Exploratory Data Analysis Report on Vesta's E-commerce Transactions

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:
https://www.kaggle.com/competitions/ieee-fraud-detection/discussion/101203

**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:

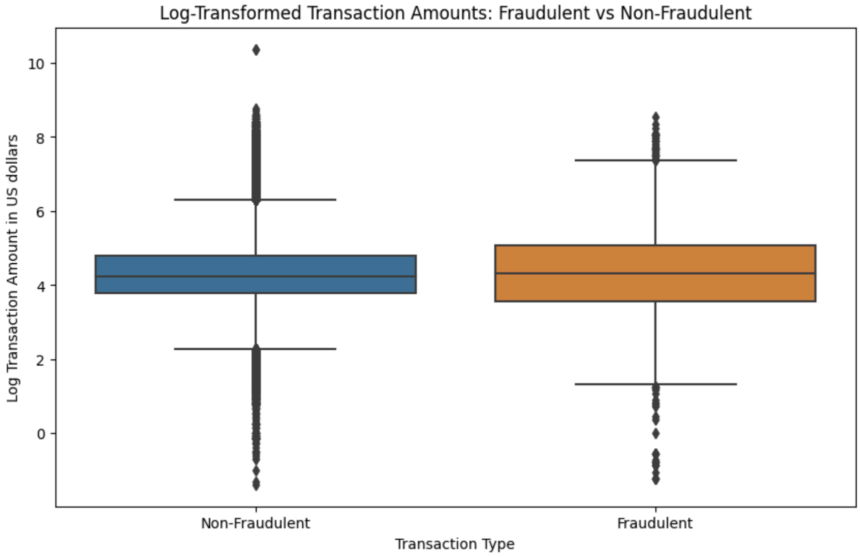
Field	***Description***
TransactionDT	*Timedelta from a given reference datetime (not a timestamp)*
TransactionAMT	*Transaction payment amount in USD*
ProductCD	*Product code, representing the product for each transaction*
card1 - card6	*Payment card information, including card type, category, etc.*
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:

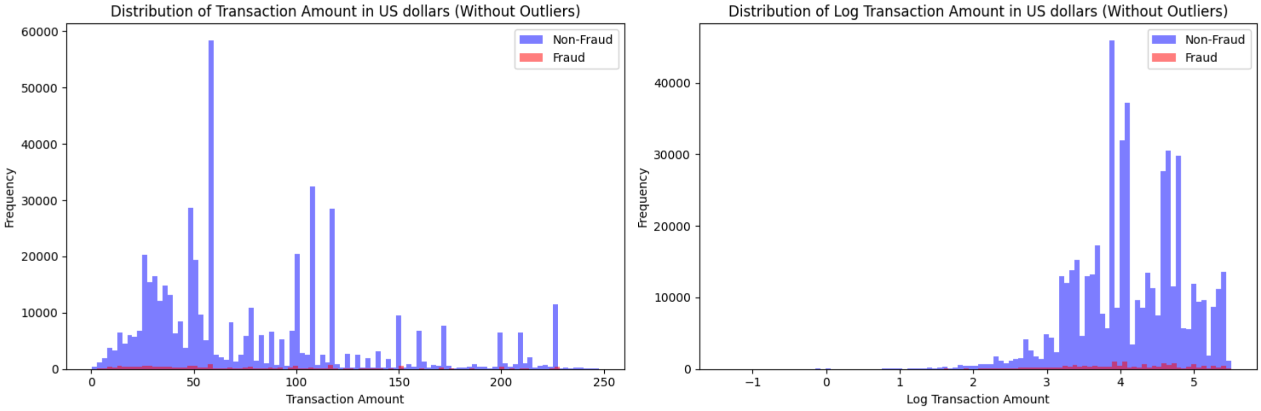
Variables in this table are identity information – network connection information (IP, ISP, Proxy, etc) and digital signature (UA/browser/os/version, etc) associated with transactions. They're collected by Vesta's fraud protection system and digital security partners.
(The field names are masked and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

Visual Analysis

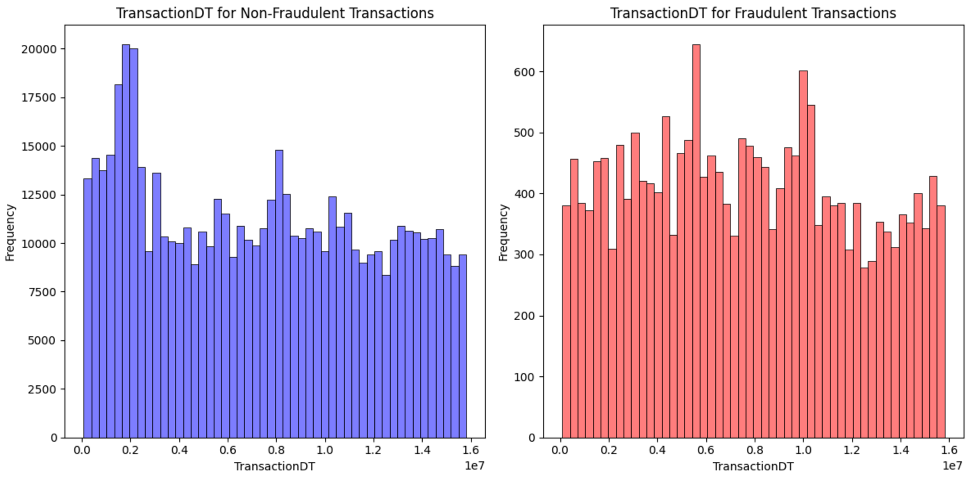
1. **Boxplot of Log-Transformed Transaction Amounts:** The log transformation of transaction amounts showcases a clear distinction between fraudulent and non-fraudulent transactions, with fraudulent ones typically involving higher amounts. It's important to note that without the log transformation, the 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.



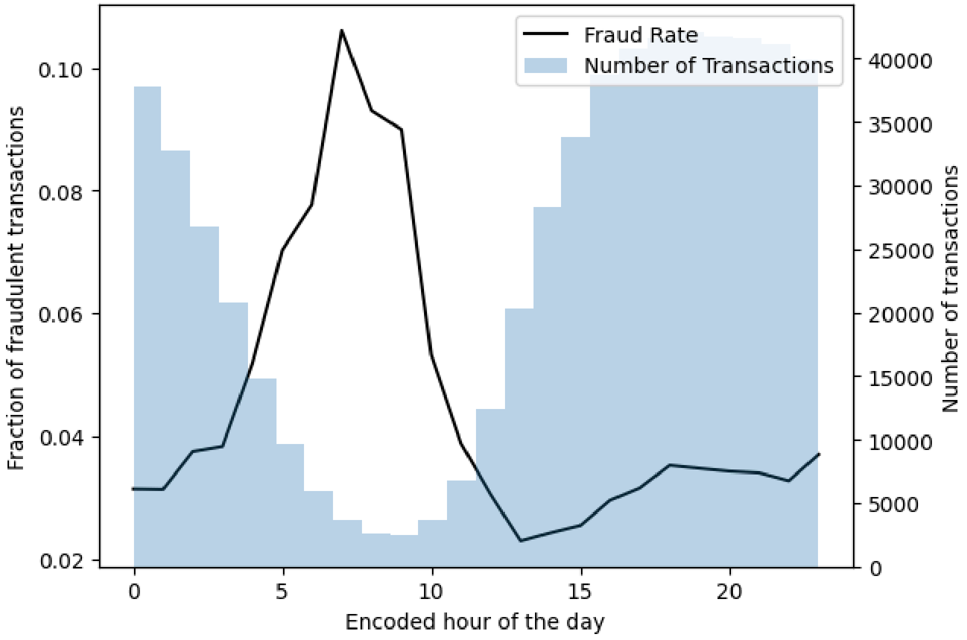
2. **Transaction Amount Distributions:** The pair of histograms titled 'Distribution of Transaction Amount in US dollars (Without Outliers)' and 'Distribution of Log Transaction Amount in US dollars (Without Outliers)' elucidate the distribution of transaction amounts for both non-fraudulent and fraudulent 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.

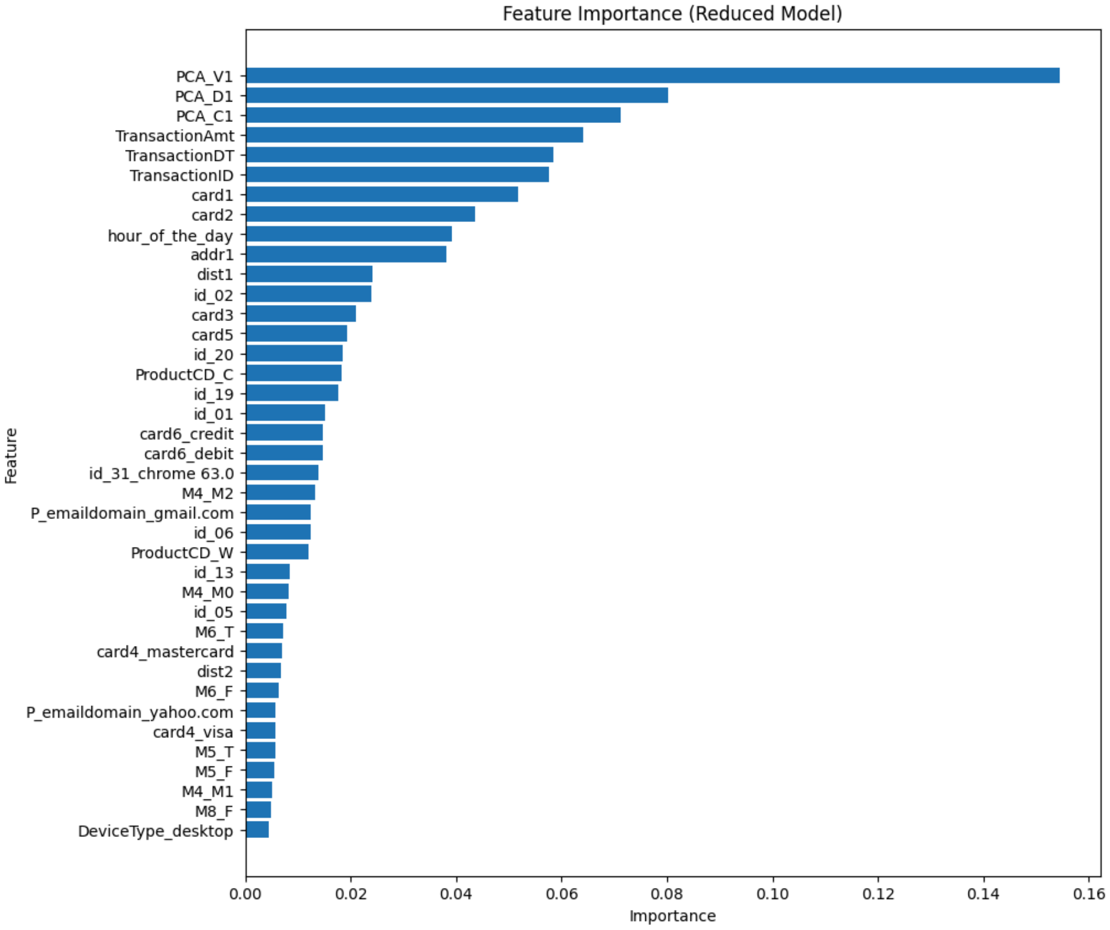


4. **Encoded Hour of the Day:** The plot of encoded hours against the fraction of fraudulent transactions uncovers possible time-based trends in fraud occurrences. The graph plotting the fraction of fraudulent transactions against the encoded hour reveals that fraud rates fluctuate throughout the day,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

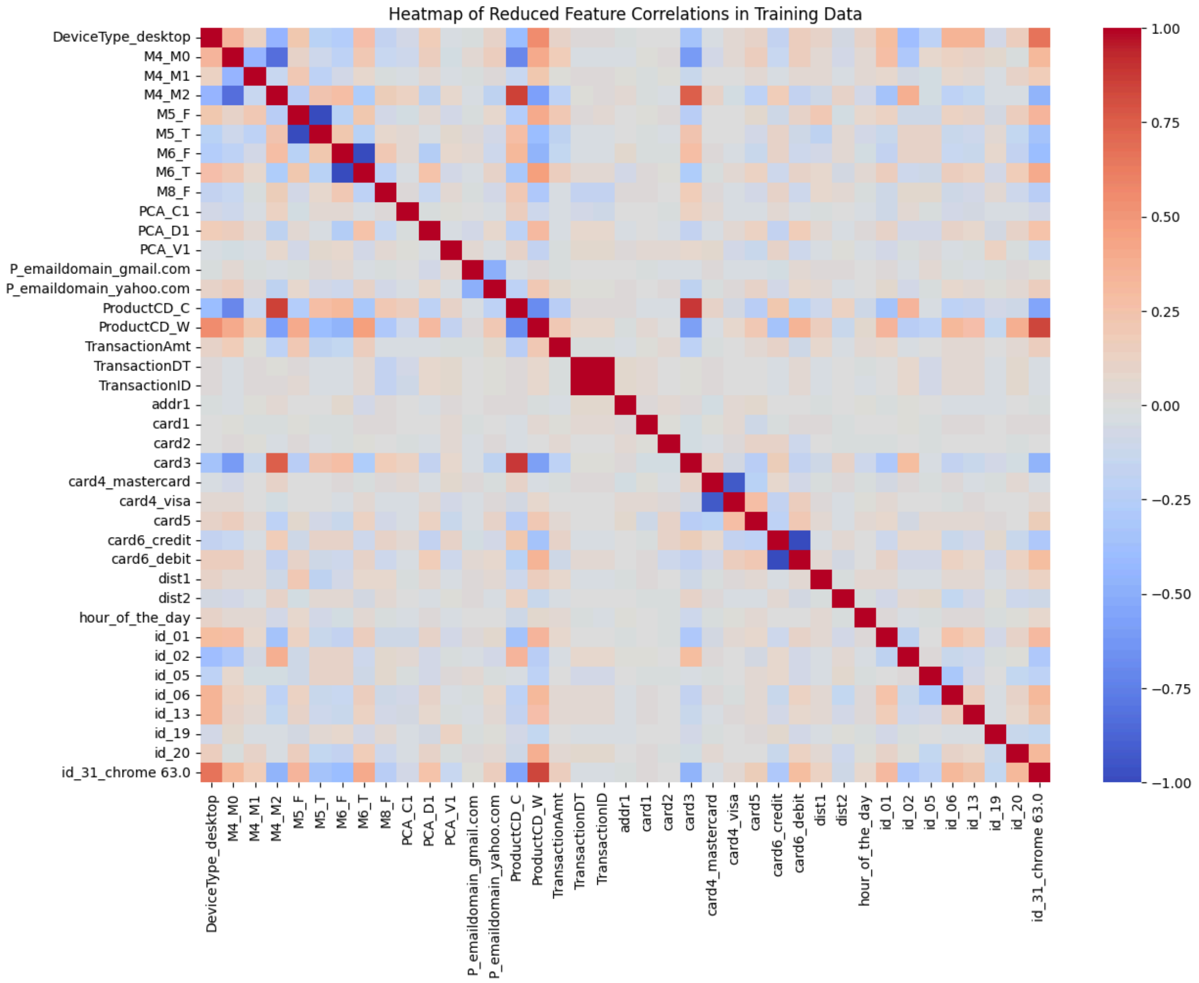


5. **Feature Importance:** As mentioned in a side note, a significant portion of the features within the dataset is anonymous. This means that they are encoded, and their actual representation and meaning are hidden. For instance, features like 'C1-C14' are used for counting purposes, such as determining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

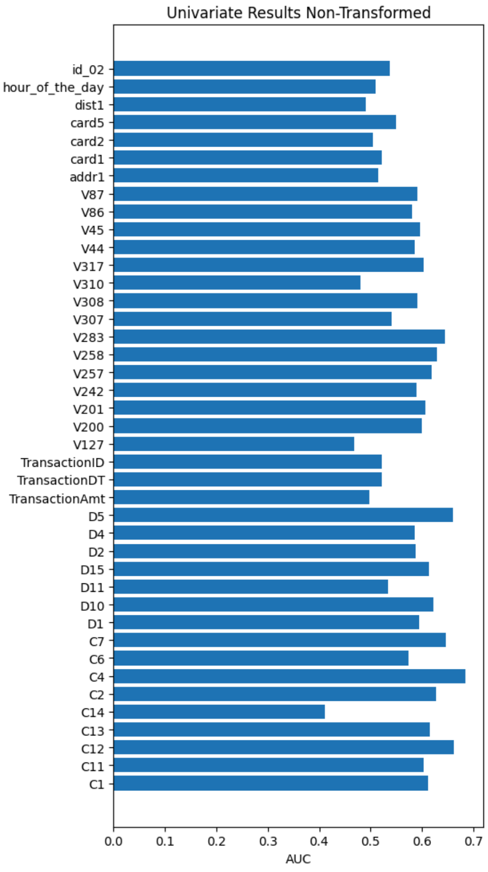


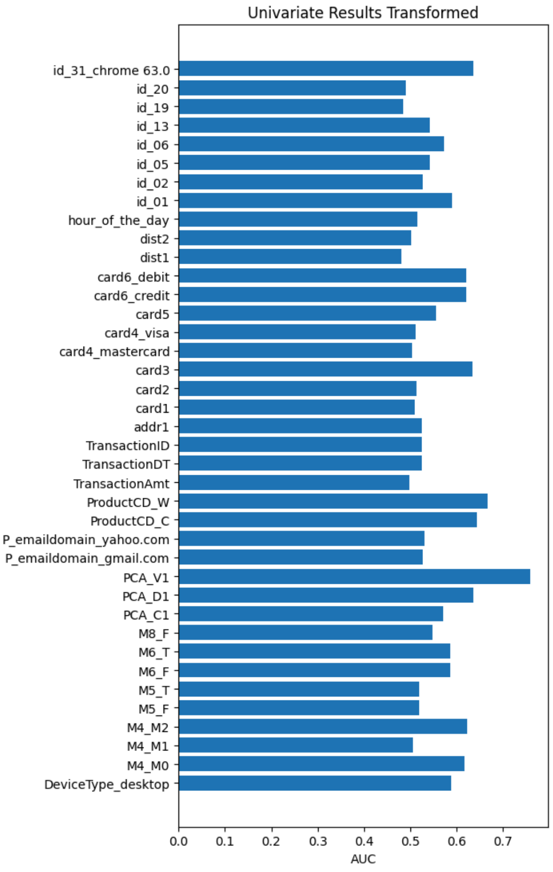
6. **Heatmap of Feature Correlations:** The heatmap visualizes the correlations between features after the application of PCA to the 'V', 'D', and 'C' features. The color gradient reveals that the previously high collinearity among these features has been significantly reduced. The lighter colors off 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 and Conclusions

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.