight the variation in risk-related metrics, with some features demonstrating distinct distributions that aid in distinguishing between the two target classes. For instance, variables such as R_2, R_5, and R_10 exhibit sharper densities for specific ranges within the defaulting group, indicating their strong predictive significance in identifying high-risk customers. Conversely, features with overlapping distributions, such as R_20 and R_25, provide less direct separation but may still contribute value when combined with other variables. These insights guided the selection of the most discriminative risk features, ensuring the model captures critical patterns associated with credit default risk while managing redundancy and noise. The KDE analysis underscores the role of risk features in improving the model's ability to identify customers with heightened default probabilities.

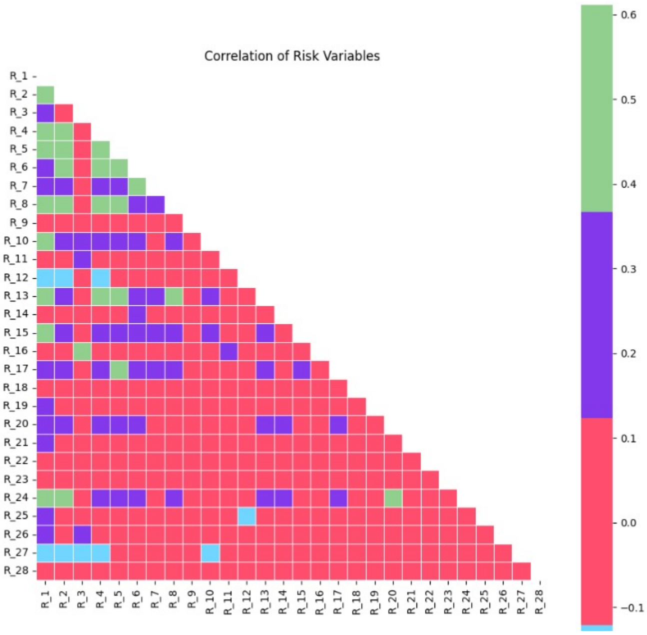

Figure 6. Correlation Heatmap of Risk Features.

3.2.6. Correlation of Each Feature with Target Variable : . The bar chart visualizing the correlation of each variable with the target provides insights into the strength and direction of relationships between individual features and the likelihood of credit default. Positive correlations indicate features that increase in value as the probability of default rises, while negative correlations highlight features that decrease as default risk grows. Notably, variables such as D_145, B_42, and D_132 exhibit the strongest positive correlations, suggesting that delinquency and balance-related features are critical predictors of default. Conversely, variables such as D_39 and D_119 show notable negative correlations, indicating their relevance in identifying non-defaulting customers. This correlation analysis informed the selection of the most predictive features, guiding the preprocessing steps and ensuring the model focused on variables with the highest impact on the target. By quantifying these relationships, the chart highlights the importance of balancing positively and negatively correlated features to enhance the model's discriminatory power.

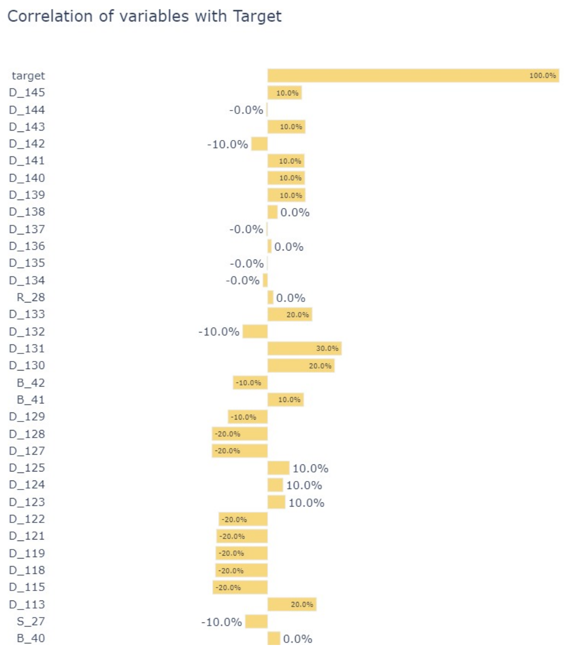

Figure 7. Correlation of Variables with Target.

3.2.7. Encoding and Feature Categorization : . Categorical features, such as B_30, B_38, and others, were encoded using LabelEncoder to transform them into numerical representations suitable for machine learning models. This encoding method preserved the ordinal nature of the categorical data without introducing unnecessary complexity.

Feature engineering was performed to categorize and group variables into meaningful categories, such as delinquency (D_⋆), spending (S_⋆), payment (P_⋆), balance (B_⋆), and risk (R_⋆). Each category was analyzed for its relevance to the target variable.

3.3. Role of EDA Outputs in Further Processes

The findings from EDA informed several downstream processes:

- **Feature Selection:** Identified key features with strong predictive potential for model training.
- **Feature Engineering:** Guided the transformation and normalization of variables to improve model interpretability and performance.
- **Model Development:** Provided insights into customer behavior, enabling the development of a model that effectively captures default risk patterns.

By leveraging EDA outputs, the study ensured a data-driven approach to feature preparation and model building.

4. Model

4.1. XGBoost

XGBoost (Extreme Gradient Boosting) is a highly optimized implementation of the gradient boosting algorithm, a technique that builds an ensemble of weak learners (decision trees) to form a strong predictive model. It is widely recognized for its speed, efficiency, and scalability, making it one of the go-to algorithms for large datasets in both classification and regression tasks. One of the reasons XGBoost stands out is its ability to handle imbalanced datasets effectively by employing techniques like weighted loss functions, which adjust the model's focus on minority classes during training. Additionally, XGBoost can handle missing values natively by treating them as a separate value during decision tree construction, eliminating the need for manual imputation. Its built-in regularization techniques (L1 and L2) further help prevent overfitting, and its parallelization capabilities allow for faster computation, especially when working with large-scale datasets. This makes XGBoost a top choice when dealing with complex, large, and imbalanced financial datasets.

4.2. CatBoost

CatBoost (Categorical Boosting) is another gradient boosting algorithm specifically designed to handle categorical features without requiring heavy preprocessing, such as one-hot encoding or label encoding. This ability to work directly with categorical data is especially valuable in financial applications where categorical variables (e.g., transaction types, customer demographics) are prevalent. CatBoost is known for its robustness in handling missing values, as it automatically deals with them during the training process, which simplifies data preprocessing significantly. It is also particularly adept at dealing with imbalanced datasets through its combination of gradient boosting and optimized handling of categorical data, which helps it maintain high accuracy without overfitting the minority class. Furthermore, CatBoost is designed to be fast, and its implementation includes techniques to avoid overfitting while ensuring that the model generalizes well on unseen data, making it a powerful tool for large and imbalanced financial datasets that require precision and efficiency.

4.3. LightGBM

LightGBM (Light Gradient Boosting Machine) is a gradient boosting framework that excels in handling large datasets efficiently. Unlike traditional boosting algorithms, LightGBM uses techniques like Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to reduce memory usage and speed up training without compromising performance. These techniques make LightGBM especially effective for big datasets, as it scales linearly with data size while maintaining fast computation times.

Like XGBoost and CatBoost, LightGBM is also known for handling imbalanced datasets well by adjusting the training process to focus more on the underrepresented classes. It employs a method called ”binary tree construction” that grows trees leaf-wise, which can significantly reduce overfitting, particularly when dealing with large datasets. Additionally, LightGBM is capable of handling missing values automatically by using a built-in mechanism that allows the model to learn optimal splits even when some data points are missing, making it particularly useful in real-world financial applications where data incompleteness is common.

5. Results

5.1. XGBoost

5.1.1. Performance Metrics : . The XGBoost model demonstrated high efficiency, achieving an F1 Score of 0.8001, which suggests a balanced precision and recall — both vital for maintaining accuracy in binary classification tasks. Precision was notably consistent at 0.8000, mirroring the F1 Score closely, while the recall matched the precision at 0.8001. The model’s overall accuracy stood at 89.74%, underlining its capability to correctly classify the majority of instances. The Area Under the Receiver Operating Characteristic (ROC) Curve (AUC-ROC) was 0.9569, indicating excellent model performance with a high true positive rate and a low false positive rate across various thresholds. The Log Loss was calculated to be 0.2297, reflecting the model’s effectiveness in probability estimation for class prediction.

5.1.2. ROC Curve : . The ROC Curve for the XGBoost model showcased a curve significantly above the diagonal reference line, which represents random guessing, further validating the model’s discriminative power. The ROC Curve is shown in Figure 8.

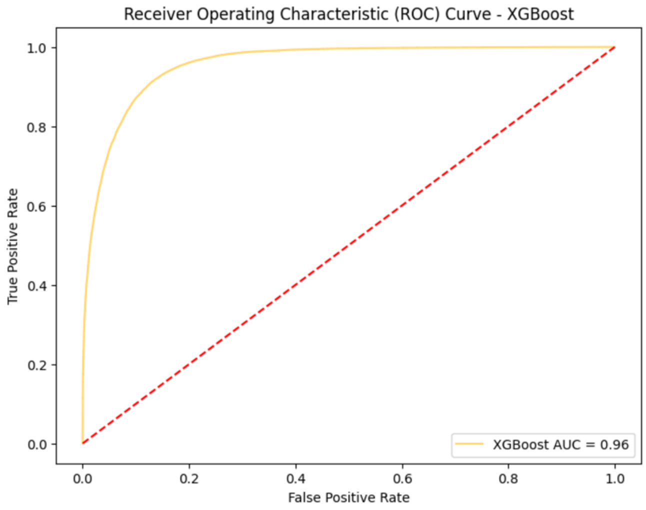

Figure 8. ROC Curve for XGBoost Model

5.1.3. Confusion Matrix : . The confusion matrix revealed a high number of true positives (18,837) and true negatives (63,531), affirming the model’s capability to distinguish between the classes effectively. The matrix also highlighted a relatively low occurrence of false positives and false negatives, which is crucial for maintaining the reliability of predictive insights in practical applications. The confusion matrix is illustrated in Figure 9.

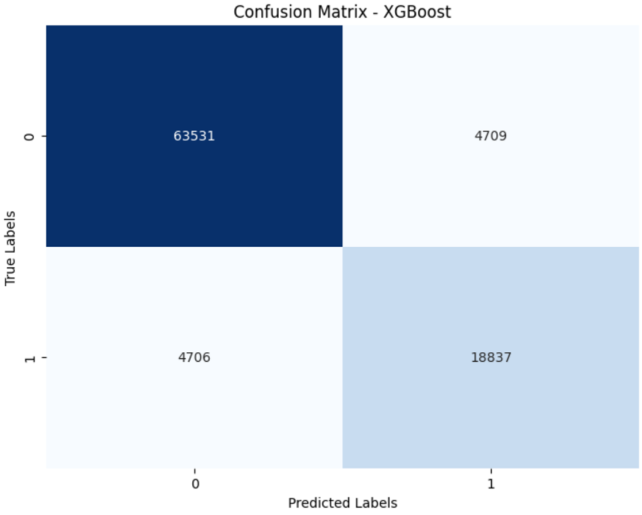

Figure 9. Confusion Matrix for XGBoost Model

5.2. CatBoost

5.2.1. Performance Metrics : . The CatBoost model slightly outperformed the XGBoost in several key metrics, indicating its robustness in handling categorical data. It achieved an F1 Score of 0.8070, Precision of 0.8094, and Recall of 0.8047, all of which are improvements over the XGBoost model, suggesting better balance and accuracy in the binary classification. The Accuracy was enhanced at 90.13%, and the AUC-ROC increased to 0.9593, highlighting superior predictability and the model’s capacity to differentiate between classes with high confidence. The Log Loss was reduced to 0.2227, indicating a more precise probability estimation by the model.

5.2.2. ROC Curve : . The ROC Curve for the CatBoost model further emphasized its superior performance, with a curve that stayed close to the upper left corner, reflecting a high true positive rate across various thresholds. The ROC Curve is depicted in Figure 10.

5.2.3. Confusion Matrix : . The confusion matrix for the CatBoost model depicted an effective classification with an increased number of true positives (18,945) and true negatives (63,779) compared to the XGBoost model, affirming the model’s efficacy in correctly predicting the class labels. The confusion matrix is shown in Figure 11.

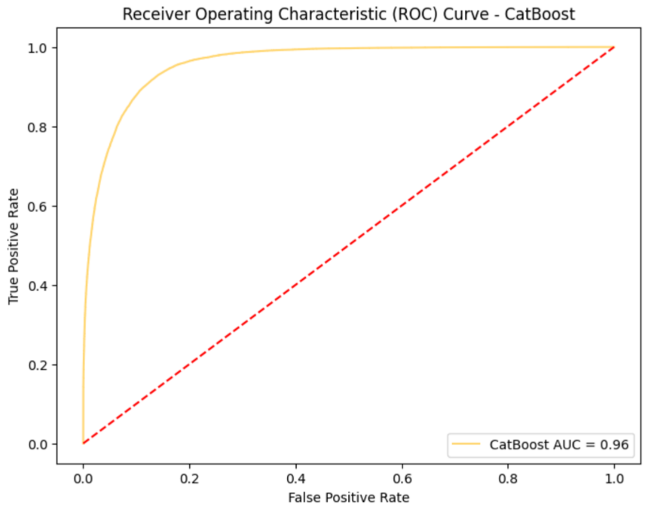

Figure 10. ROC Curve for CatBoost Model

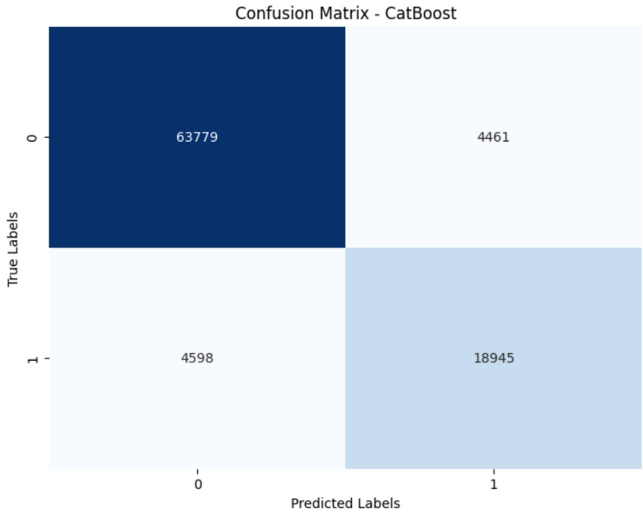

Figure 11. Confusion Matrix for CatBoost Model

5.3. LightGBM

5.3.1. Performance Metrics : . LightGBM showed commendable results with an F1 Score of 0.8043, indicative of a strong balance between precision and recall, crucial for the reliable classification in skewed datasets. The precision achieved was 0.8001, closely aligning with its F1 score, and a recall of 0.8084, suggesting a slightly better sensitivity towards positive class detection compared to XGBoost. The accuracy stood at 89.91%, which is competitive with the other models evaluated, asserting its robustness in general classification tasks. The AUC-ROC reached 0.9582, nearly matching that of CatBoost, which underscores its capability to perform well under various threshold settings. The log loss was recorded at 0.2259, showing efficient probability predictions.

5.3.2. ROC Curve : . The ROC curve for LightGBM exhibited an excellent performance characteristic with a curve well above the diagonal line, indicative of a high true positive rate and a low false positive rate, similar to the other models discussed. The ROC curve is shown in Figure 12.

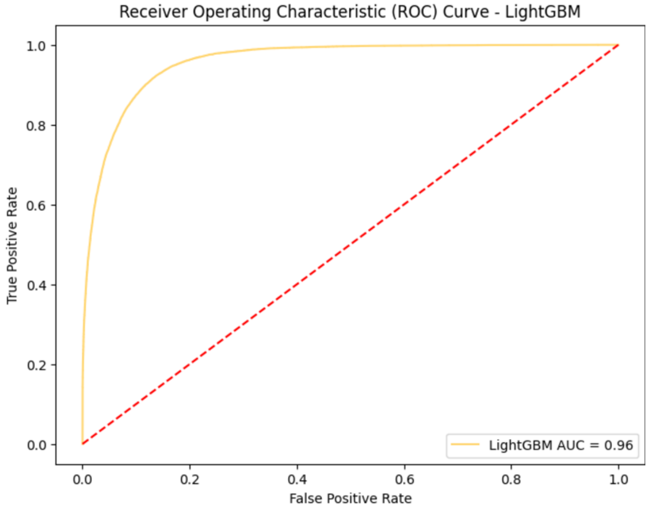

Figure 12. ROC Curve for LightGBM Model

5.3.3. Confusion Matrix : . The confusion matrix displayed an effective classification capability, with a significant number of true positives (19,033) and true negatives (63,485). This matrix also highlighted the model's ability to minimize false positives (4,755) and false negatives (4,510), which is essential for applications where the cost of misclassification is high. The confusion matrix is shown in Figure 13.

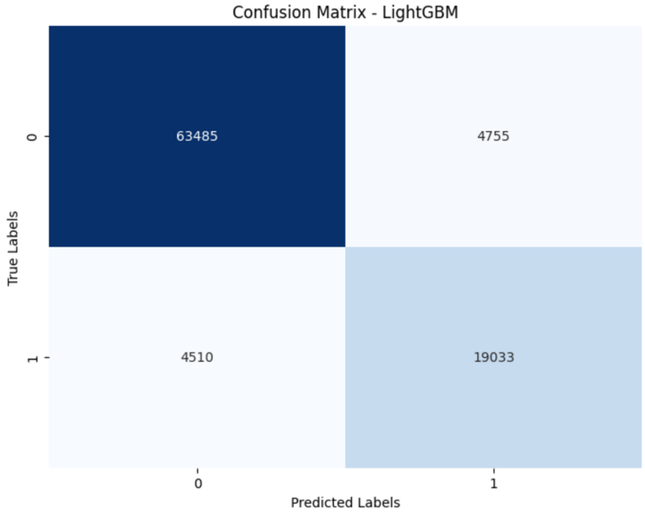

Figure 13. Confusion Matrix for LightGBM Model

5.4. Discussion

The comparative analysis of XGBoost, CatBoost, and LightGBM reveals distinct advantages for each model, depending on dataset characteristics and task requirements. CatBoost excels in handling categorical features without extensive preprocessing, making it ideal for datasets with abundant categorical variables, such as customer behavior analytics. LightGBM demonstrates strong performance with competitive metrics, offering advantages in scenarios with limited computational resources and time, while XGBoost stands out for its reliable performance in general classification tasks.

Overall, the models' high AUC-ROC scores and balanced precision-recall metrics highlight their effectiveness in financial prediction tasks, where accurate class distinction is crucial. Each model provides unique benefits—LightGBM offers slightly better recall, while CatBoost excels in precision and log loss. These strengths allow for tailored model selection based on operational constraints, ensuring the best-fit solution for specific use cases in financial applications.

6. Related Work

Several research studies have explored the application of machine learning techniques to predict credit card defaults, aiming to enhance risk management and decision-making in financial institutions.

One study by [3] Li, Junhong and Kang, Jijia Wu, Jie et Wang, ”Research on Credit Card Default Repayment Prediction Model,” compares various machine learning models, including XGBoost and Extreme Learning Machine (ELM), for predicting credit card default repayment. The findings indicate that both XGBoost and ELM models demonstrate superior predictive performance, offering valuable insights for commercial banks in developing effective frameworks to address credit card default payments.

Another research by [4] Gaganis, C., Papadimitri, P., Pasiouras, F. et al, ”Social Traits and Credit Card Default: A Two-Stage Prediction Framework,” introduces a two-stage framework that incorporates collective social traits to predict credit card delinquencies. By segmenting the market into homogeneous sub-populations based on social traits, the study enhances the accuracy of delinquency prediction models, highlighting the importance of social factors in credit risk assessment.

Another work by [1] H. Yash, Affan, K. Saurav and S. S. Dhanda, ”Credit Card Default Prediction Using Machine Learning Models,” the authors propose a machine learning-based strategy to predict credit card default using a dataset of credit card clients from Taiwan. The study evaluates various machine learning algorithms, demonstrating the effectiveness of these models in accurately predicting default probabilities, thereby assisting banks and credit card companies in better managing risk and optimizing lending strategies.

The research work by [2] Husejinovic, Admel and Kečo, Dino and Masetic et Zerina, ”Application of Machine Learning Algorithms in Credit Card Default Payment Prediction” evaluates the performance of machine learning methods, including logistic regression, decision trees, support vector machines, and ensemble learning methods, in predicting credit card default payments. The study highlights the potential of these algorithms to improve the prediction of default payments, aiding financial institutions in minimizing credit risk.

These studies collectively underscore the growing importance of machine learning techniques in predicting credit card defaults, offering valuable insights and methodologies for enhancing credit risk assessment and management in the financial sector.

7. Challenges

This project came with its own set of challenges, just like any other project. Identifying and solving these challenges played a vital role in executing the project smoothly and successfully. The challenges encountered during this project were:

7.1. Handling Missing Values and Imbalances

The data offered by American Express contained a lot of missing values, and the dataset was imbalanced. This required extensive data preprocessing to balance the dataset and drop columns or rows that contained mostly missing or null values.

7.2. Optimizing Dataset Size and Feature Selection

The dataset provided by Amex was too large to process, with the combined size of the training and testing datasets being over 50 gigabytes. To address this, we used a smaller feather dataset that contains only the last credit card statements of each of the 450,000 unique Amex customers. This change made the execution of the project simpler and more feasible.

7.3. Addressing Data Quality and Feature Redundancy

The columns in the dataset were unlabelled, making it more challenging to identify which columns were more important. To solve this, we used a correlation technique to determine the relationships between variables. This helped us identify columns that were identical or redundant, enabling us to discard redundant features and reduce the dataset size.

8. Conclusion

This study addressed the challenge of credit card default prediction by employing advanced machine learning models, including XGBoost, CatBoost, and LightGBM. The methodology involved robust data preprocessing techniques to manage missing values, class imbalance, and categorical features while focusing on the most recent customer records to ensure data relevance. Exploratory Data Analysis (EDA) revealed critical patterns in customer behavior, guiding effective feature selection and model development. Among the models, CatBoost demonstrated superior performance in metrics like F1-score, precision, recall, accuracy, and AUC-ROC, showcasing its strength in handling categorical data and imbalanced datasets.

Despite achieving competitive results, challenges such as managing large datasets, handling missing values, and addressing feature redundancy persisted. Future enhancements could involve exploring neural networks and additional ensemble methods to capture more complex patterns, incorporating time-series analysis for improved temporal predictions, and using interpretability tools like SHAP or LIME to provide transparent insights into model decision-making. These steps would not only enhance predictive performance but also aid financial institutions in making informed and explainable credit risk assessments.

9. Division of work

Jeena Mole(G01460309) :

Responsibilities:

- Importing necessary libraries and packages (e.g., `pandas`, `numpy`, etc.).
- Loading the training, test, and submission datasets.
- Cleaning the datasets by grouping and setting the index.
- Handling missing values in the dataset.
- Plot graphs to understand feature distributions and their relationships with the target variable.

Parshwa Gandhi(G01511122) :

Responsibilities:

- Perform exploratory data analysis (EDA) to identify key patterns in the dataset.
- Generate visualizations such as correlation heatmaps, KDE plots, and custom visualizations.
- Plot graphs to understand feature distributions and their relationships with the target variable.
- Use the trained models to make predictions on the test set and prepare submission files.

Krish Sanghvi(G01521041) :

Responsibilities:

- Perform exploratory data analysis (EDA) to identify key patterns in the dataset.
- Implement machine learning models such as `XGBoost`, `CatBoost`, and `LightGBM`.
- Train the models on the preprocessed data and evaluate them using appropriate metrics (e.g., F1-score, AUC-ROC).
- Use the trained models to make predictions on the test set and prepare submission files.

References

[1] H. Yash, Affan, K. Saurav, and S. S. Dhanda, “Credit Card Default Prediction Using Machine Learning Models,” in *2023 3rd International Conference on Innovative Sustainable Computational Technologies (CISCT)*, Dehradun, India, 2023, pp. 1–5. doi: 10.1109/CISCT57197.2023.10351316.

[2] A. Husejinovic, D. Kečo, and Z. Masetic, “Application of Machine Learning Algorithms in Credit Card Default Payment Prediction,” *International Journal of Scientific Research*, vol. 7, no. 10, pp. 425–426, 2018. Available: https://ssrn.com/abstract=3848590.

[3] J. Li, J. Kang, J. Wu, H. Wang, and X. Yang, “Research on Credit Card Default Repayment Prediction Model,” Available at SSRN, 2023. Available: https://ssrn.com/abstract=4732387.

[4] C. Gaganis, P. Papadimitri, and F. Pasiouras, “Social traits and credit card default: a two-stage prediction framework,” *Annals of Operations Research*, vol. 325, pp. 1231–1253, 2023. doi: 10.1007/s10479-022-04859-1.