AMERICAN EXPRESS CREDIT DEFAULT PROJECT

Table of Contents

1. Introduction

Business Context

Credit cards have become a cornerstone of modern financial transactions, offering a blend of convenience and security that is unparalleled. The global embrace of credit cards has not only propelled consumer spending but also fostered business growth by streamlining payment processes.

American Express, with a market capitalization of $121.80 billion, is a key player in this sector, boasting a robust rewards system that has endeared it to both individuals and businesses. However, the proliferation of credit card usage also ushers in the risk of credit defaults, a challenge that is not unique to American Express but is a shared concern across the financial industry.

Recent data elucidates the gravity of credit defaults, with credit card delinquencies reaching 3.8% and credit card and car loan defaults hitting a 10-year high in 2023 (Equifax, 2023 as cited in NY Post, 2023). The scenario is further exacerbated as credit card balances soared by $45 billion from Q1 to Q2 2023, marking a troubling 4.6% quarterly increase and pushing the total credit card debt past the $1 trillion mark (Federal Reserve Bank of New York, 2023). Moreover, the aftermath of aggressive loan growth has set the stage for more pronounced changes in the consumer credit market in 2023 (TransUnion, 2023).

American Express has been navigating through these financial headwinds, as evidenced by its recent financial performances. In Q2 2023, the company reported a 12% rise in earnings per share to a record $2.89, with a net income of $2.2 billion (American Express Company, 2023a). Despite the challenges, the company's revenue in Q1 2023 increased by 22% to a record $14.3 billion, primarily driven by strong card member spending (American Express Company, 2023b). This resilient performance underscores the company’s adept management and innovative

strategies, a testament being its ranking as the No. 1 on Fast Company's list of the World's Most Innovative Companies for 2023 in the 'Personal Finance' category (American Express Company, 2023c).

In light of the above, this project aims to leverage data science to mitigate the risk of credit defaults for American Express. Specifically, we aspire to develop a predictive model to forecast the likelihood of a customer defaulting on their credit card payments, defined as failing to settle their balance for 120 days or more. Through this endeavor, we intend to provide actionable insights that could aid American Express in proactively managing credit default risks, thereby safeguarding its revenue streams and bolstering its financial stability in a fluctuating economic landscape.

A. Business Problem

As I mentioned above, credit defaults pose a significant risk to American Express, affecting its financial stability and the trust built with its clientele. Specifically, the failure of cardholders to settle their outstanding balances beyond 120 days results in financial losses and potentially tarnishes the company's credit rating. In the fiercely competitive financial market, a proactive approach to mitigate credit default risks is imperative for sustaining a robust customer base and ensuring financial health.

To address this challenge, our project aims to harness data science to develop a classification model using an American Express customers data from 2021. Leveraging customer characteristics, this model intends to accurately predict the likelihood of a customer defaulting on their credit card payments. This initiative is a stride towards enabling American Express to devise targeted strategies for preventing credit defaults, thereby fostering a more financially secure and trust-worthy customer-issuer relationship.

2. Data Wrangling

In this stage, we load the data from a csv file, we then perform techniques to clean the data so that it is optimally designed for futures steps such as EDA, preprocessing and modeling. The dataset contained 45,528 records spread across 19 columns, meaning, there are 18 variables to consider as possible features for predictions. Preliminary analysis using methods like **info()** and **describe()** showcased a blend of object, int64, and float64 data types. Some columns were found to have missing data. Attributes in the dataset encompass **customer_id, age, gender, owns_car, owns_house, no_of_children, net_yearly_income, occupation_type, credit_limit, credit_score**, amongst other variables.

To get a tidy dataset for exploratory data analysis, we needed to standardize nomenclatures, profile it and handle some of the missing values if possible. Diving deeper into the nullity of the matrix presented in the data, several columns had missing entries, with **no_of_children** having the highest percentage of missing values (1.7%). Below is a figure taken out of the pandas profiling report I created as HTML file to portrait an overview of the data.

Overview | Alerts 43 | Reproduction

Dataset statistics

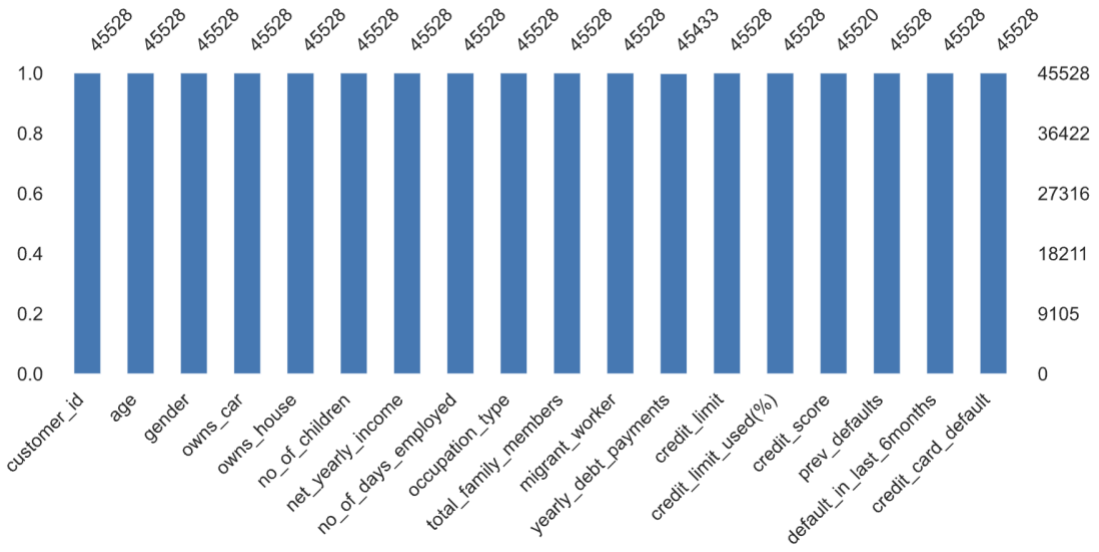
Number of variables	18
Number of observations	45528
Missing cells	103
Missing cells (%)	< 0.1%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	6.3 MiB
Average record size in memory	144.0 B

Variable types

Numeric	9
Categorical	7
Boolean	2

The columns **owns_car**, **no_of_days_employed**, **yearly_debt_payments**, **migrant_worker**, and **total_family_members** also had missing values, ranging from 0.18% to 1.2%. The column **name** was dropped as it was inconsistent and redundant due to the presence of **customer_id**. The **gender** column had an 'XNA' value, which was imputed with the mode of the gender column. The **occupation_type** column had 31.4% of 'Unknown' values, which might need special attention in subsequent steps.

The missing values, in categorical variables, like **owns_car**, **migrant_worker**, and **total_family_members** were imputed with mode. For the numerical variable **no_of_days_employed**, missing values were imputed with the median of **no_of_days_employed** grouped by **occupation_type**. We decided to proceed to handle some of the missing values as the nullity was quite low, thus, the overall development of the project wouldn't be affected by the techniques applied. Below is a graphical representation of the null values found in the data, the figure was taken out of my pandas' profile which can be found on the reports folder under figures.

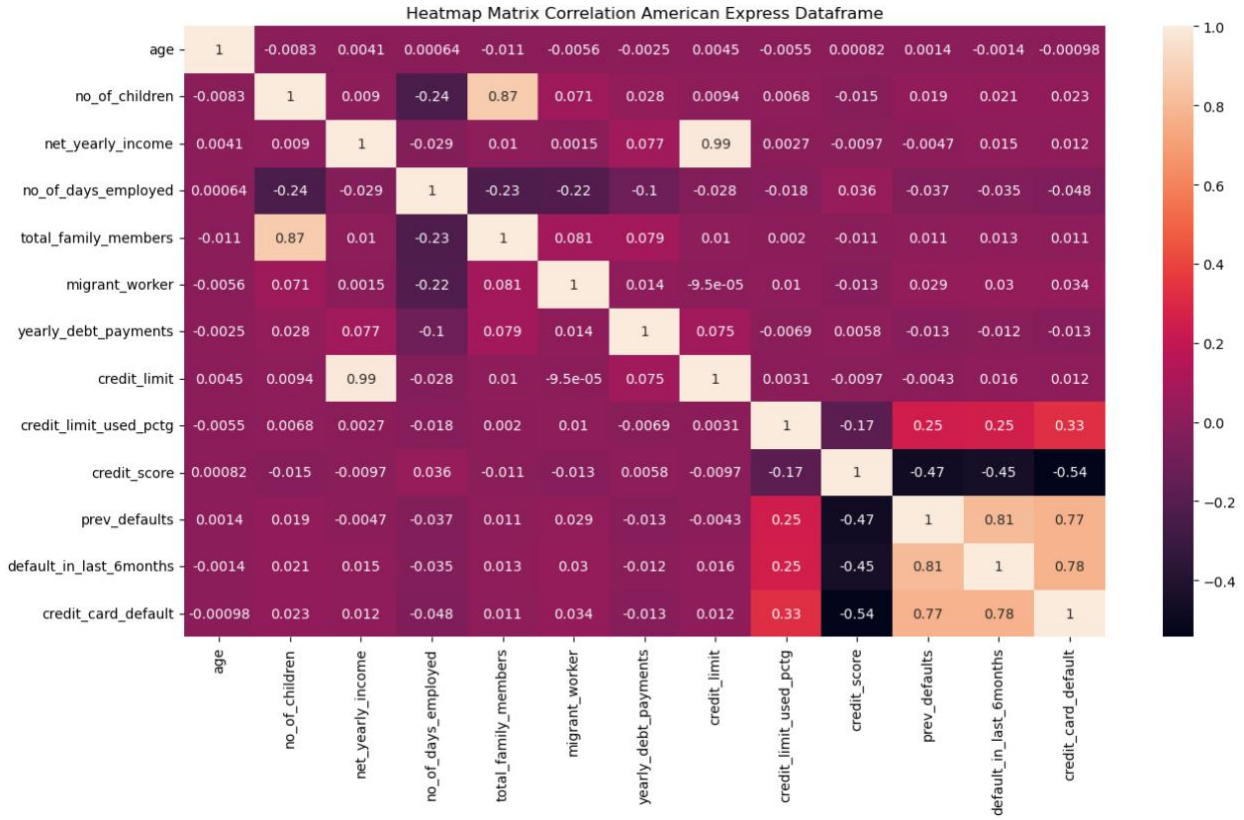


3. Exploratory Data Analysis

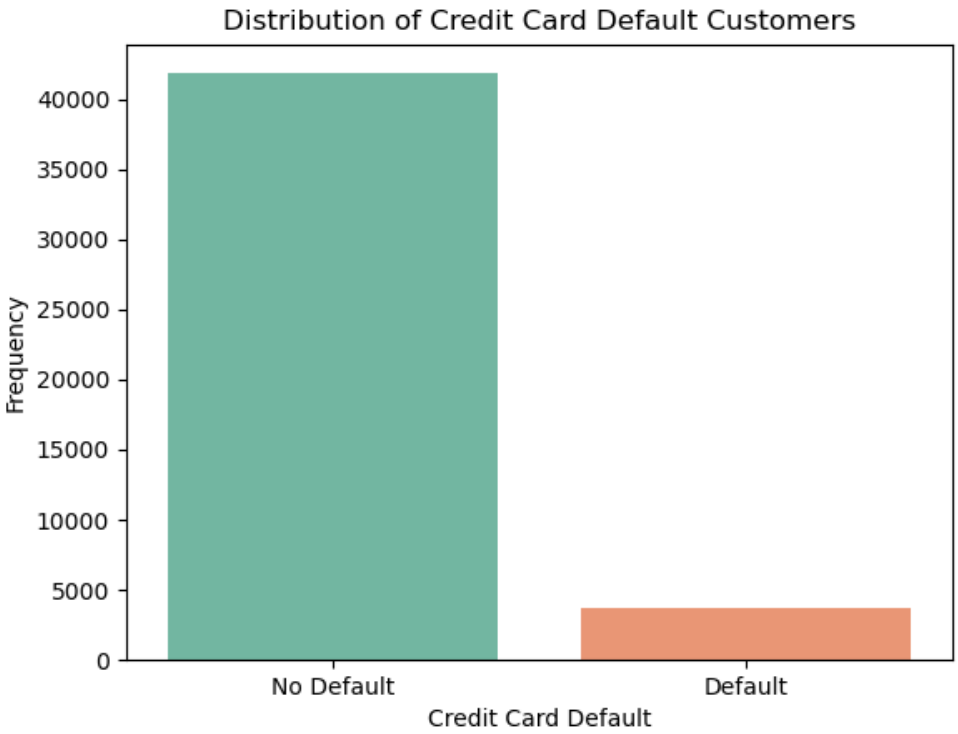
In this stage, it is crucial to visualize the dataset, therefore, it will show different plots that supported my exploration of the American Express dataset. The method .describe() offered a foundational understanding of the inherent dynamics within the data:

- The average age of individuals in this dataset is approximately 39 years, with a standard deviation of around 9.54 years, indicating a moderate spread around the mean. This spread represents a relatively diverse age group, which is essential for understanding different credit behaviors across different life stages.
- The standard deviation in **net_yearly_income** is considerably high, showing a value of approximately $669,074.03, which indicates a vast disparity in income levels among the individuals. The minimum and maximum values further elucidate this point, with the range stretching from $27,170.61 to a staggering $140,759,000. This vast range suggests the presence of outliers or high-income individuals, which could significantly skew the average income upwards.
- Examining the credit behavior, the credit limit used percentage has a mean of 52.23%, indicating that, on average, individuals are utilizing slightly more than half of their available credit.

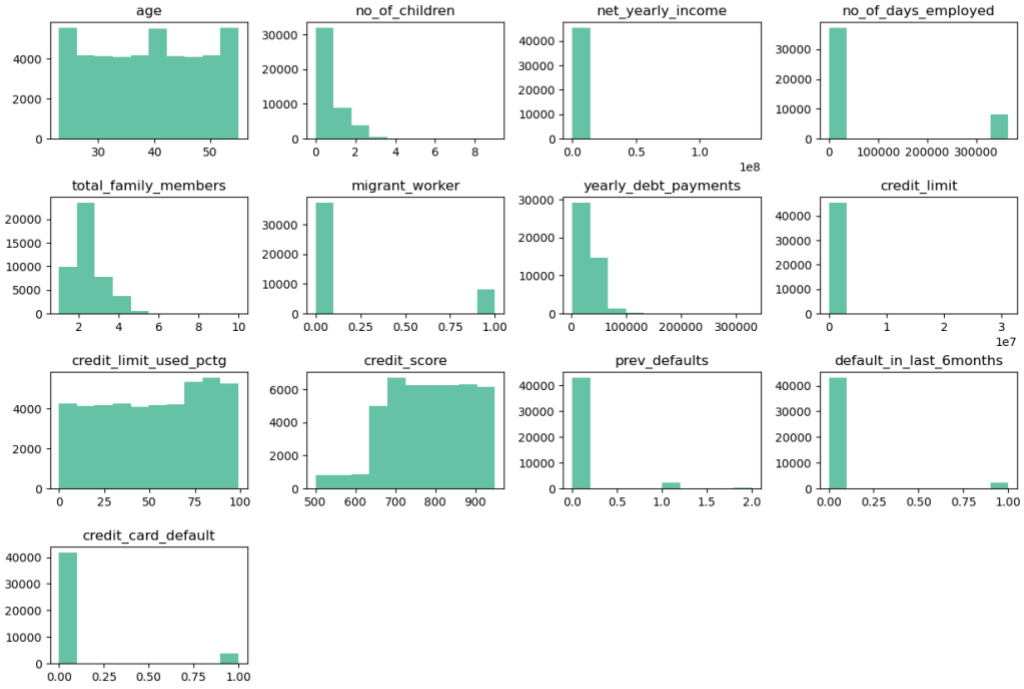
The insights above paved the way for a more in-depth analysis though visual exploration of variables. One of the takeaways of the visualizations is the presence of multicollinearity which indicates high correlation between dependent variables. This phenomenon can impact the development of the model and its performance to unseen data; thus, it is crucial to handle it properly while preprocessing the data for modeling. The following figure shows the correlation between these dependent variables: **net_yearly_income** and **credit_limit**, **total_family_members** and **no_of_children,** and **prev_defaults** and **default_in_last_6months**.

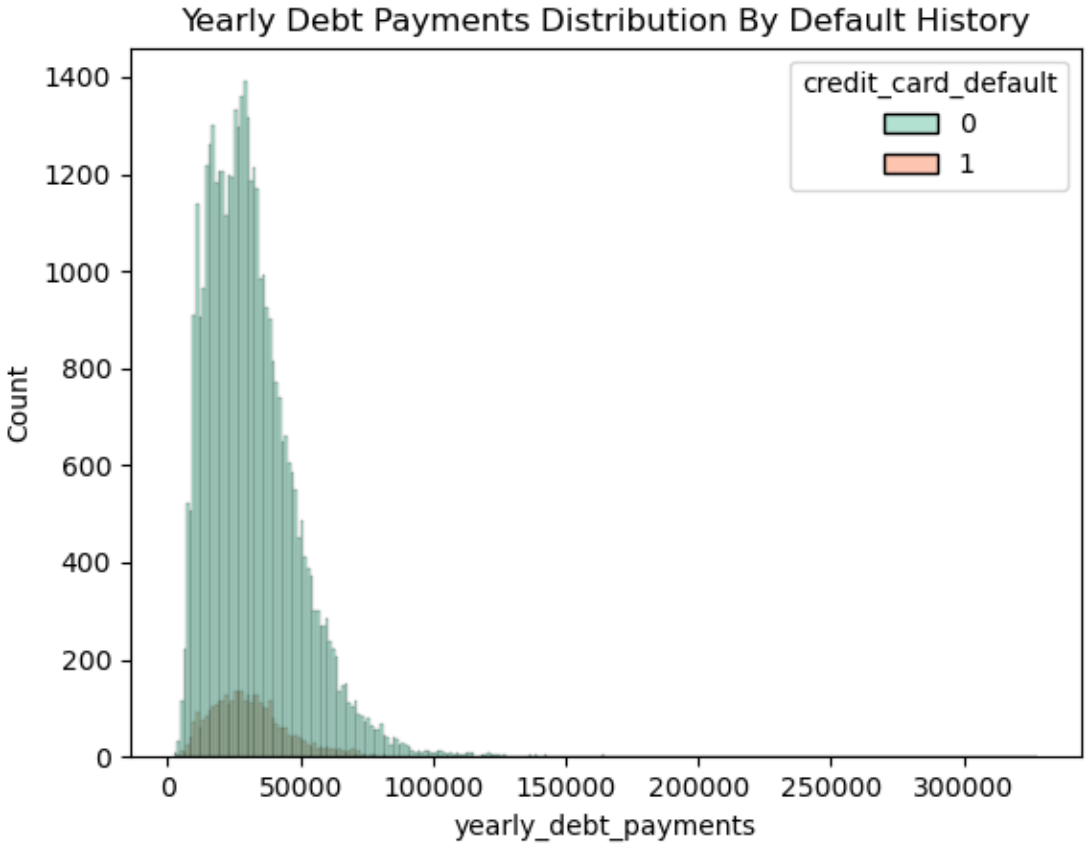


Furthermore, by plotting the distribution of the dependent variable **credit_card_default** as a countplot, the dataset also shows a strong class imbalance issue which will need to be solved in the next step before modeling. Imbalanced classes in a credit card default classification project are common due to the nature of the data. However, it should be balanced so that the model is properly trained without biases towards the majority class.



The dependent variable shows signs of Bernoulli distribution if we consider the mean and standard deviation. Also, since the Bernoulli distribution is a discrete probability distribution of a binary variable, it matches with the nature of the **credit_card_default** variable amongst other in the American Express dataset. To visualize it, I performed a for loop to plot histograms of the variables:

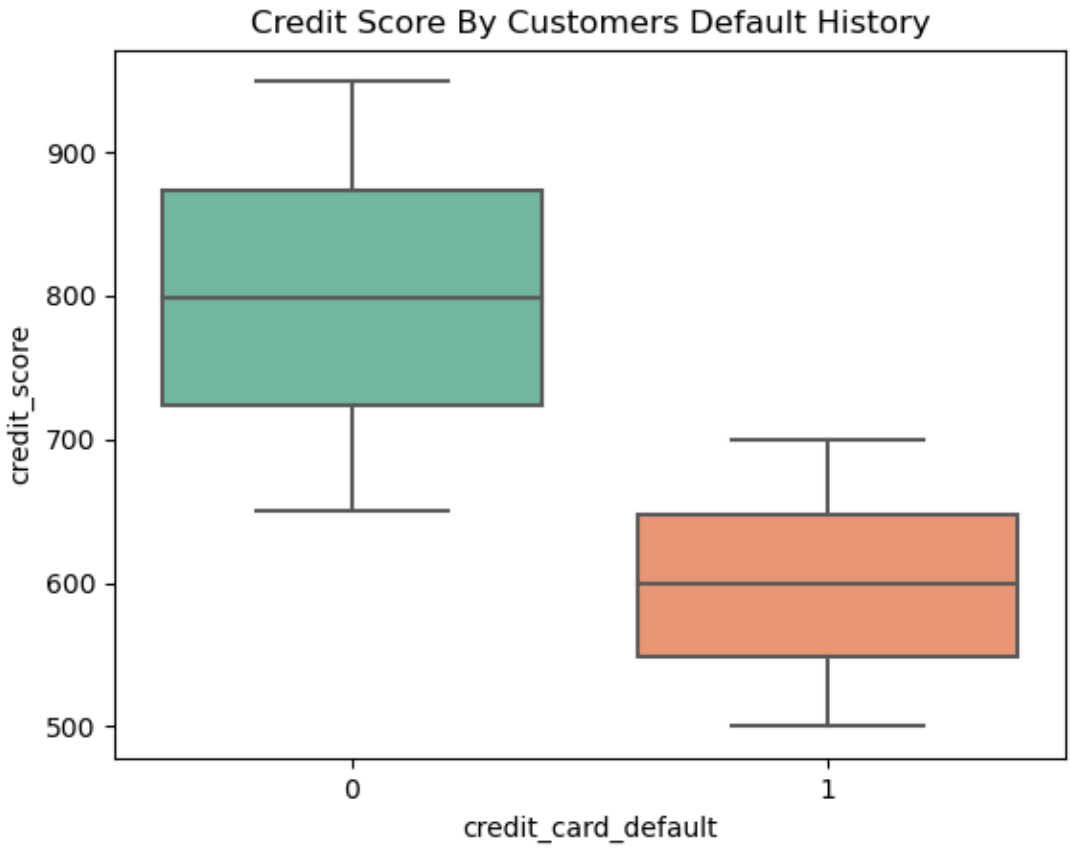




The histograms show a visual representation of the distribution among variables. We

further analyzed **yearly_debt_payments** by plotting a histogram with the imbalanced classes. We can observe how both classes follow a Bernoulli distribution as suspected. This is an important insight to keep in mind in the modeling stage as some algorithms perform better if a distribution is specified in their hyperparameters. Several other variables were analyzed by visualizing them, the process can be found in this notebook: EDA American Express.

Below there is a boxplot of **credit_score**, we decided to show this visualization as it is usually relevant to credit card defaults reports. The boxplot represents a relatively high average credit score among the individuals, which is a positive indicator of creditworthiness.



4. Data Preprocessing and Training

In this project, we navigated through several critical stages of data preprocessing for our American Express Default Classifier model, ensuring our data was not just fit for use, but primed for efficiency. The steps in the stage are explained below:

a. Multicollinearity: Feature Selection

In our analysis, we identified substantial correlations between **net_yearly_income** and **credit_limit**, as well as **total_family_members** and **no_of_children**. To mitigate multicollinearity's effects, we dropped **credit_limit** and **no_of_children**, reducing redundancy without compromising the information quality.

Furthermore, we encountered a unique scenario with **prev_defaults** and **default_in_last_6months**, two variables significantly correlated with each other and the dependent variable, **credit_card_default**. Instead of dropping one, we ingeniously combined them, creating a composite feature encapsulating both variables' information. This strategy not only preserved valuable information but also enhanced our model's interpretability by providing a more comprehensive view of a customer's default history.

b. Scaling Dataset: MinMaxScaler()

The next phase of our preprocessing was feature scaling, an essential step to harmonize the range of different features, ensuring that no particular feature dominates the model due to its scale. We employed MinMaxScaler, a technique that re-scales features to a uniform range, typically between 0 and 1, while maintaining the original distribution and relationships amongst variables.

By scaling features like **net_yearly_income**, which originally had a much larger range compared to others like 'age', we provided our model with a balanced dataset, where each feature had an equal opportunity to influence the model. This uniformity is particularly crucial for algorithms that calculate distances between data points, like k-NN, which we will use as part of our next stage, Modeling.

c. Class Imbalance: SMOTE

The final challenge we addressed was class imbalance, a common problem in classification tasks where the classes are not represented equally. As mentioned before, the **credit_card_default** instances were significantly fewer than the non-default cases. Such imbalance could lead models to over-predict the majority class and undermine the minority class's predictive power.

To counter this, we utilized SMOTE (Synthetic Minority Over-sampling Technique), an innovative oversampling method that synthesizes new examples in the minority class. SMOTE works by selecting similar instances and altering them slightly, avoiding exact duplicates, thus enriching our dataset with more **credit_card_default** cases without adding bias.

5. Modeling

a. Metrics:

In predictive modeling, the choice of evaluation metrics is critical, especially in binary classification tasks like predicting customer credit card default. We used AUC-ROC and AUC-PRC as primary metrics. AUC-ROC assesses the trade-off between sensitivity and specificity, ideal for imbalanced datasets, providing a measure for the model's discriminative capacity. However, it can be optimistic in extremely imbalanced cases, hence the additional use of AUC-PRC.

AUC-PRC evaluates the trade-off between precision and recall, important when false positives carry high costs, or the focus is on the positive class. Despite AUC-PRC's relevance, we chose AUC-ROC as the baseline for model selection due to its robustness across various thresholds and class distributions, offering a comprehensive view of model performance.

b. Hyperparameter Tunning:

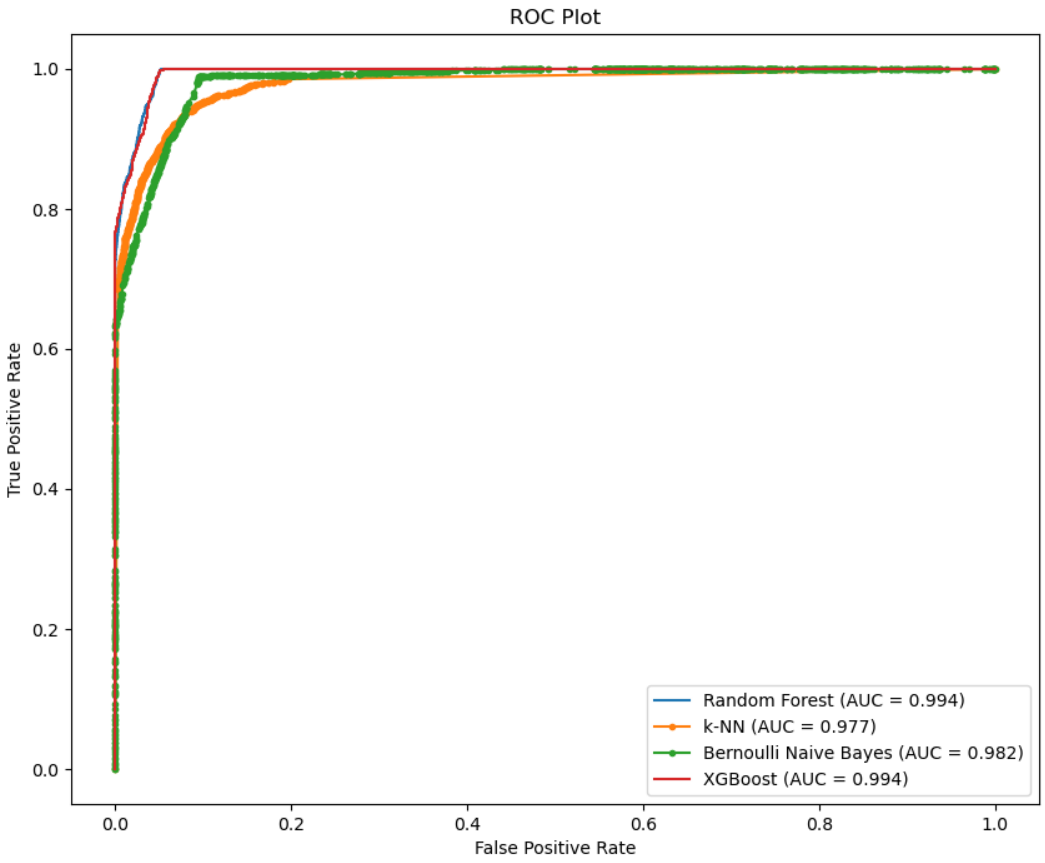
Hyperparameter tuning, crucial for optimizing a model's learning characteristics, was a key step in enhancing model performance. We used RandomizedSearchCV, an efficient tool for this process, sampling from a predefined space of hyperparameters, saving computational resources compared to exhaustive methods like GridSearchCV. This approach was vital for navigating the high-dimensional space, enhancing model performance without exhaustive computational demand.

c. Model Selection:

The final model selection was based on AUC-ROC scores, interpretability, computational efficiency, and generalizability. The XGBoost Classifier was selected due to its high performance in cross-validation stages. The framework can handle various data types and incorporates techniques to prevent overfitting. Its efficiency, scalability, and robustness to imbalanced datasets like ours make it suitable for predicting customer defaults. Therefore, XGBoost was chosen for its performance, interpretability, and flexibility in handling complex predictive tasks.

Model AUC-ROC Scores

Model	AUC-ROC
k-NN	0.976874
Bernoulli Naive Bayes	0.982254
XGBoost	0.994303
Random Forest	0.994454



6. Conclusions and Recommendations

Conclusions:

- **Prevalence of Credit Defaults:** the analysis underscores the critical challenge that credit defaults pose in the contemporary financial landscape. As credit card delinquencies and defaults are on an upward trajectory, the need for advanced predictive mechanisms cannot be overstated. This project has illuminated the exigency for sophisticated models to preempt potential credit risks.

- **Robustness of XGBoost Model:** through rigorous evaluation, the XGBoost Classifier has proven to be the most efficacious model, demonstrating not only superior performance metrics in cross-validation but also an impressive blend of computational efficiency and model generalizability. These attributes are indispensable when transitioning from a theoretical model to real-world applications where scalability and adaptability are paramount.

- **Significance of Hyperparameter Tuning:** the incorporation of RandomizedSearchCV in the hyperparameter tuning stage has distinctly amplified the model's predictive performance. This enhancement underscores the pivotal role of meticulous hyperparameter optimization in the machine learning development lifecycle.

Recommendations:

- **Continuous Model Evaluation and Improvement:** The financial domain is marked by its dynamism, with incessant shifts in consumer behavior patterns and macroeconomic indicators. It is imperative to institute a regimen of periodic model retraining with fresh datasets. This approach ensures the model's evolutionary adaptation to contemporaneous trends and sustains its predictive precision.

- **Feature Engineering and Exploration:** there is room for further exploration into alternative features that can potentially influence credit default predictions. Delving into economic indices or granular customer transaction details could unveil insights that significantly bolster the model's predictive prowess.

- **Risk Mitigation Strategies:** Beyond predictive modeling, these insights must be synthesized into comprehensive risk management strategies. These strategies may encompass setting prudent credit limits, initiating custom-tailored follow-up programs, or proffering financial literacy resources to clientele identified as high-risk.

- **Model Interpretability and Fairness:** it is important to ensure that the model's operational mechanisms are transparent and its decisions unbiased.

A concerted effort must be made to evaluate the model for any partiality to guarantee that its predictions do not unjustly disadvantage any demographic group.

- **Regulatory Compliance:** operating within the realm of credit and personal finance mandates stringent adherence to several regulations and privacy statutes. It is crucial to ascertain that all facets of the model and its deployment are in unwavering compliance with pertinent financial regulations and data protection laws.

7. References:

Equifax. (2023). Credit card and car loan defaults statistics. As cited in NY Post. Retrieved from https://nypost.com/2023/credit-card-and-car-loan-defaults-hit-10-year-high/

Federal Reserve Bank of New York. (2023). Total Household Debt Reaches $17.06 Trillion in Q2 2023; Credit Card Debt Exceeds $1 Trillion. Retrieved from https://www.newyorkfed.org/microeconomics/hhdc.html

TransUnion. (2023). More Pronounced Changes Expected in Consumer Credit Market in 2023. Retrieved from https://newsroom.transunion.com/more-pronounced-changes-expected-in-consumer-credit-market-in-2023

American Express Company. (2023a). Second-quarter earnings per share rose 12% to record $2.89. Retrieved from https://ir.americanexpress.com/earnings/earnings-releases-and-quarterly-reports/2023/2q2023

American Express Company. (2023b). First-quarter revenue increased 22% to record $14.3 Billion. Retrieved from https://about.americanexpress.com/press-releases/pr/2023/first-quarter-revenue-increased-22-to-record-14.3-billion

American Express Company. (2023c). American Express Ranks No. 1 on Fast Company's Most Innovative Companies for 2023. Retrieved from https://about.americanexpress.com/press-releases/pr/2023/american-express-ranks-no-1-on-fast-companys-most-innovative-companies-for-2023