# Vesta E-Commerce Fraud Detection

## Capstone 2 Final Report

By: Valentina Sanchez

# Introduction

In the realm of e-commerce, the fine line between security and convenience is navigated through the adept implementation of fraud prevention systems.
As you face the inconvenience of a declined card at the checkout, it's these systems that are the unsung heroes guarding against unauthorized transactions. Partnering with IEEE-CIS, Vesta Corporation is on the forefront of refining these systems, leveraging a vast dataset from real-world transactions to benchmark machine learning models.
This EDA report delves into the intricacies of transaction data, aiming to enhance the precision of fraud detection and, by extension, your shopping experience.

# Data-set Overview

The dataset is a hefty collection of 590,540 transactions each described by 434 features. The statistical summary of the 'TransactionAmt' feature reveals an average transaction amount of $134.92 with a standard deviation of $231.89, indicating a wide variation in transaction values. The following is the link to the dataset used.

****Side note:** The dataset thus hefty is composed mostly of anonymous features. Thankfully, someone who works at Vesta did provide some insight on some of the features which I will share below but can also be found with the following link:

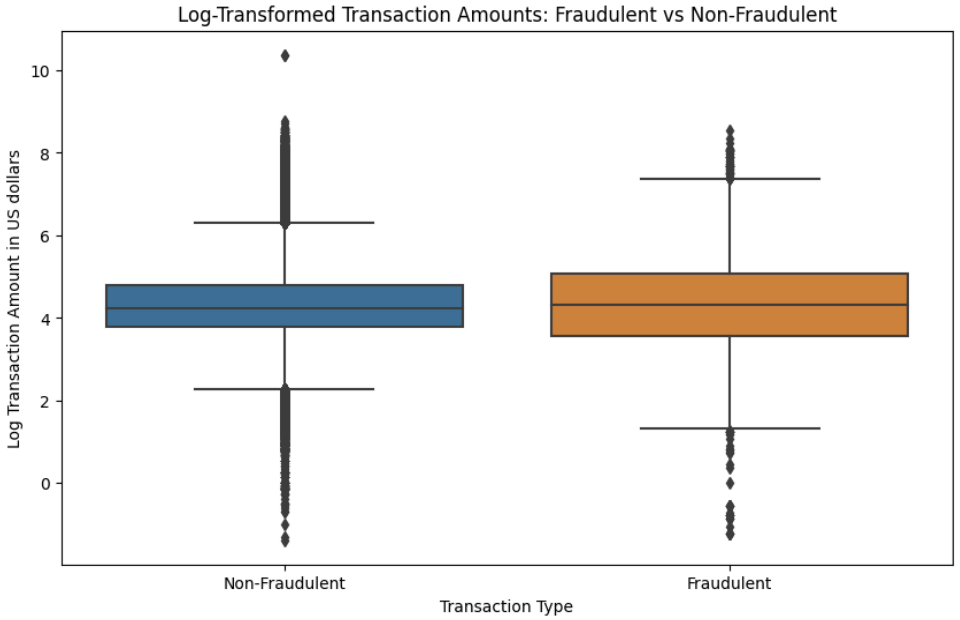
*Transaction Table:*

| ***Field*** | ***Description*** |
|---|---|
| *TransactionDT* | *Timedelta from a given reference datetime (not a timestamp)* |
| *TransactionAMT* | *Transaction payment amount in USD* |
| *ProductCD* | *Product code, representing the product for each transaction* |
| *card1 - card6* | *Payment card information, including card type, category, etc.* |
| *addr* | *Address* |
| *dist* | *Distance* |
| *P_ and (R__)* | *Email domains for purchaser and recipient* |
| *C1 - C14* | *Counting information, such as associated addresses, etc.* |
| *D1 - D15* | *Timedelta information, including days between transactions* |
| *M1 - M9* | *Match information, such as card names and addresses* |
| *Vxxx* | *Vesta engineered rich features, including ranking, counting,and other entity relations* |

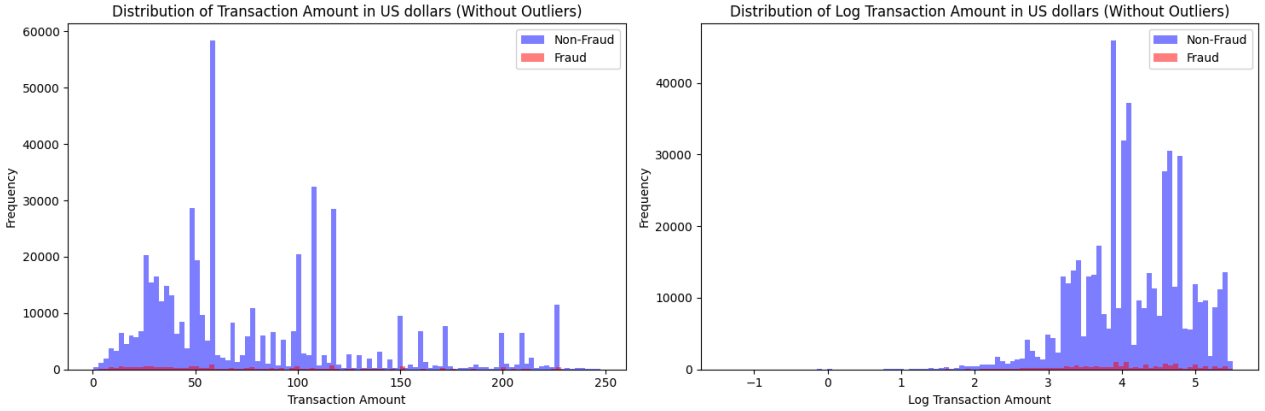
*Identity Table:*
Variables in this table are identity information – network connection information (IP, ISP, Proxy, etc) and digital signature (UA/browser/os/version, etc) associated with transactions. They're collected by Vesta's fraud protection system and digital security partners.
(The field names are masked and pairwise dictionary will not be provided for privacy protection and contract agreement)

# Exploratory Data Analysis

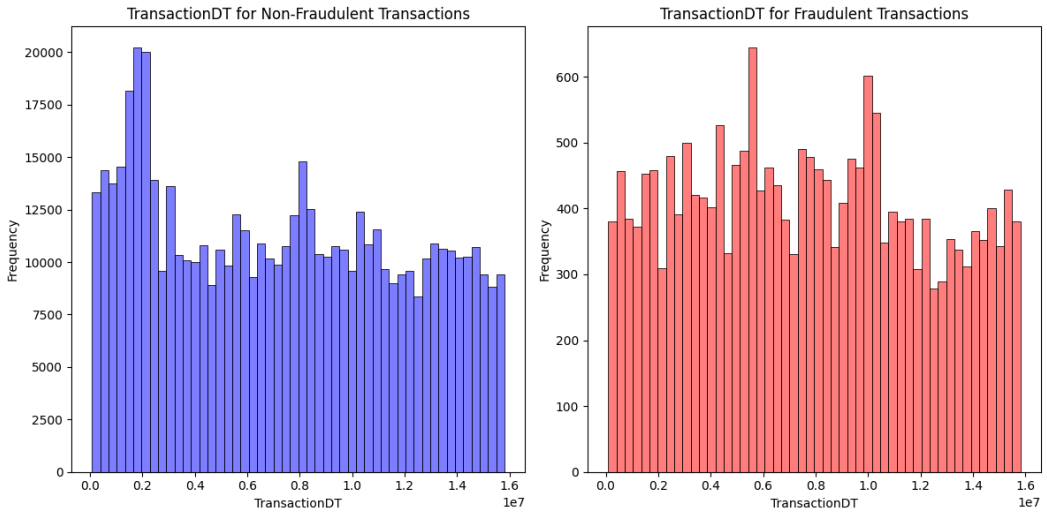
1. **Boxplot of Log-Transformed Transaction Amounts:** The log transformation of transaction amounts showcases a clear distinction between fraudulent and non-fraudulent transactions, with fraudulent ones typically involving higher amounts. It's important to note that without the log transformation, the difference in transaction amounts might be even more apparent. In addition to the higher amounts associated with fraudulent transactions, there also seems to be a concentration of lower amounts, which may become more evident when the data isn't log-transformed.



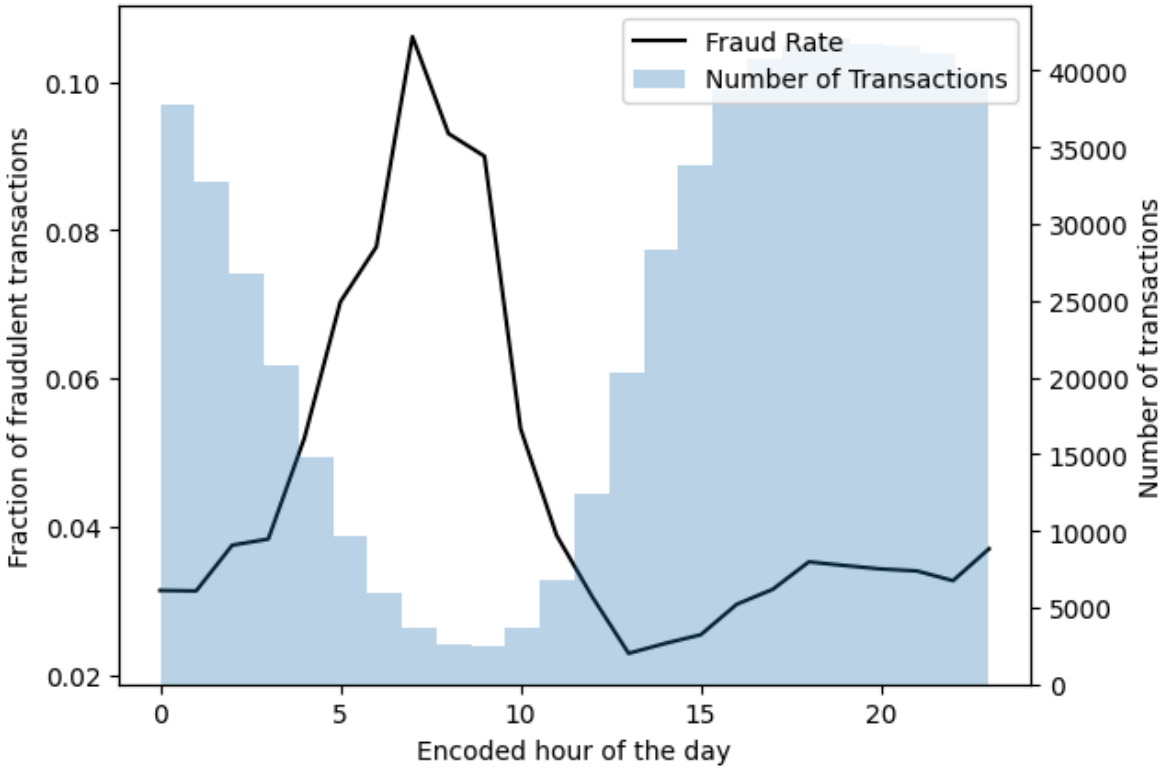
2. **Transaction Amount Distributions:** The pair of histograms titled 'Distribution of Transaction Amount in US dollars (Without Outliers)' and 'Distribution of Log Transaction Amount in US dollars (Without Outliers)' elucidate the distribution of transaction amounts for both non-fraudulent and fraudulent transactions. The first histogram indicates a skew towards lower transaction amounts, with spikes in frequency at certain intervals, which may represent common transaction thresholds. The second histogram, which presents the log-transformed transaction amounts, reveals a more normalized distribution. This transformation accentuates that fraudulent activities are not confined to higher transaction values but are rather distributed across a wide range of amounts. It is evident that fraudulent transactions can occur at any scale, underscoring the need for vigilance across all transaction levels.




3. **TransactionDT Distributions:** The conversion of 'TransactionDT' to hours has uncovered patterns in transaction frequency. For non-fraudulent transactions, the frequency appears relatively consistent, with slight variations indicating typical shopping hours. Conversely, fraudulent transactions display a more erratic pattern, suggesting that fraudsters may operate at times when detection is less likely, or they exploit specific time-based vulnerabilities in the transaction process.

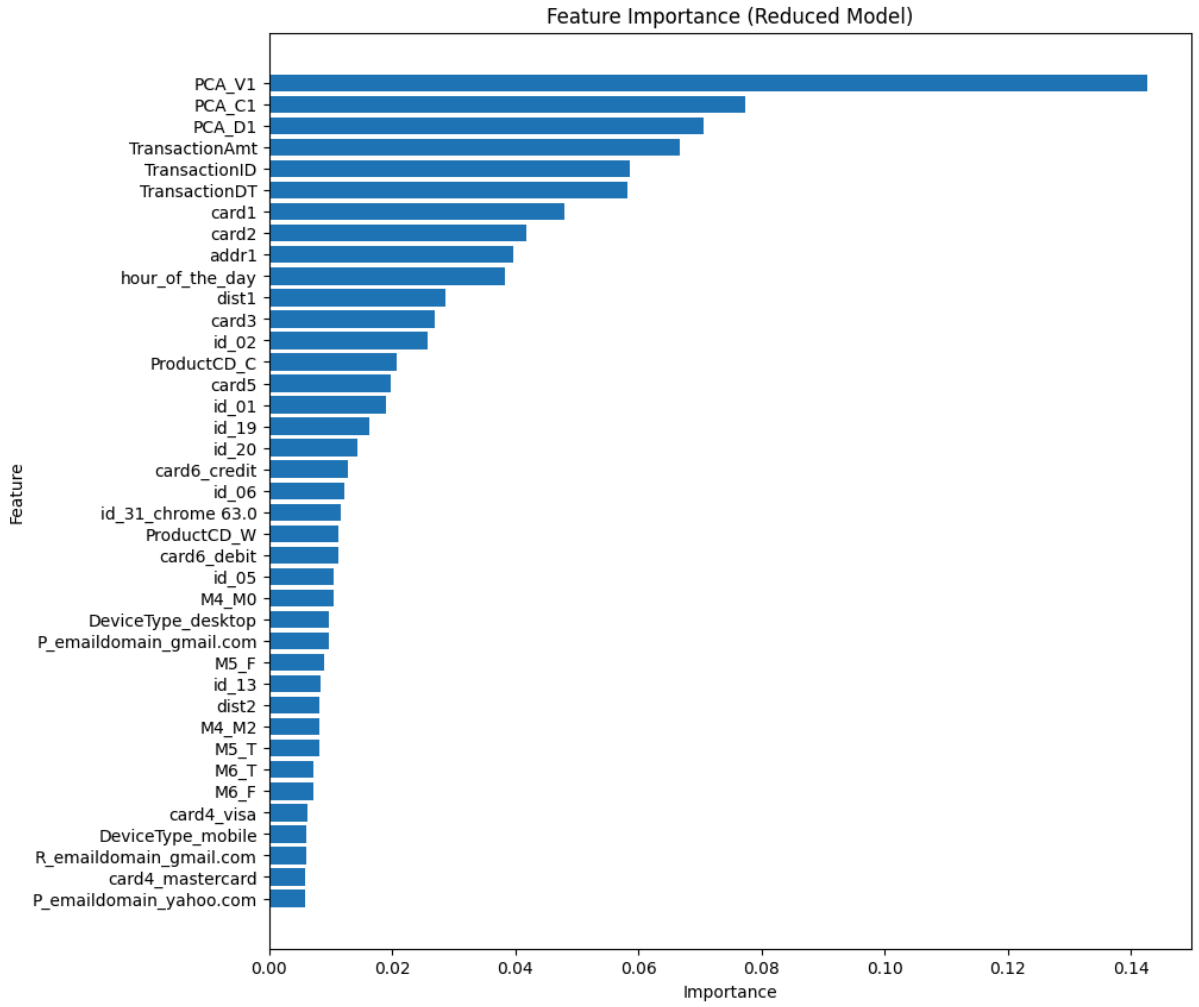


4. **Encoded Hour of the Day:** The plot of encoded hours against the fraction of fraudulent transactions uncovers possible time-based trends in fraud occurrences. The graph plotting the fraction of fraudulent transactions against the encoded hour reveals that fraud rates fluctuate throughout the day, peaking at certain hours. This trend can inform security measures, such as increasing monitoring during peak fraud hours, to prevent fraudulent activities more effectively.

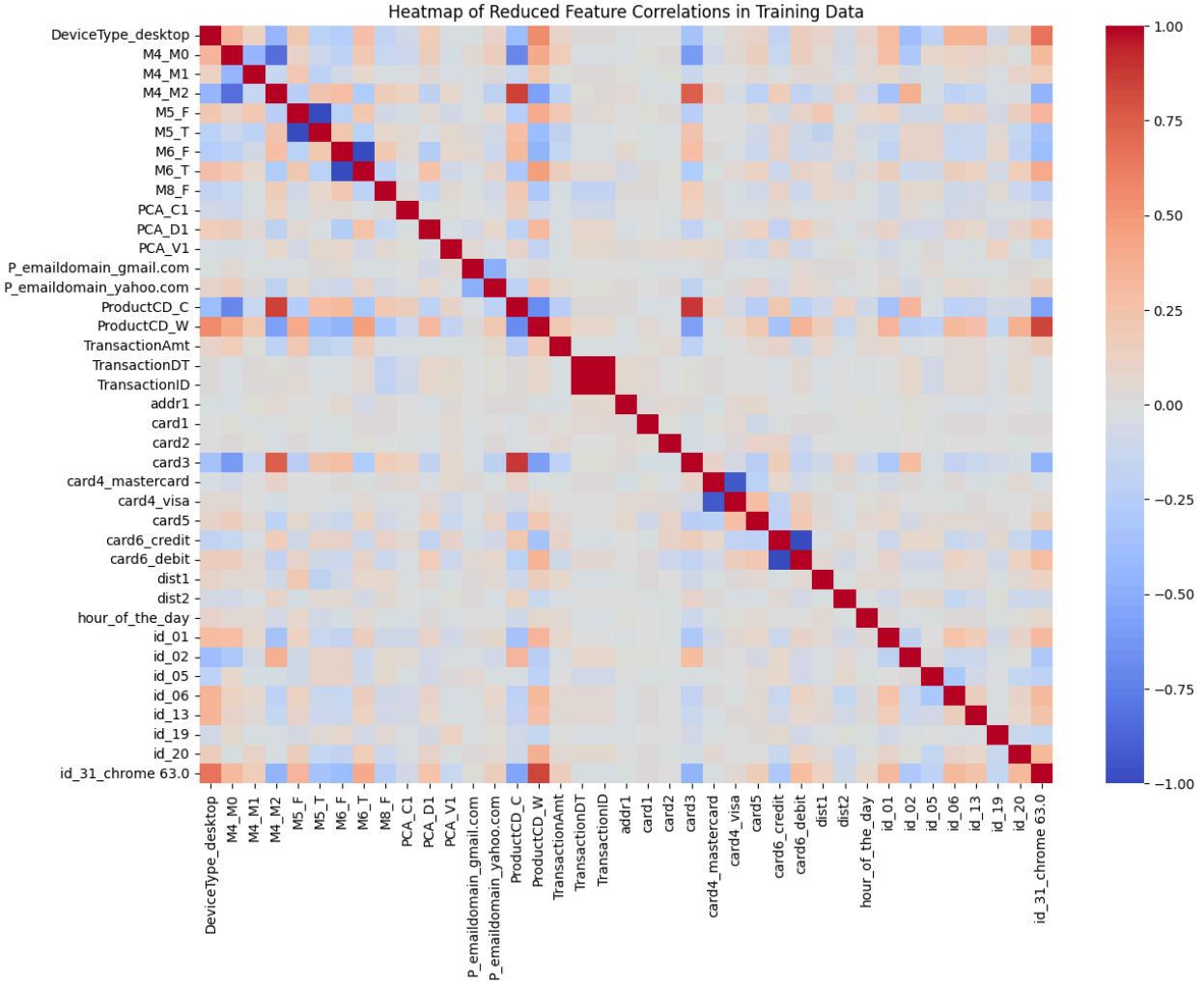


5. **Feature Importance:** As mentioned in a side note, a significant portion of the features within the dataset is anonymous. This means that they are encoded, and their actual representation and meaning are hidden. For instance, features like 'C1-C14' are used for counting purposes, such as determining how many addresses are associated with a payment card, but their true significance remains concealed.

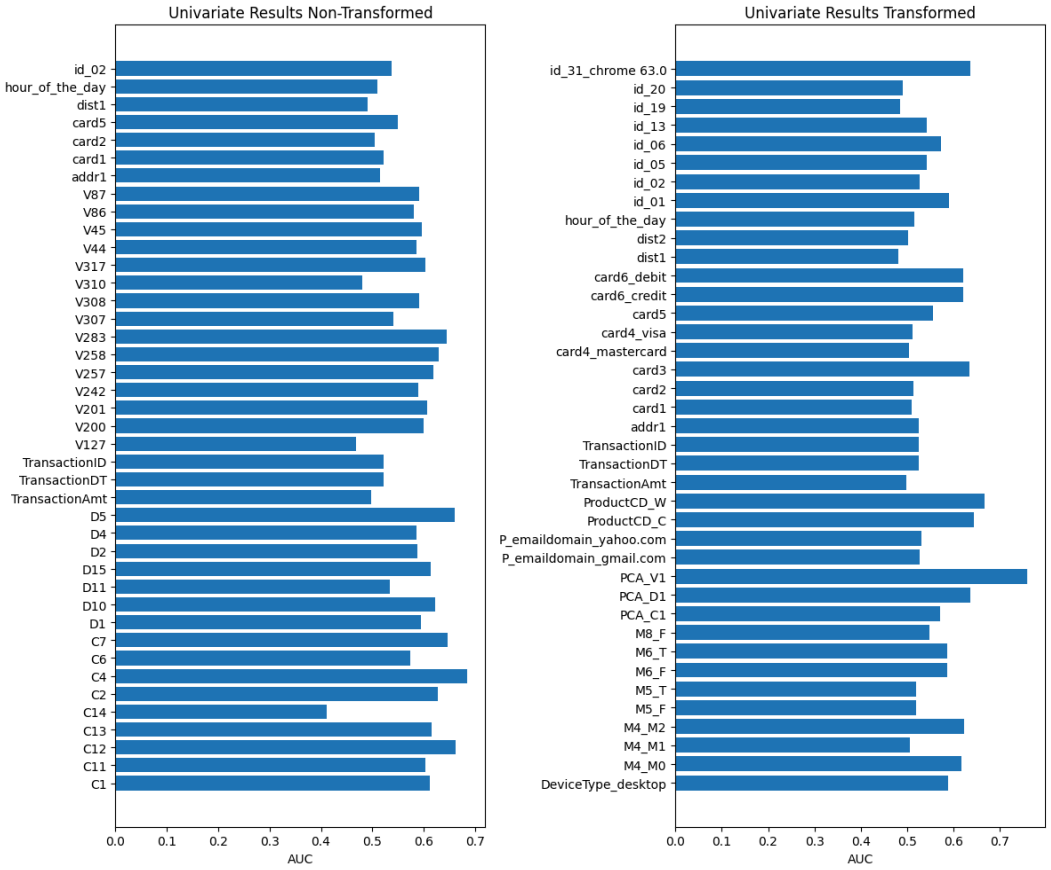
The bar graph 'Feature Importance (Reduced Model)' continues to highlight the pivotal role of principal components such as 'PCA_V1', 'PCA_D1', and 'PCA_C1' following the dimensionality reduction. These components signify the encapsulation of crucial information from the original 'V', 'D', and 'C' features, reducing multicollinearity while retaining their predictive power. The 'TransactionAmt' and 'TransactionDT' remain influential, along with other temporal features like 'hour_of_the_day', indicating that transaction timing and amount are key indicators of fraudulent activity. The persistence of these features at the top of the importance chart reaffirms their critical role in the predictive modeling of fraud detection.



6. **Heatmap of Feature Correlations:** The heatmap visualizes the correlations between features after the application of PCA to the 'V', 'D', and 'C' features. The color gradient reveals that the previously high collinearity among these features has been significantly reduced. The lighter colors off the diagonal suggest that PCA has successfully minimized redundant information, allowing for a clearer understanding of the unique contribution of each feature. The heatmap confirms that PCA was instrumental in extracting the essence of the 'V', 'D', and 'C' features, enhancing the model's ability to discern patterns indicative of fraud while avoiding issues associated with collinearity.



7. **Features by AUC:** The graph 'Univariate Results Non-Transformed' versus 'Univariate Results Transformed' presents the Area Under the Curve (AUC) of individual features before and after applying Principal Component Analysis (PCA) to the 'C' and 'V' features. Overall, the transformation through PCA seems to have shifted the relative importance of features, possibly reducing overfitting and enhancing the generalizability of the model by capturing underlying patterns within the 'C' and 'V' features. The consistency of 'hour_of_the_day' in both graphs underscores its robustness as a predictor regardless of feature transformation.




**Insights on Exploratory Data Analysis**

Our exploratory data analysis has honed in on pivotal transaction characteristics. PCA transformation has streamlined collinear features, bolstering their predictive value. Time of transaction remains a consistent predictor, while analysis of transaction amounts shows fraud's prevalence across various sums. PCA features like 'PCA_V1', 'PCA_D1', and 'PCA_C1' have proven vital in distinguishing fraud, potentially cutting down on incorrect flags. Leveraging these insights, Vesta can sharpen its fraud detection, balancing robust security with user convenience.

# Model Selection - Choosing a Classifier

## Objective

The primary objective of our modeling efforts is to strike an optimal balance between recall and precision, ensuring a robust fraud detection model. Our goal is to fine-tune the model's parameters such that it minimizes false positives—thereby reducing the inconvenience of legitimate transactions being incorrectly flagged—while simultaneously achieving a high recall to capture the maximum number of fraudulent transactions. This balanced approach is crucial for maintaining user trust by providing a seamless transaction experience and safeguarding against fraud.

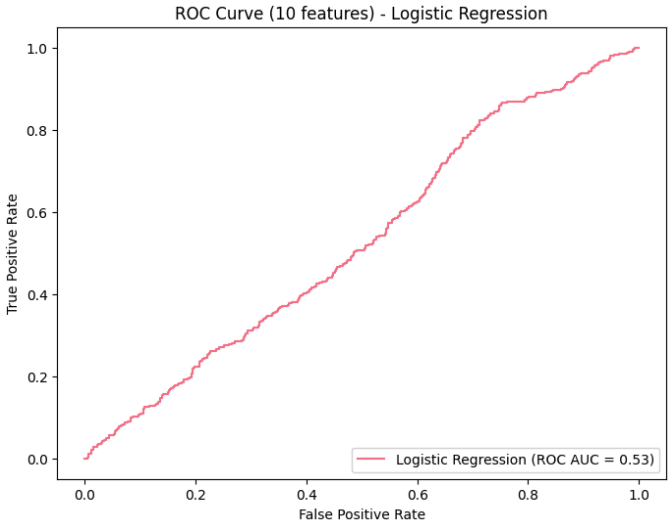
## Evaluating the Hyperparameters

A comprehensive grid search was conducted to determine the optimal hyperparameters, where we tested various feature subsets. These subsets, derived from the top features identified through Random Forest Feature Importance analysis, ranged from the top 6, 10, 20, to the complete set of top 39 features. Our approach was to assess the models' Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) scores alongside finding the hyperparameters that maximized each model's performance.

| ***Classifier*** | ***ROC-AUC Score*** | ***Best Hyperparameter Values*** |
|---|---|---|
| *Random Forest* | *0.89* | *{'C': 0.001, 'class_weight': None}* |
| *Logistic Regression* | *0.51* | *{'max_depth': 20, 'min_samples_split': 2, 'n_estimators': 400}* |
| *XGBoost* | *0.88* | *{'learning_rate': 0.1, 'max_depth': 6, 'n_estimators': 300}* |

## Logistic Regression

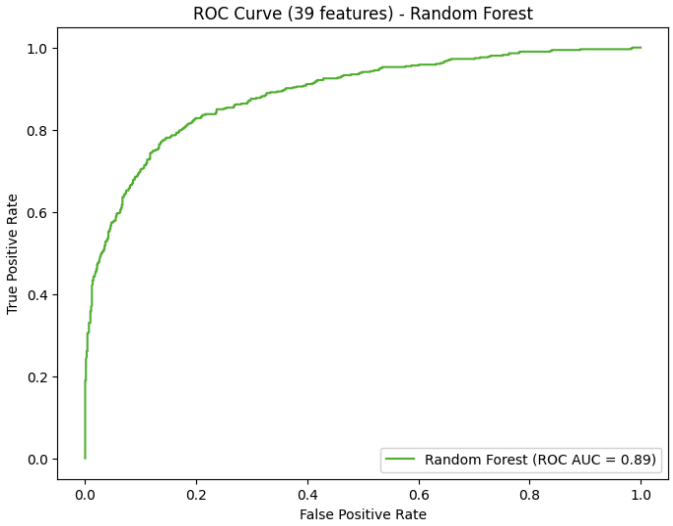
The Logistic Regression model exhibited the highest ROC AUC score of 0.53 when using both the top 6 and 10 features; however, the overall lower scores suggest its performance is largely attributable to random chance rather than predictive accuracy.




```
Model: Logistic Regression

Accuracy: 0.6528

Precision: 0.75

Recall: 0.17

F1 Score: 0.28

Confusion Matrix: [[731 27]
[407 85]]
```

## Random Forest

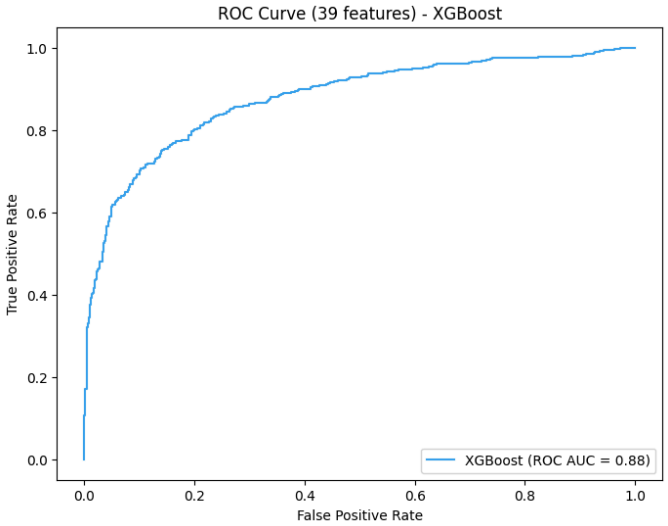
The Random Forest model demonstrated a higher ROC AUC score of 0.89 when utilizing all 39 features, indicative of its robust capability to differentiate between fraudulent and legitimate transactions.




```
Model: Random Forest

Accuracy: 0.81

Precision: 0.79

Recall: 0.71

F1 Score: 0.75

Confusion Matrix: [[666 92]
[143 329]]
```

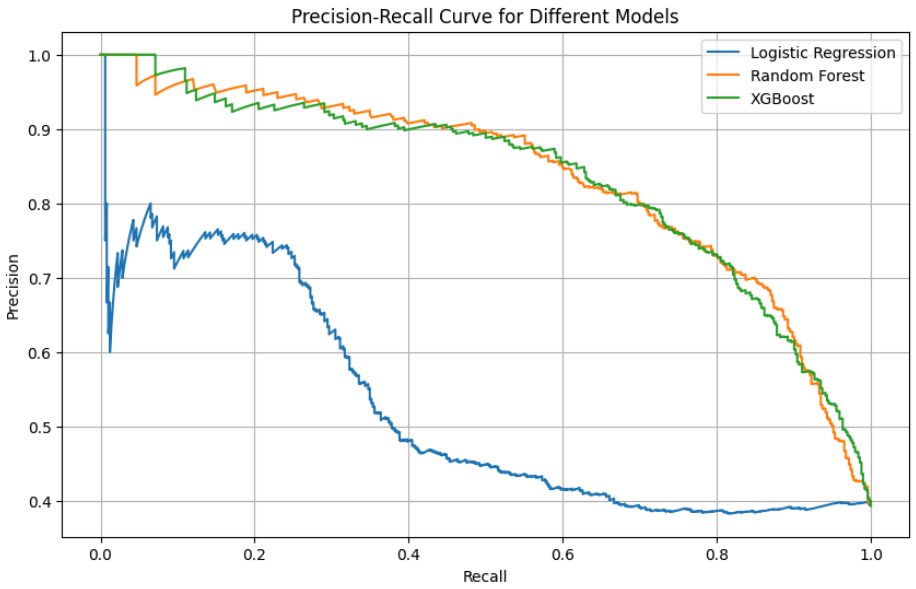
## XGBoost

XGBoost achieved a ROC AUC score of 0.88, similar to Random Forest, indicating its strong performance or a very good ability to distinguish between fraudulent and non-fraudulent transactions.

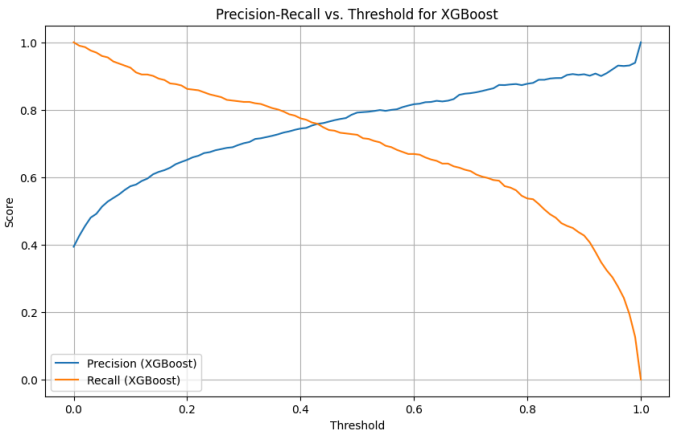


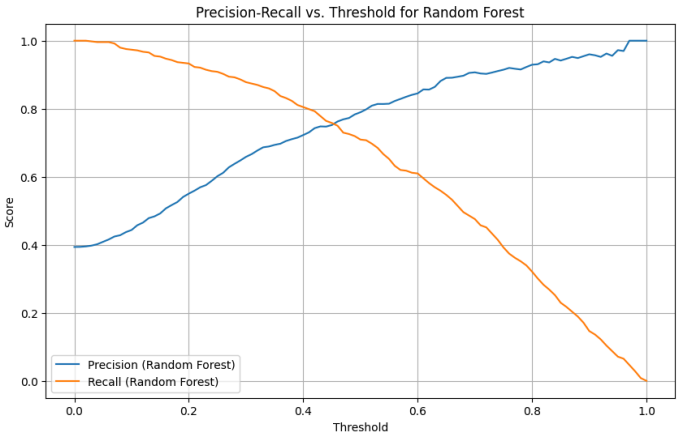

```
Model: XGBoost

Accuracy: 0.82

Precision: 0.79

Recall: 0.73

F1 Score: 0.76

Confusion Matrix: [[664 94]
[135 357]]
```

The Precision-Recall curve below compares the performance of three different classification models: Logistic Regression, Random Forest, and XGBoost. Logistic Regression is shown to have significantly lower precision across all levels of recall compared to the other two models, indicating it may not be as effective for this particular task. Both Random Forest and XGBoost, on the other hand, display a much closer performance with high precision at higher levels of recall. This suggests that they are more capable of correctly identifying positive cases (high precision) while also capturing a large proportion of the actual positive cases (high recall). Overall, Random Forest and XGBoost are both strong contenders for the classification task, and the choice between them may depend on the specific recall-precision trade-off that is desired for the application.



## Precision-Recall vs. Threshold Analysis




### Threshold Selection

Analyzing the Precision-Recall vs. Threshold graphs for XGBoost and Random Forest models reveals a trade-off between precision and recall. Initially, at low thresholds, XGBoost exhibits higher recall but lower precision, while Random Forest shows a more gradual decrease in recall. As thresholds rise, both models improve precision but decrease recall. However, XGBoost experiences a steeper drop in recall at higher thresholds, indicating greater sensitivity to threshold changes. Finding the balance between precision and recall is what will help us determine which model to use in the end, therefore we will then experiment with the different threshold to observe which can be more balanced.

**Threshold Experimentation:**

**XGBoost**

| *Threshold* | *Precision (Fraud)* | *Recall (Fraud)* |
|---|---|---|
| *0.3* | *0.70* | *0.82* |
| *0.43*** | *0.76* | *0.76* |
| *0.5* | *0.79* | *0.73* |
| *0.7* | *0.85* | *0.62* |

**Random Forest**

| *Threshold* | *Precision (Fraud)* | *Recall (Fraud)* |
|---|---|---|
| *0.3* | *0.66* | *0.88* |
| *0.45*** | *0.75* | *0.76* |
| *0.5* | *0.79* | *0.71* |
| *0.7* | *0.91* | *0.48* |

***Intersection of precision and recall

## Final Model Choice

XGBoost and Random Forest show similar precision and recall at intersection points, but XGBoost's slightly higher precision at a lower threshold (0.3) suggests better performance in high-recall scenarios. XGBoost also exhibits a more gradual decline in recall with increasing thresholds, indicating greater stability. Given our pursuit of a model that demonstrates stability and maintains equilibrium between recall and precision across various threshold settings, XGBoost stands out as a viable option. Particularly at a threshold of 0.43, XGBoost achieves a harmonious balance between recall and precision, suggesting it as a robust model choice for our fraud detection objectives.

**Performance Metrics**

Upon applying the threshold where the intersection between recall and precision is, the XGBoost model yielded the following classification metrics:

```
Classification Report of XGBoost:
              precision    recall  f1-score   support

   Not Fraud       0.84      0.84      0.84       758
       Fraud       0.76      0.76      0.76       492

    accuracy                           0.81      1250
   macro avg       0.80      0.80      0.80      1250
weighted avg       0.81      0.81      0.81      1250

Confusion Matrix (in percentages):
 [[84.30079156 15.69920844]
 [24.18699187 75.81300813]]
```

**Conclusion**

In conclusion, the XGBoost model has proven to be highly effective in the classification of fraudulent transactions, exhibiting a strong balance between precision and recall, as evidenced by the classification report. The model has an F1-score of 0.76 for detecting fraud, highlighting its reliability in identifying fraudulent activities. The nearly equal precision and recall scores for both classes, along with an overall accuracy of 81%, highlight the model's ability to perform consistently across different types of transactions.
However, the confusion matrix indicates that there is potential for further refinement, reducing the 24.18% of false negatives and the 15.69% of false positives.
Overall, the XGBoost model's solid statistical underpinnings suggest that it is a robust and dependable choice for fraud detection systems.

# Further work

This project represents an initial stride into the realm of banking systems, a field encompassed by numerous factors. Despite the considerable work that remains, I am confident that the following steps can be taken to advance this endeavor further.

Explore the possibility of enhancing the predictive accuracy and resilience of our fraud detection system by integrating the two top-performing models, Random Forest and XGBoost, through methods like ensembling or stacking. Utilizing the collective strengths of different models may lead to improved performance in detecting fraudulent activities. Additionally, investigate the practicality and necessary infrastructure to deploy this integrated model in a live, real-time fraud detection environment. This entails evaluating the model's performance in handling live transaction data, with a focus on its adaptability to the fast pace and high volume of real-time operations.

By focusing on these aspects, we aim to progressively refine the fraud detection system's precision and dependability. This will ensure the system's effectiveness in countering emerging fraud strategies and maintaining its efficiency, all while reducing false positives and minimizing disruptions for genuine users.