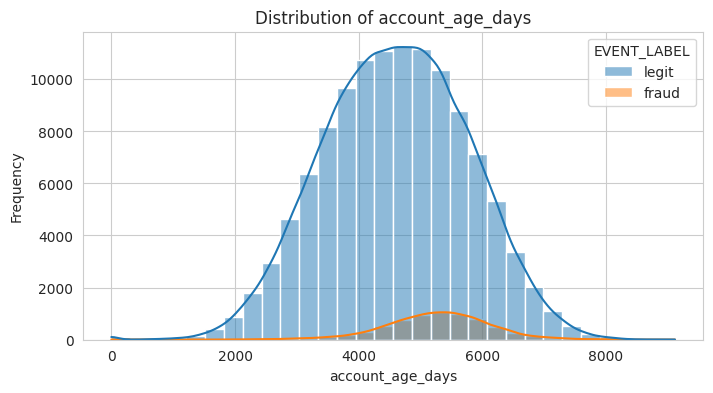
Finding Fraud Faster

# Executive Summary

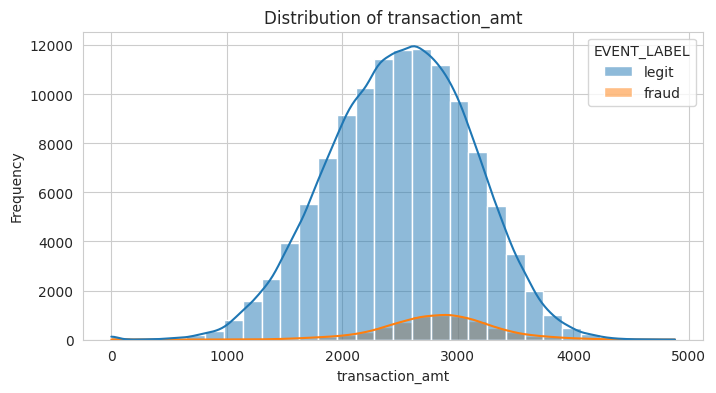
The dataset consists of 125,000 entries with 27 columns, including both numerical and categorical features. The target variable is EVENT\_LABEL, indicating whether an event is "legit" or not. There are missing values in several columns, as indicated by the non-null count being less than the total number of entries for those columns.



A diagram of a distribution of a number of transactions with Ryugyong Hotel in the background

Description automatically generated with medium confidence

The histogram for transaction\_amt (which is the first image you provided) shows the frequency distribution of transaction amounts, separated by event label (legit and fraud). The KDE overlay suggests that most legitimate transactions are concentrated around a median amount, while fraudulent transactions are relatively fewer in number and have a lower and flatter distribution, indicating they are more uniformly spread across different transaction amounts.



The second histogram for transaction\_adj\_amt (the additional image you provided) shows a similar pattern but on a different scale, indicating that transaction\_adj\_amt is likely a normalized or adjusted version of transaction\_amt. The distribution for legitimate transactions again shows a peak with a normal distribution, while fraud transactions are much less frequent and spread out across the range of adjusted transaction amounts.

A graph with blue and orange lines

Description automatically generated

Account age in days shows a wide distribution for both legit and fraud events, with a higher median for legit events.

There are outliers for both legit and fraud events, with legit events showing outliers on both the lower and higher ends.

Fraud events seem to have a tighter IQR but also exhibit outliers, particularly on the higher end, indicating that there are accounts tagged as fraud that have existed for a very long time.A graph of a box plot

Description automatically generated with medium confidence

A diagram of a box plot

Description automatically generated

Box Plot of transaction\_adj\_amt:

The median adjusted transaction amount for both legit and fraud events is relatively low compared to the overall range, with legit events having a slightly higher median.

There are outliers on both ends for legit events, with more extreme values on the higher end.

Fraud events also have outliers, but they are less extreme compared to legit events.

The interquartile range (IQR) for fraud events is narrower than for legit ones, indicating less variability in adjusted transaction amounts for fraud.

A diagram of a box

Description automatically generated

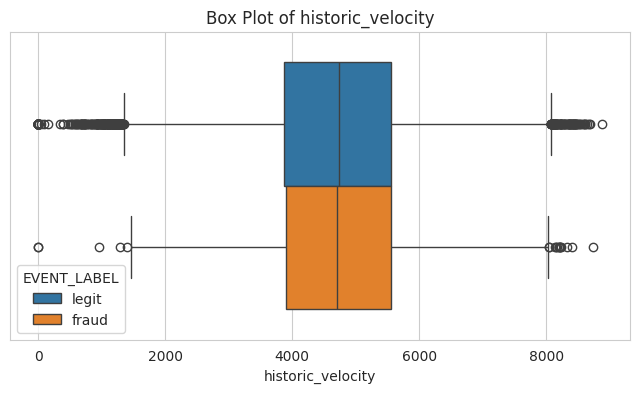
Box Plot of transaction\_amt:

The transaction amount has a wider range for legit events, with a higher median transaction amount compared to fraud events.

There are many outliers for legit events, especially on the higher end, indicating that there are quite a few legit transactions with significantly higher amounts than typical.

Fraud events show outliers primarily on the lower end.

The distribution of transaction amounts for fraud events is skewed towards the lower end, as seen by the position of the median closer to the first quartile.



**Model Development**

1. **Model Training**: Develop models using Logistic Regression, Random Forest, and GBM/XGBoost on the training data.
2. **Parameter Tuning**: Optimize model parameters to enhance performance.
3. **Feature Selection**: Identify and retain the most informative features for the models.

**Logistic Regression**

**A screenshot of a computer screen

Description automatically generated**

**Random Forest**

**A screenshot of a computer code

Description automatically generated**

**Gradient Boosting**

**A screen shot of a black screen

Description automatically generated**

A graph of a curve

Description automatically generated

Here's what the ROC curve tells us about each model:

Logistic Regression: The logistic regression model shows good performance, but it is slightly outperformed by the other two models.

Random Forest: The random forest model's ROC curve is very close to the top-left corner, indicating excellent performance. which means it has a higher true positive rate for most thresholds compared to logistic regression.

Gradient Boosting: The gradient boosting model's ROC curve is the closest to the top-left corner, suggesting it has the best performance among the three models. This indicates a higher true positive rate and a lower false positive rate for most thresholds.

A table with numbers and symbols

Description automatically generated

Operational Strategy at 5% FPR

IF predicted\_proba >= 0.280229 THEN fraud.

Expected True Positive Rate (TPR): This rule will catch 74.0291% of all frauds.

FPR decreases the precision increases, meaning that a higher proportion of the transactions flagged as fraudulent will actually be fraudulent.

**Understanding 5% False Positive Rate**:

A 5% FPR strikes a balance between catching fraudulent activity and not overwhelming the system or customers with false alarms. It's important to note that this rate also corresponds to a 74.0291% TPR, meaning that approximately 26% of fraudulent transactions might go undetected. This trade-off must be carefully considered against the cost and impact of fraud versus the cost and impact of false positives.

**Recommendations:**

While the current threshold achieves the desired FPR, it's important to continuously analyze the impact and adjust the threshold to optimize the balance between catching fraud and minimizing inconvenience to customers. Or develop a comprehensive customer education campaign about the measures in place to protect them from fraud. Explain the possibility of false alarms, their role in verification, and how these measures keep their accounts safe. Lastly, regularly review the model's performance metrics, such as precision, recall, and FPR, to ensure that it continues to meet the business's objectives and adjust strategies as needed.