Neela Ropp

Analyzing Credit Card Default Risk: Predictive Modeling and Insights

Summary

Credit card defaults pose a significant challenge for banks and financial institutions, because when people fail to make their payments on time, it can lead to financial loss for banks and lenders, and a damage to the users credit score and finances. In a world where credit cards are being increasingly used, it is becoming more and more important to conduct studies to try and predict who is likely to default in order to serve the best interest of lenders and borrowers. The goal of this study was to analyze a dataset on credit card defaults to try and better understand what features are the most important in predicting whether someone will default on their credit card, how do factors like age, gender, marital status, and education influence default risk, what patterns in payment behavior are most closely linked to defaults, and which machine learning models give us the best predictions. To answer these questions, I used the "Default of Credit Card Clients" dataset from the UC Irvine Machine Learning Repository. This dataset includes 30,000 records of credit card holders, each with 23 variables containing the demographics of the people and their credit card statement information. Now let's dive into how the study was conducted in order to answer the research questions, and explain why it was so important to understand the metrics that were used.

Problem Statement and Data Layout

Credit card defaults happen when a customer can't make the minimum payment on their account on time, putting both the customer and the bank at risk. In this study, the data was broken down into 23 columns in the following format: X1 (LIMIT_BAL): The amount of the given credit in NT dollars. This includes both the individual consumer credit and their family

(supplementary) credit, X2 (SEX): Gender of the client (1 = male; 2 = female), X3 (EDUCATION): Education level of the client (1 = graduate school; 2 = university; 3 = high school; 4 = others), X4 (MARRIAGE): Marital status of the client (1 = married; 2 = single; 3 = others), X5 (AGE): Age of the client in years, X6 to X11 (PAY_0 to PAY_6): History of past payments. These columns represent the repayment status for the last six months (PAY_0 being the most recent month), X12 to X17 (BILL_AMT1 to BILL_AMT6): Amount of bill statements in NT dollars for the last six months (BILL_AMT1 being the most recent month), X18 to X23 (PAY_AMT1 to PAY_AMT6): Amount of previous payments in NT dollars for the last six months (PAY_AMT1 being the most recent month), Y (default payment next month): Whether the client defaulted on their payment in the following month (1 = Yes, 0 = No).

Given all these attributes in our data it posed the challenge of sifting through a lot of data to figure out which factors are the best indicators of whether someone will default. My goal was to find these key indicators and build models that could help banks predict defaults more accurately. By using the "Default of Credit Card Clients" dataset, I aimed to build models that not only predict defaults but also offer insights into how these predictions can be applied in the real world.

Data Overview and Preparation

The dataset used was well-structured, but there were extraneous dummy variables in some of the demographic columns. I took this into account and made sure that we were only using data for 0,1 or 2 dummy variables to simplify the study. Then, I began by engineering several new features to capture more detailed patterns in the data. For instance, I calculated the average, maximum, and sum of payment statuses across six months. This allowed us to see not just whether a customer was late on a payment, but how often and by how much they were late.

Similar calculations were made for bill amounts and payment amounts across six months, providing a comprehensive view of each customer's financial behavior. When I began looking at my data though, I knew I quickly had to narrow down what methods would be best for understanding and visualizing my data.

Exploratory Analysis

With the data prepared, I conducted an exploratory data analysis to visualize and find statistics to see patterns about when people were defaulting and who was defaulting the most. Initially, I tried methods like k-mean clustering to see if any groups appeared to distinguish who would default, but I quickly realized histograms bar charts worked much better.

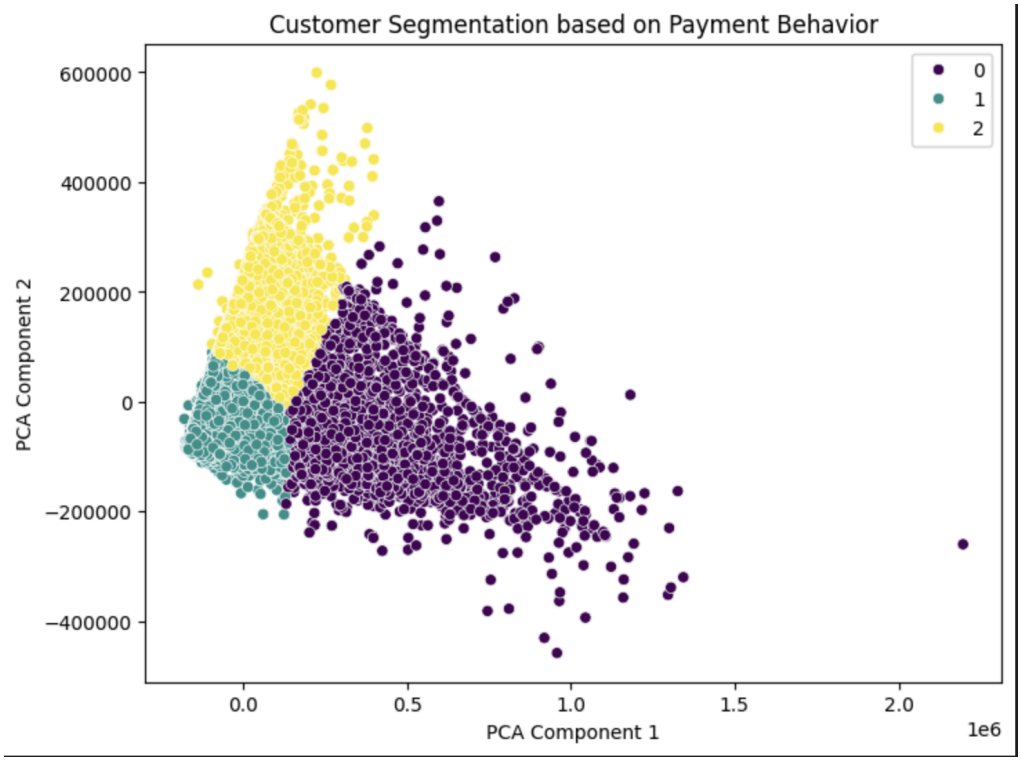

Figure 1. K-means clustering

The analysis revealed several key insights. First, we found that gender had a significant relationship with default status, with males slightly more likely to default than females.

Specifically, 22.63% of male customers defaulted, compared to 20.77% of female customers. This finding was supported by a chi-square test, which showed a significant relationship between gender and default status ($chi2 = 47.71$, p-value < 0.001). Education level also played a role in default risk. Customers with graduate-level education had a median credit limit of 200,000 NT dollars, significantly higher than those with only a high school education, who had a median credit limit of 90,000 NT dollars. This suggests that higher education levels correlate with better creditworthiness. Additionally, older customers generally had higher credit limits. For example, customers aged 60-69 had a median credit limit of 200,000 NT dollars, while those aged 20-29 had a median credit limit of 70,000 NT dollars.

Payment behavior emerged as the strongest predictor of default. The most recent payment status had a correlation of 0.32 with default, making it the most significant feature in predicting whether a customer would default. This means that customers who were late on their most recent payments were much more likely to default. We also found that customers who had high bill amounts but made low payments were more likely to default. For instance, one customer had a total bill amount of 17,077 NT dollars but made total payments of only 5,000 NT dollars,

indicating difficulty in managing their debt.

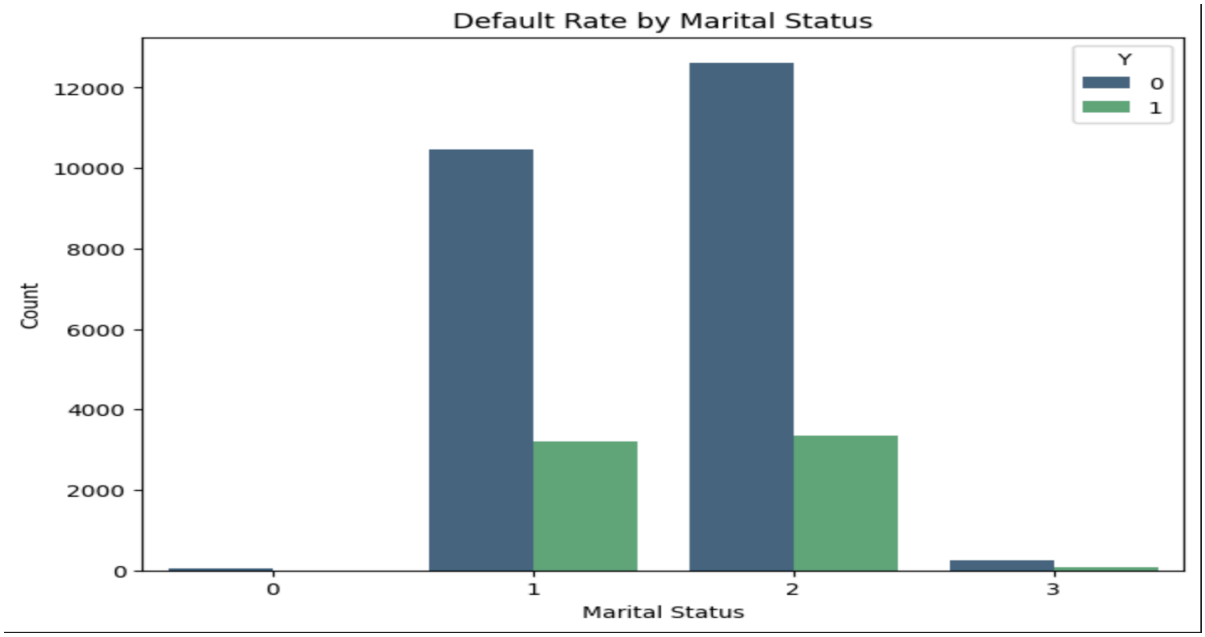

Figure 2. Marital Status and Credit Card Payments.

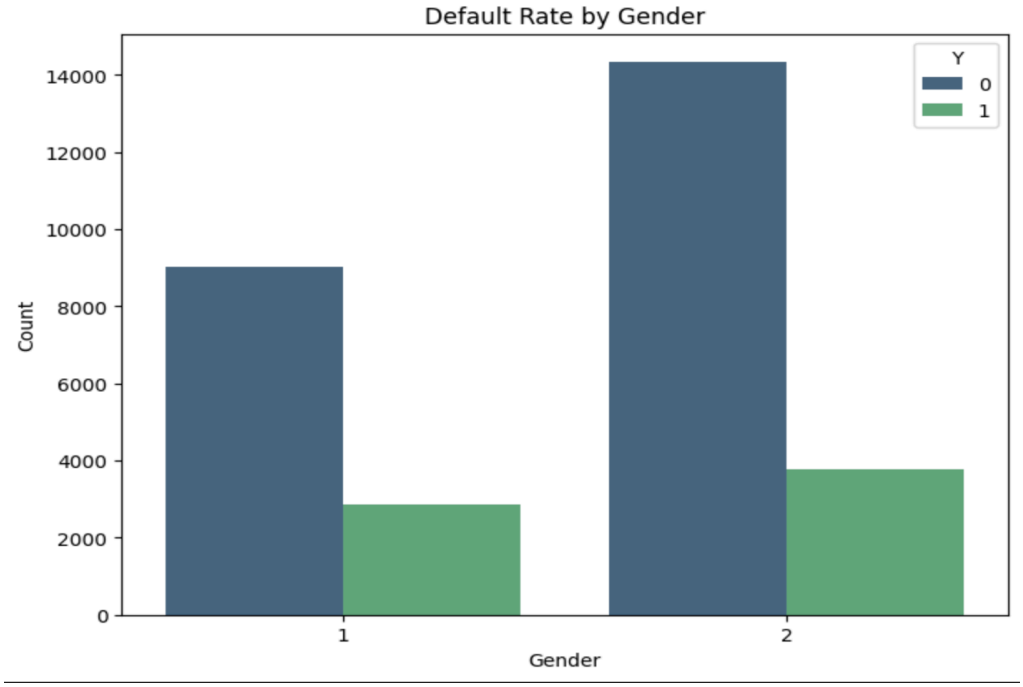

Figure 3. Gender and Credit Card Payments

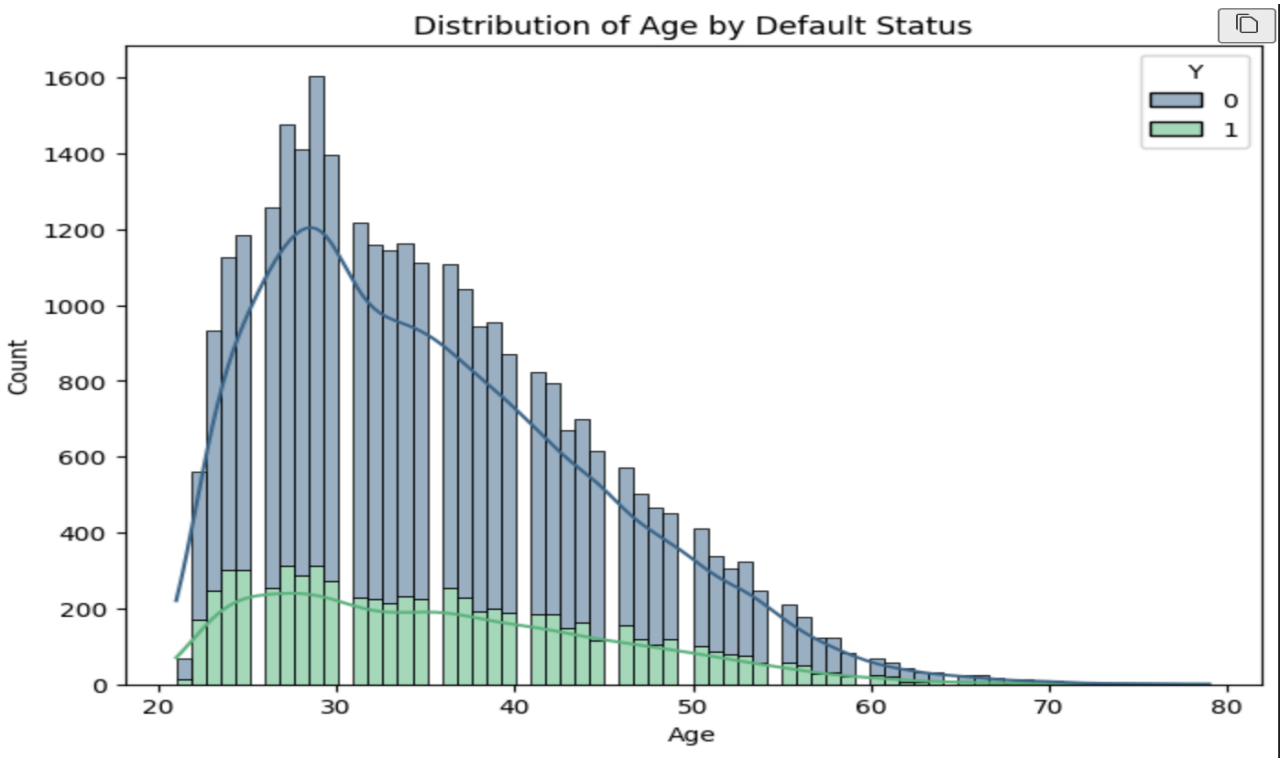

Figure 4. Age and Credit Card Payments

Model Selection and Evaluation

To predict defaults, we tested several machine learning models, including Logistic Regression, Random Forest, and Principal Component Analysis (PCA). The Logistic Regression model gave us an AUC-ROC score of 0.69, indicating a moderate ability to distinguish between defaulters and non-defaulters. However, the model faced some convergence issues, suggesting that further iterations or better data scaling might improve its performance. The model's accuracy was about 81%, with a precision of 0.62 for predicting defaults.

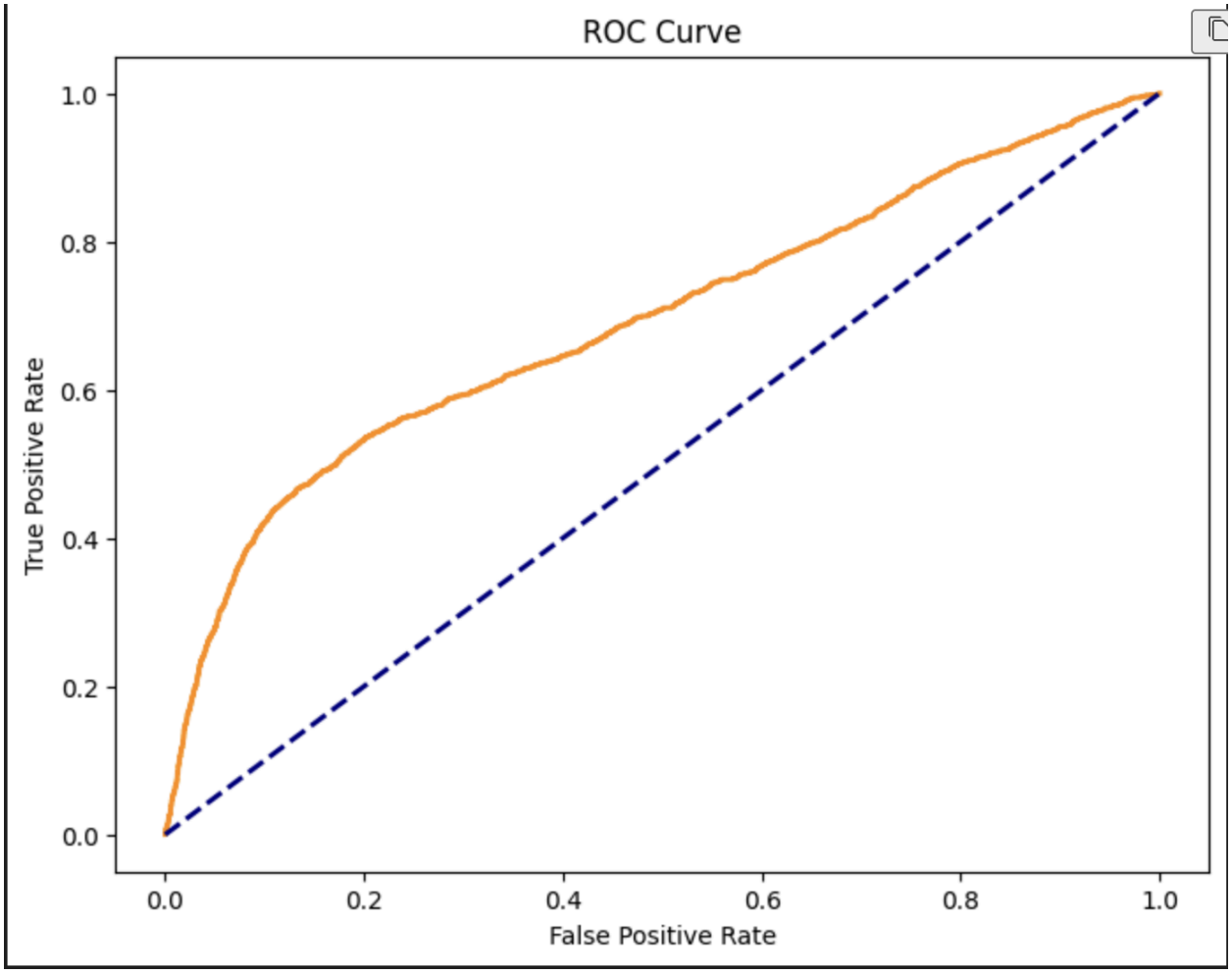

Figure 5. ROC curve

The Random Forest model, however, was the standout performer. It achieved an accuracy of 81% and identified key features like payment status, age, and credit limit as the most important predictors. Also important to note in the random forest study, the 6th column was noted to be the biggest indicator of whether or not someone would default, as it was the number of people who didn't make their payment the first month after it was due. Columns 7-11 were the next 5 months after where there was a steep decline in people who didn't pay their credit card payment. Closely following the 6th column were the fifth and first column. The 5th column was the person's age, and the first column was the amount of credit a person was given.

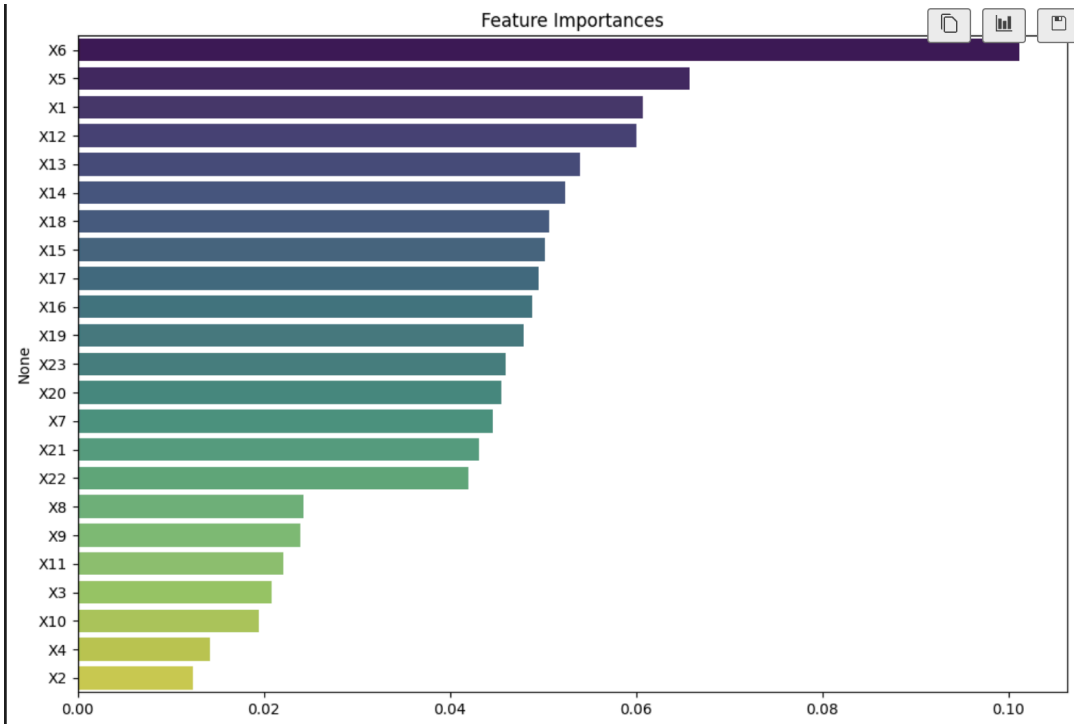

Figure 6. Feature Importance

The Gini impurity analysis showed that payment status was the top feature, with an importance score of 0.16. This model's ability to handle complex, non-linear relationships made it the best choice for this dataset.

PCA was used to reduce the dimensionality of the data, revealing that the first three components explained over 90% of the variance. This method helped us focus on the most relevant features, improving the performance of other models.

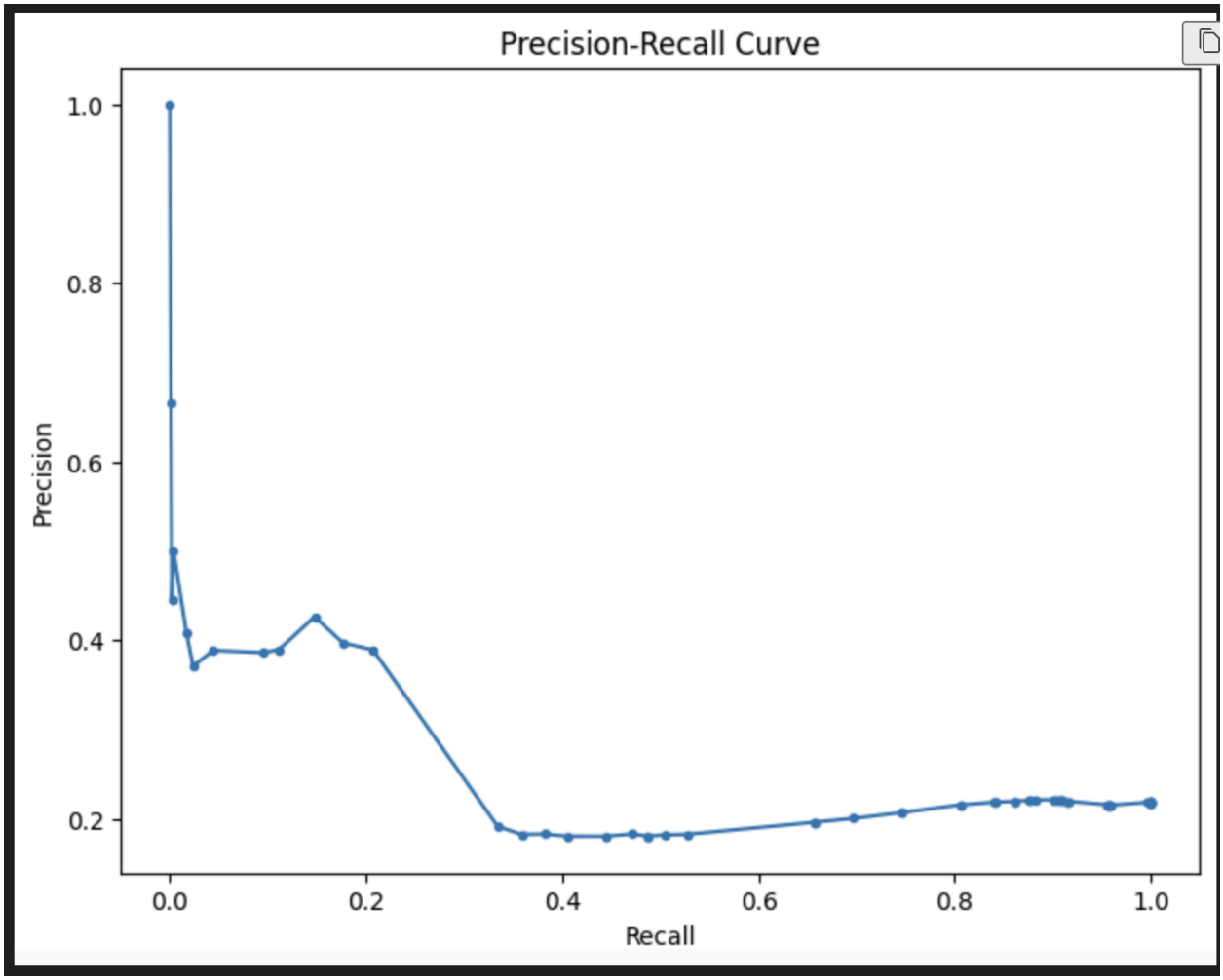

Figure 7. Recall Curve

Conclusion and Application of Results

In the end, the Random Forest model was the most effective at predicting defaults, with the highest accuracy and the most insightful feature importance rankings. The study's results highlighted that payment status is the strongest predictor of whether a customer will default on their credit card. Customers who consistently made late payments were much more likely to default, especially if they also had high bill amounts and low payments relative to their debt.

The findings of this study have significant implications for credit risk management. By focusing on key predictors like payment status, age, and credit limit, financial institutions can better manage their credit portfolios and reduce the incidence of defaults. The Random Forest model, with its strong performance, could be integrated into real-world systems to monitor customers and take preemptive action before defaults occur. For instance, customers identified as high-risk could be offered financial counseling or adjusted credit terms to help them avoid defaulting.

In conclusion, this study has provided valuable insights into the factors that lead to credit card defaults and demonstrated the effectiveness of machine learning models in predicting these defaults. The Random Forest model, in particular, stood out as a powerful tool for identifying high-risk customers based on their payment behavior and credit history. By using data-driven approaches to understand and predict credit card defaults, banks can minimize risk and offer customers the opportunity to avoid financial hardship. The insights gained from this study provide a solid foundation for improving credit risk management and ensuring financial stability for both lenders and borrowers.