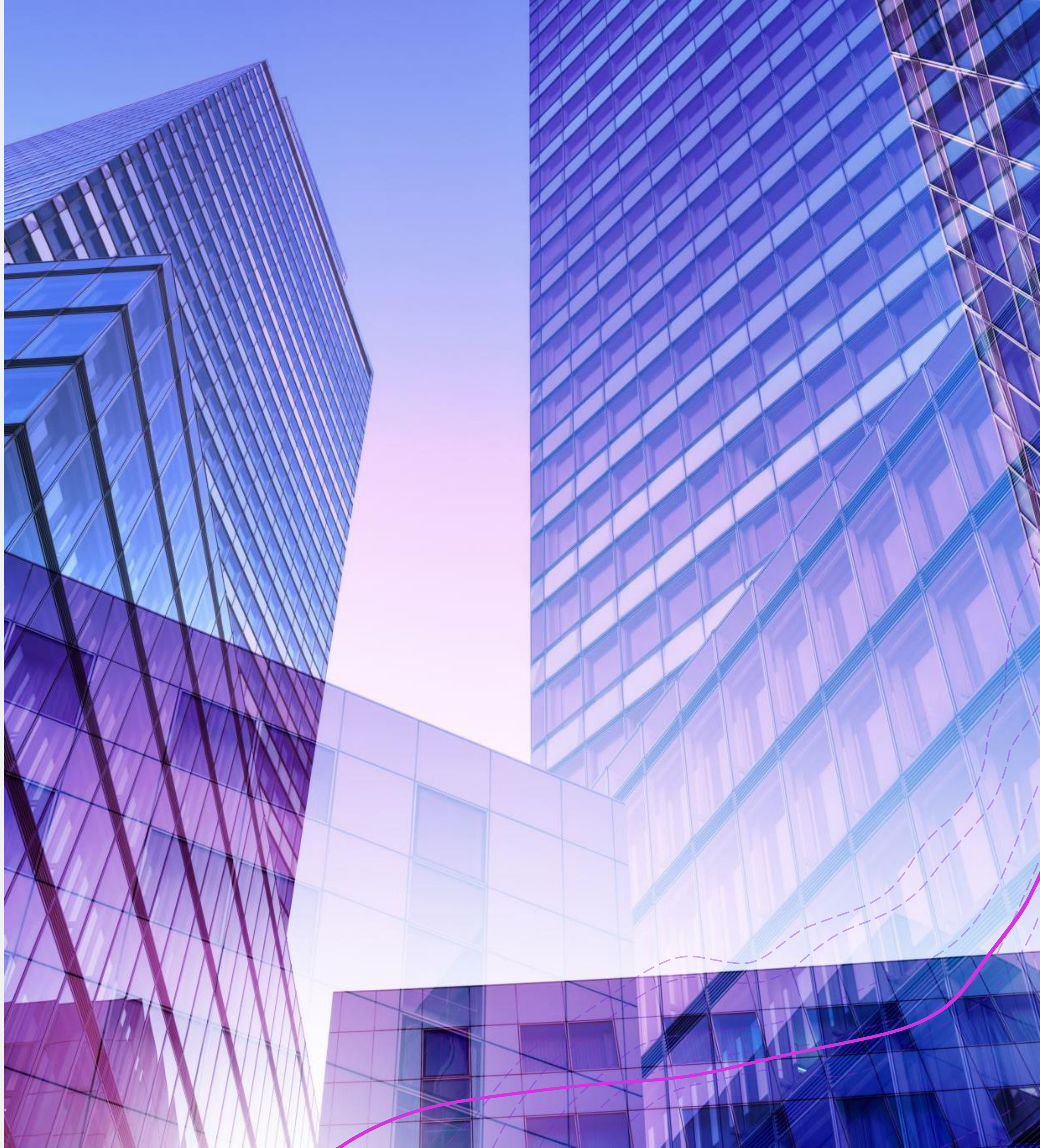
FICO Analytic Challenge

By Ameen Rufai, Khalil Greene & Saniya Isaac

Table of Contents

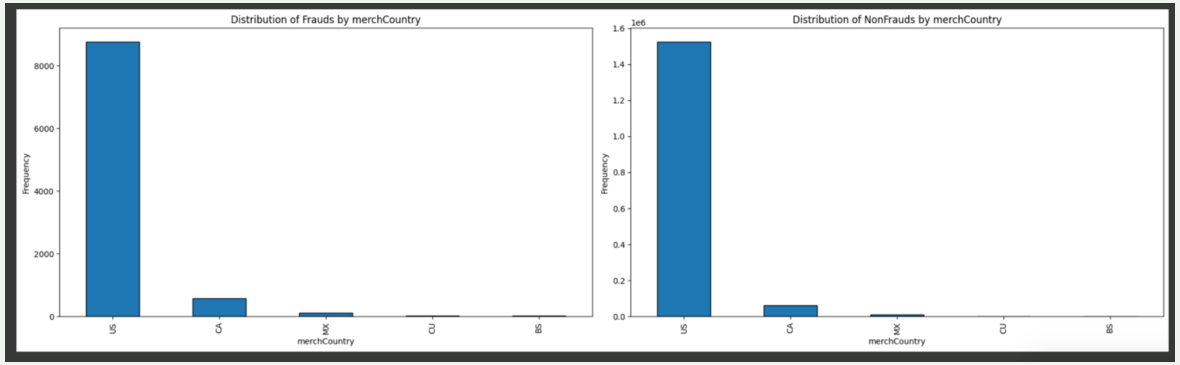


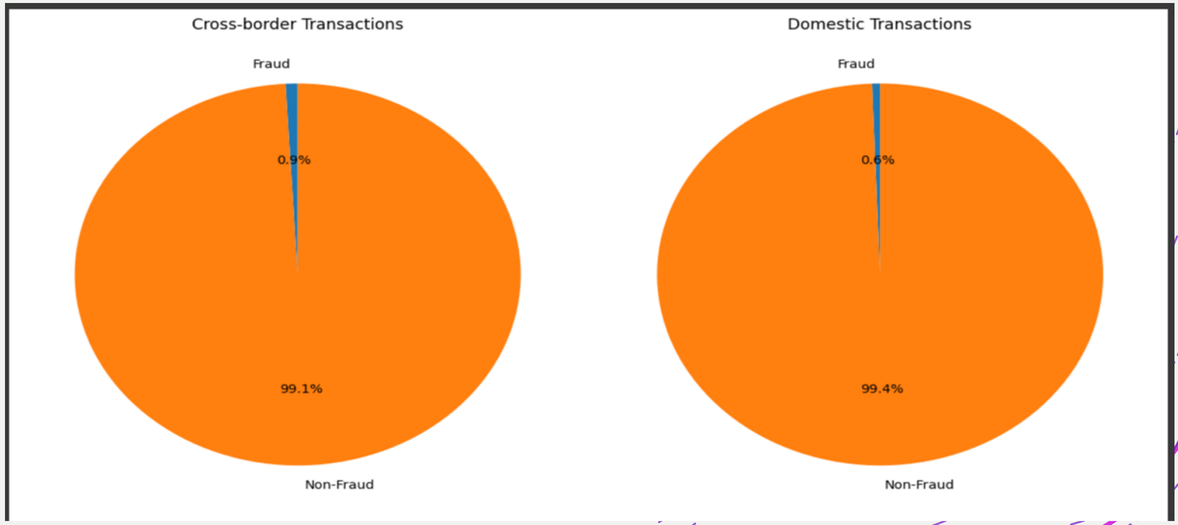
Introduction

+ **Who We Are**: Team Sneak Attack, a group of three curious data enthusiasts.
+ **Problem Statement**: Develop a model to accurately detect fraudulent transactions by analyzing patterns in transaction data, minimizing false positives, and preventing financial losses.
+ **Key Achievements**: Developed and refined models over 11 weeks.
+ Focused on innovative feature engineering and neural network modeling.
+ Achieved significant accuracy in detecting fraudulent transactions.

Data Analysis

- **Exploring Patterns in Transaction Data provided.**
- **Key Insights**:
- The U.S. has the highest fraud rate (57%), significantly more than other countries.
- Fraudulent transactions are rare, accounting for less than 1% of all domestic and cross-border transactions.
- **Visuals Created**:
- Histogram showing fraud and non-fraud distribution by country.
- Pie chart breaking down fraud by transaction type.





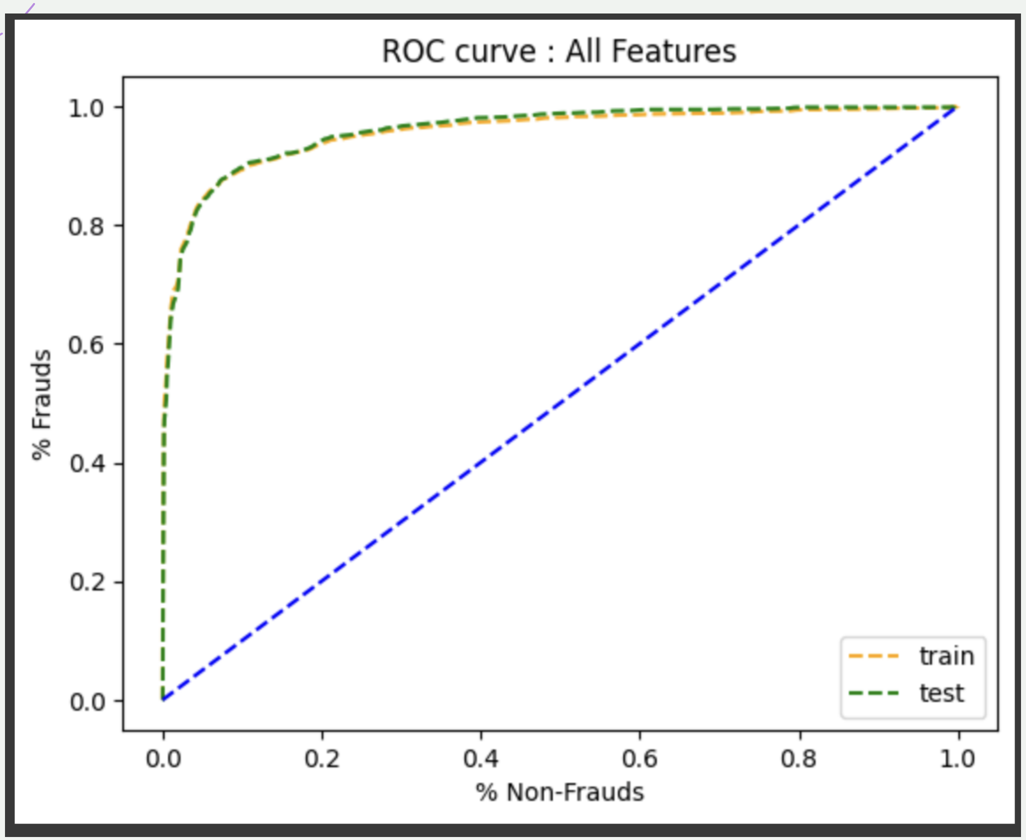
Feature Engineering and Selection

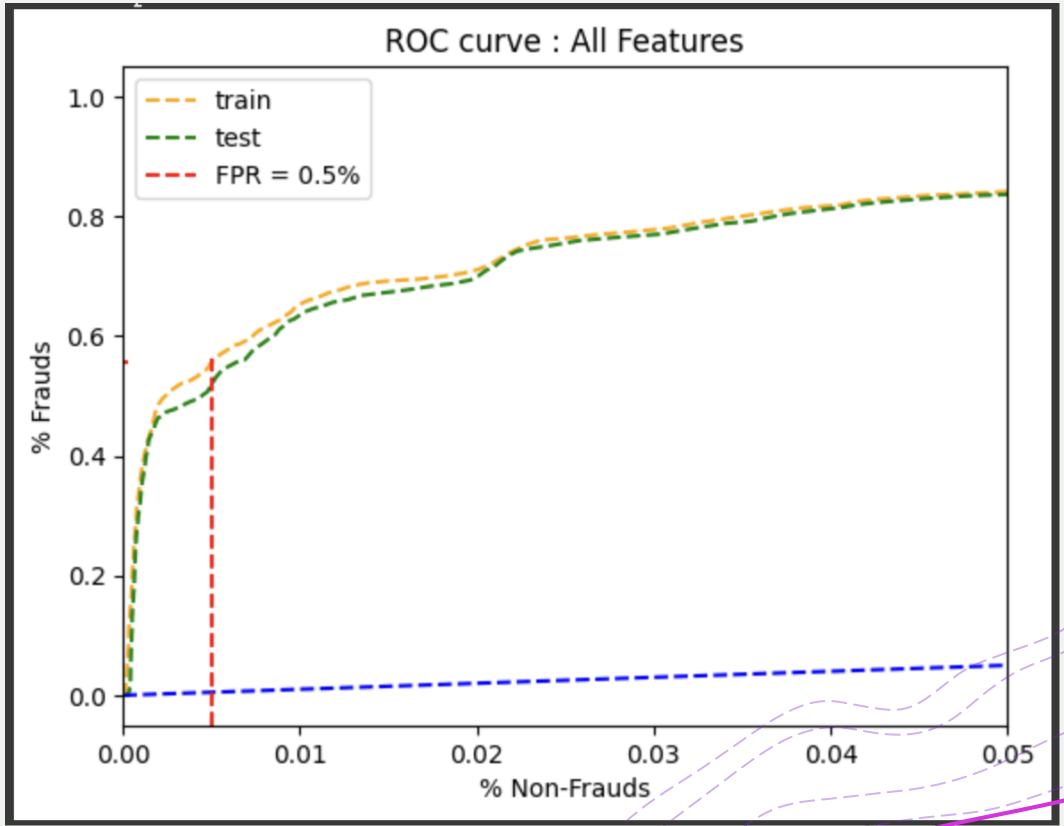
- **Total Input Features:** A total of 19 features were selected and used to train the neural network;
- The features were; [HighValueHourDeviation, HighValueTransactionRate, IS_0_TO_5AM, IS_12_TO_2PM, IsHighValue, RelativeAmount, amount_diff, amt_trend_24h, amt_trend_5e, category_ratio, count_trend_1h, is_cnp, is_international, is_late_night, num_hi_amt_last_hour, num_last_24_hours, repeat_amt, transactionHour, user_avg_amount]
- **Key Team-Developed Features:**
- IS_12_TO_2PM
- HighValueHourDeviation
- HighValueTransactionRate
- IsHighValue:
- Consistency Score
- Spending Spike Score
- Time Since Last Transaction Score
- Transaction Diversity Score
- **Feature Selection:** Chose features that combined transaction behavior, time, and user patterns to maximize fraud detection accuracy.
- “Recent High-Value Transaction Rate”, ran into conversion errors while trying to input this feature.

Feature Engineering and Selection cont.

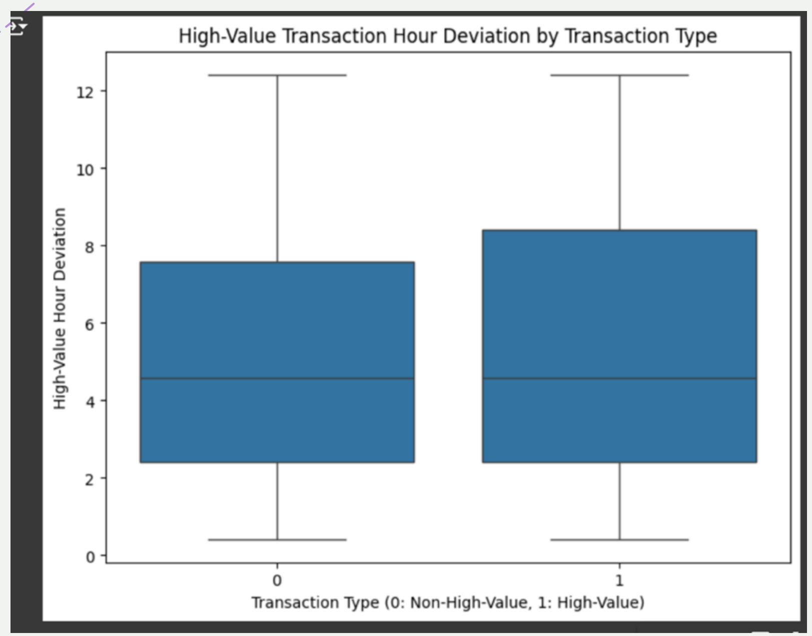
- **Logistic Regression Overview:**
- Machine learning algorithm used for binary classification, ideal for predicting fraud vs. non-fraud transactions.
- **Performance Metrics:**
- Fraud Capture Rate (at 0.005 threshold):
- Train Data: 55.29% of fraud cases correctly identified.
- Test Data: 50.62% of fraud cases correctly identified.
- Indicates consistent performance across datasets.
- **AUC Scores (Area Under the ROC Curve):**
- Card Present (CP): 0.949 – strong ability to distinguish between fraud and non-fraud when the card is physically present.
- Card Not Present (CNP): 0.984 – even better fraud detection for online/remote transactions, where risks are higher.

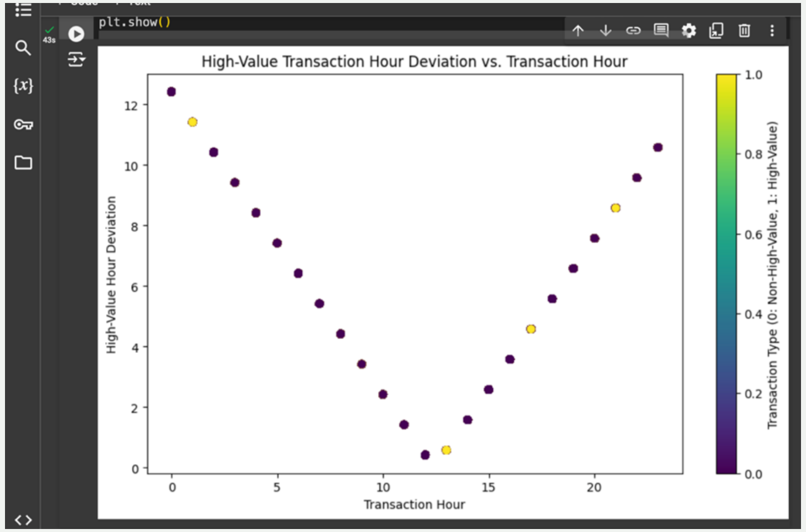
Logistic regression plot

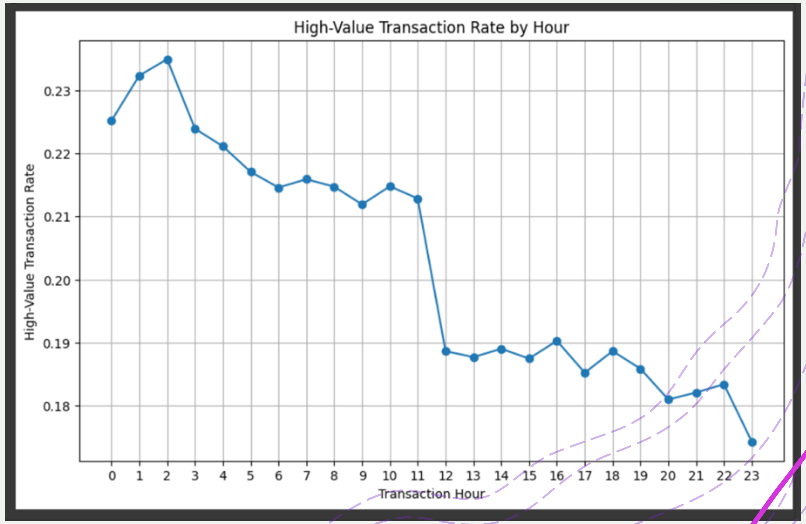




Plots of Features

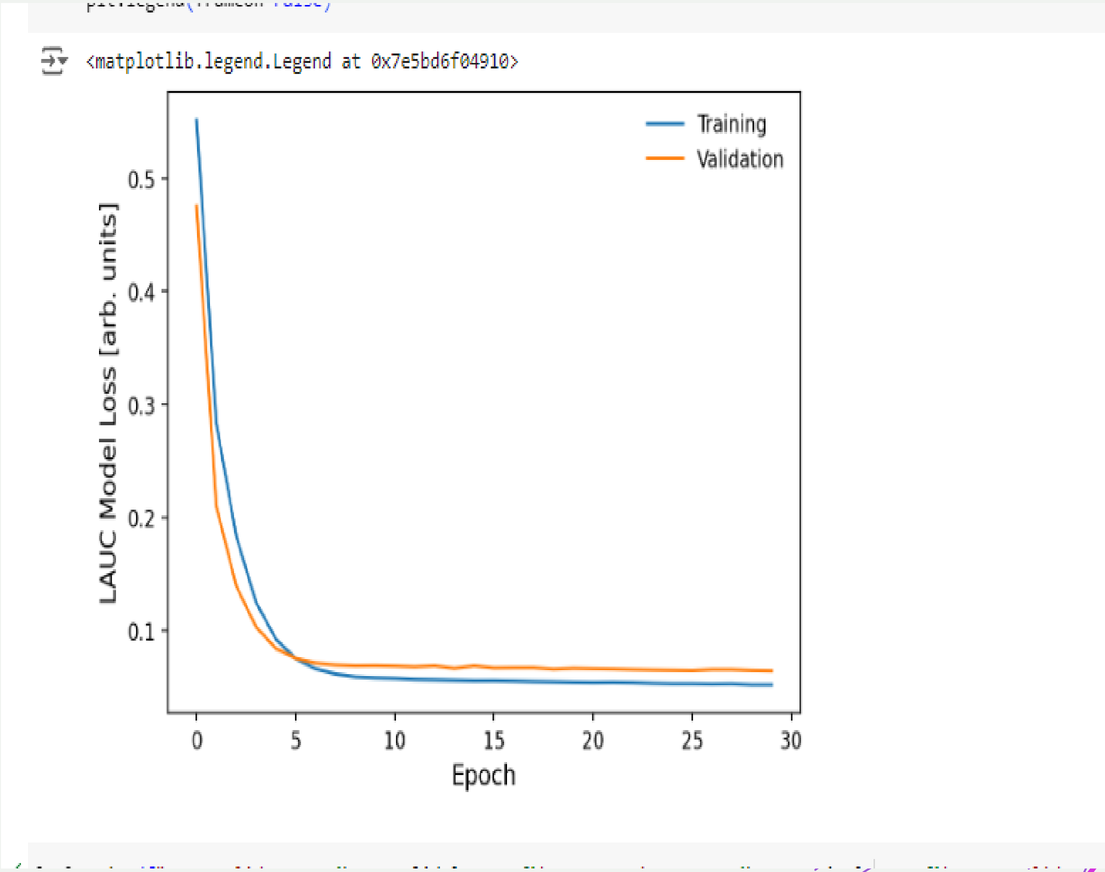




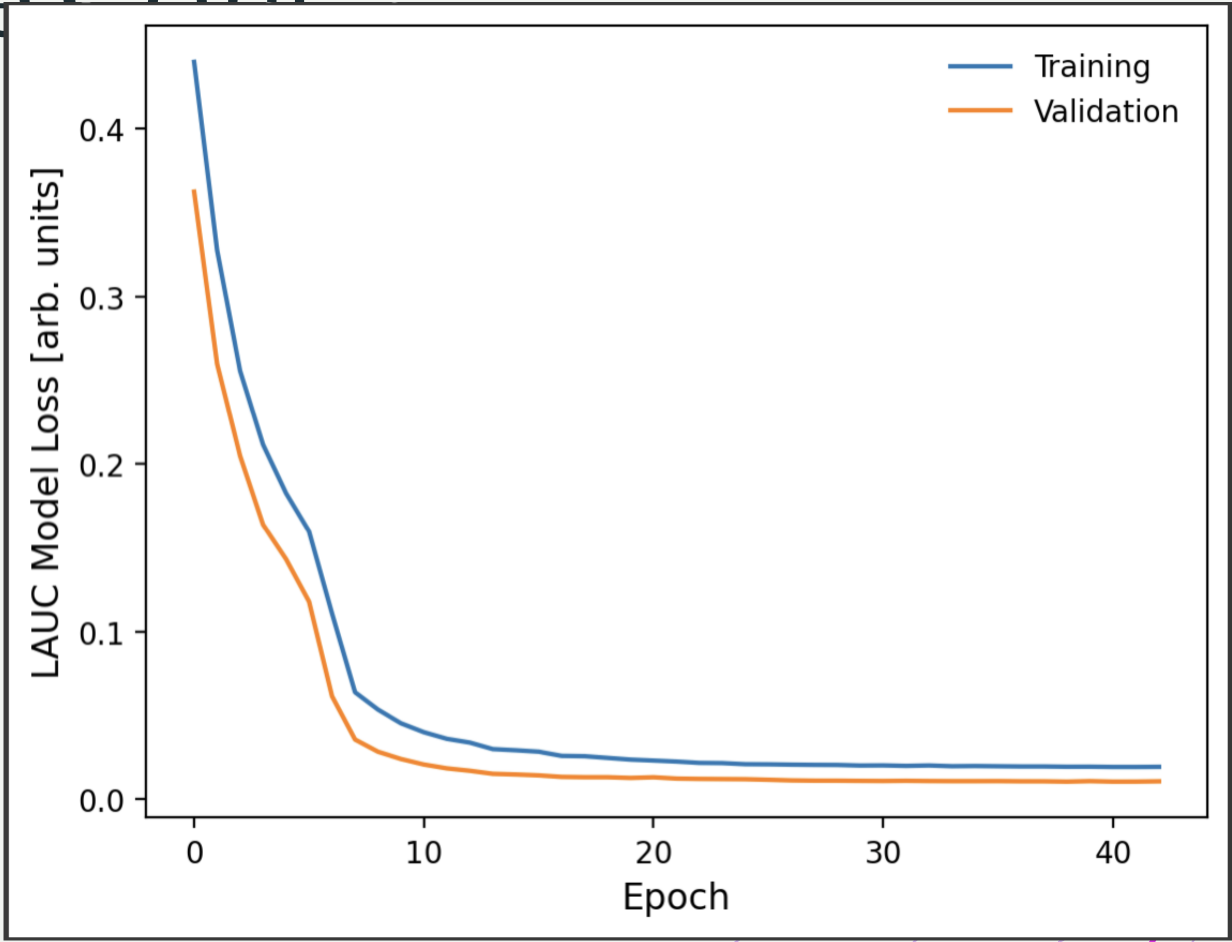


Model Architecture

- Our first model went through several iterations (epoch's) of training to tackle the problem statement of eliminating false positives.
- By training the neural network or the brain of the model with multiple intervals, the model is less prone to give false positives of fraud due to our features.



Model Architecture Cont.



- Our second model includes features engineered from transactional data, such as HighValueTransactionRate, HighValueHourDeviation, and IS_12_TO_2PM.
- Trained with 202 epochs to minimize loss, using LAUC (Logarithmic AUC Loss) as the metric.
- The training and validation losses were recorded for each epoch.
- Key Observations (Epoch Log and Metrics):
- Best Validation Loss: 0.010321, achieved during training.
- Best Train LAUC: 0.915897.
- Best Validation AUC: 0.991837, indicating strong generalization capabilities.
- Best Validation LAUC: 0.941922.

Model Architecture Final

- **Conclusion:**
- This neural network model demonstrated robust performance, achieving a high AUC of 0.991837 with a minimal validation loss.
- Training over 202 epochs ensured stability and consistent improvement, with the best metrics aligning closely during the training process.

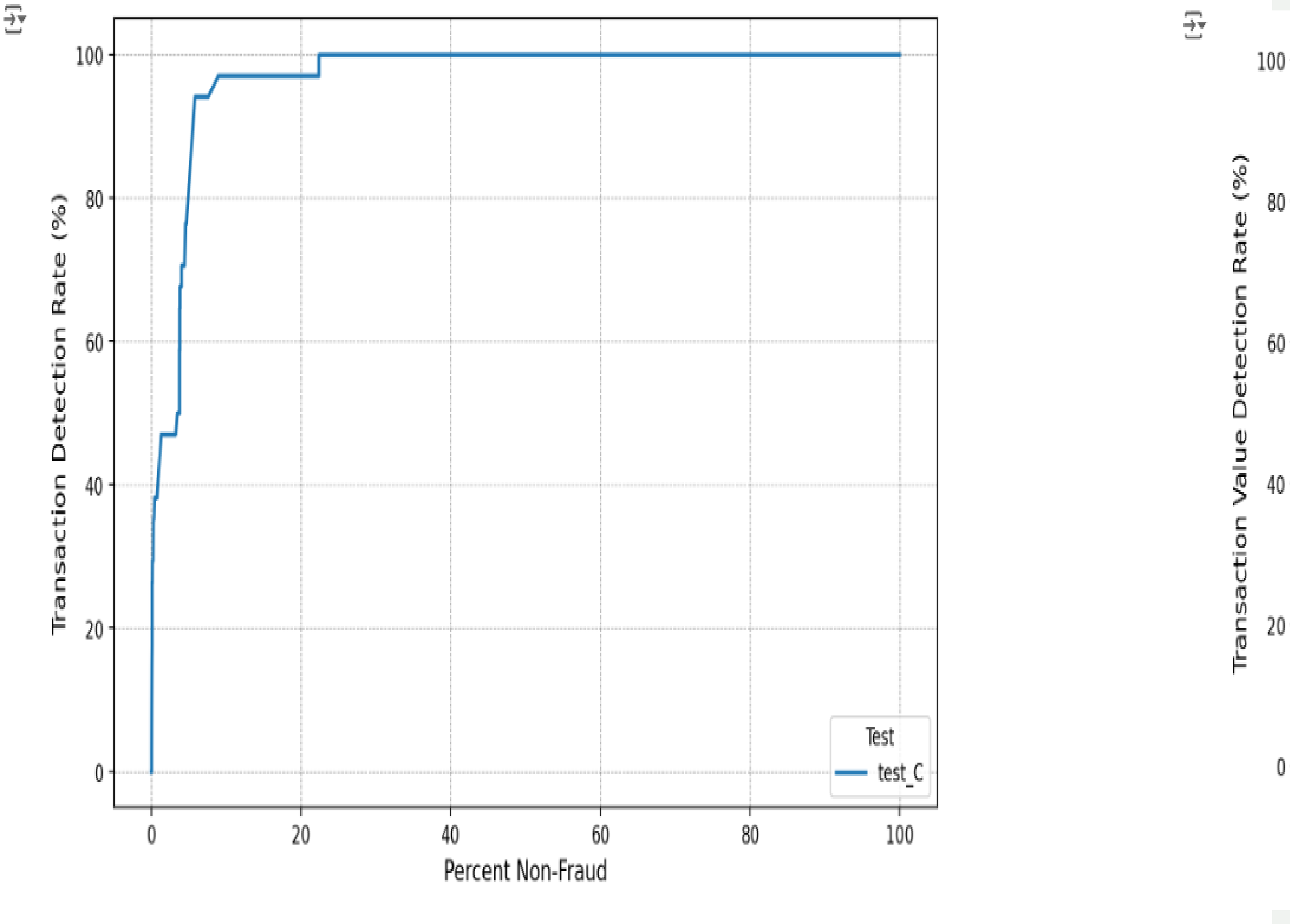
Performance Metrics

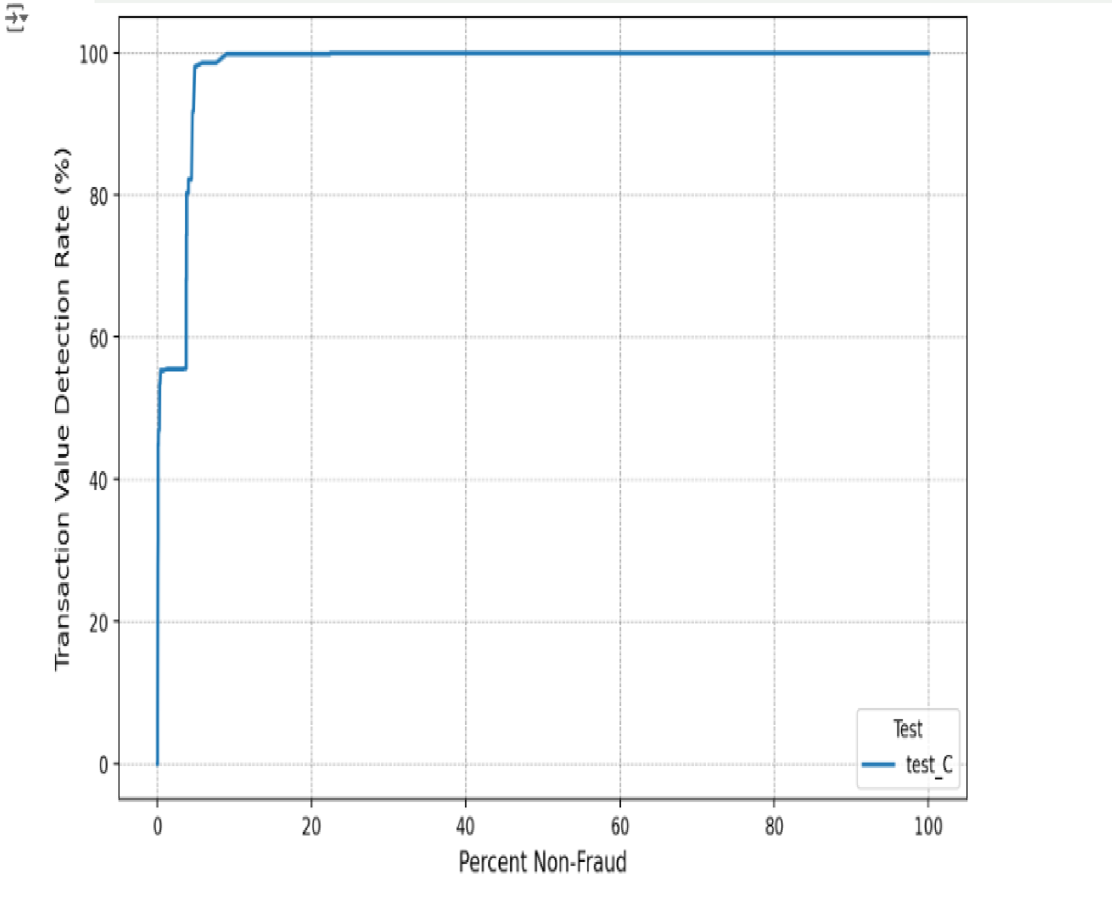
- For the True detection rate, our model captured 82.35% Fraud Transactions and prevented 87.49% Fraud Loss at a 0.5% NF review rate
- For the Acceptance detection rate, our model captured 77.42% Fraud Account and prevented 85.41% Fraud Loss at a 0.87% NF Account review rate
- Our second model captured 16.3% Fraud Account and prevented 27.9% Fraud Loss at a 0.87% NF Account review rate

Performance Metrics Cont'd

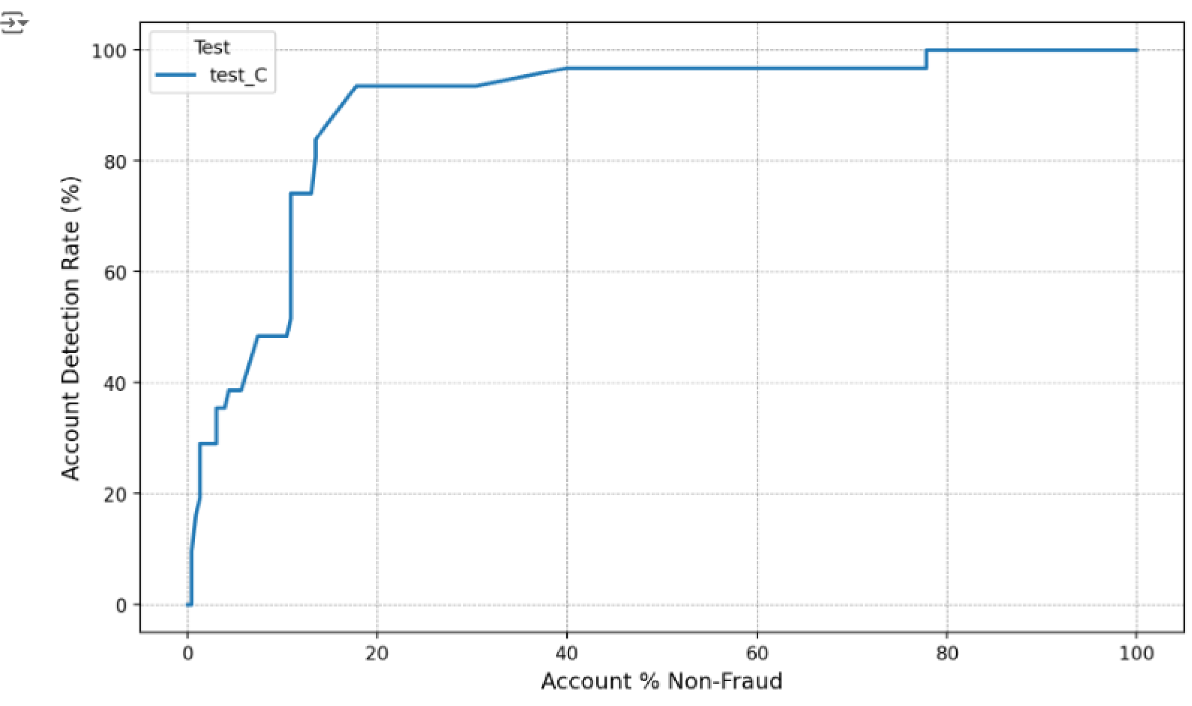
+ Our score out ran two iterations of % non-fraud and account % non-fraud transactions
+ % non-fraud calculates the the number of transactions from non-fraud accounts that scored above the suspect threshold over the total number of transactions from the non-fraud accounts
+ Account % non-fraud calculates this at the account level

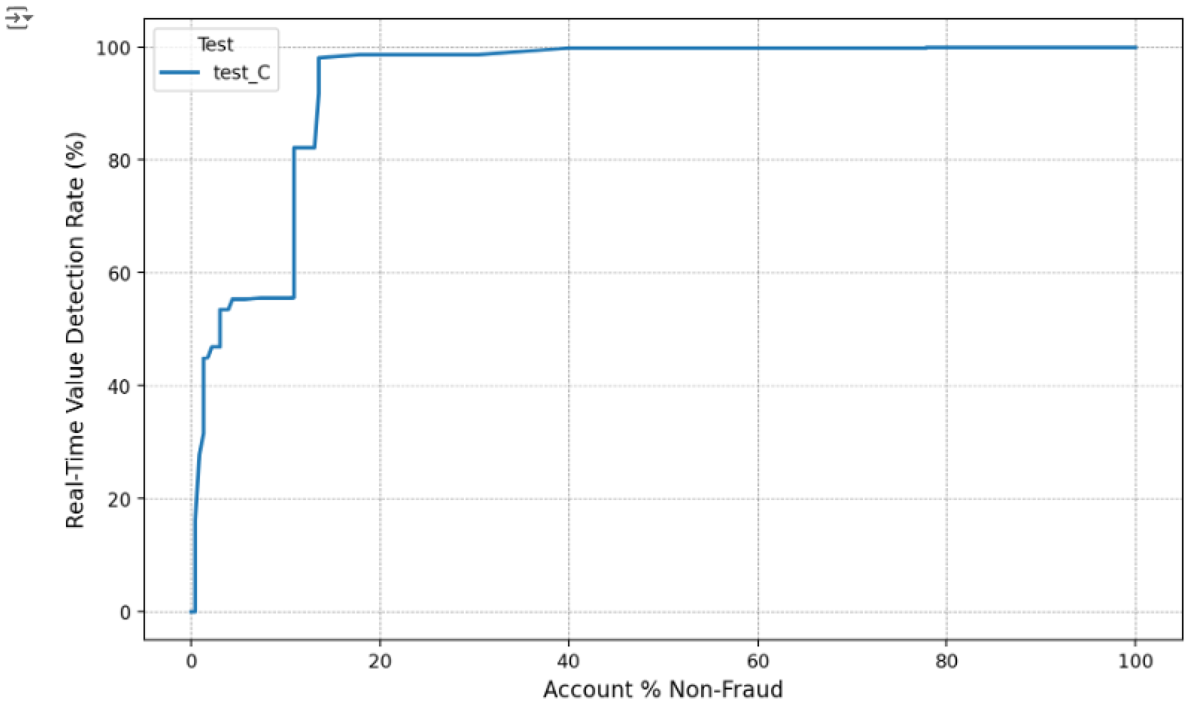
Performance Metrics Cont'd



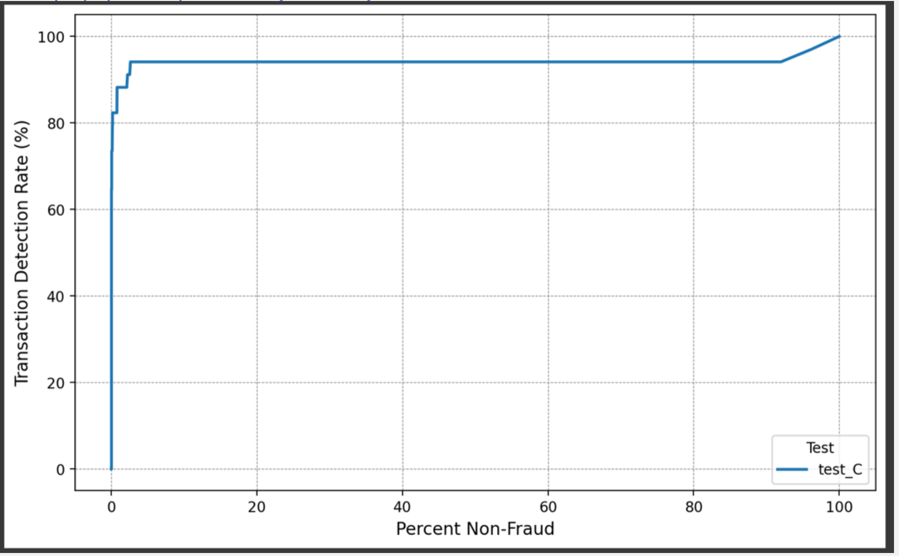


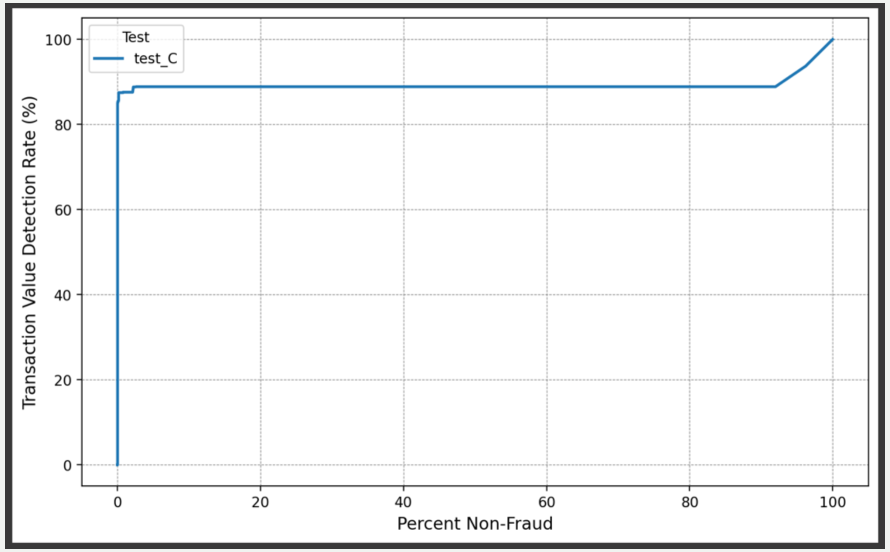
Performance Metrics Cont'd

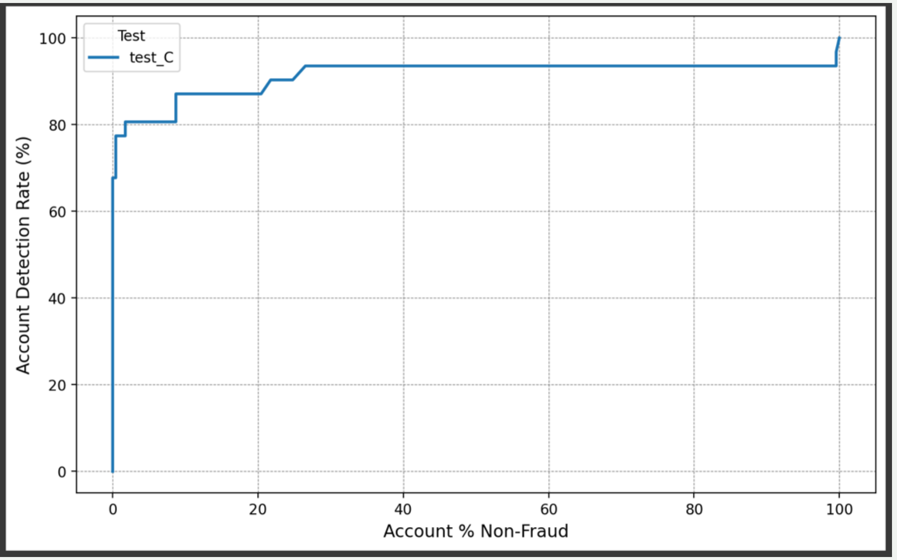


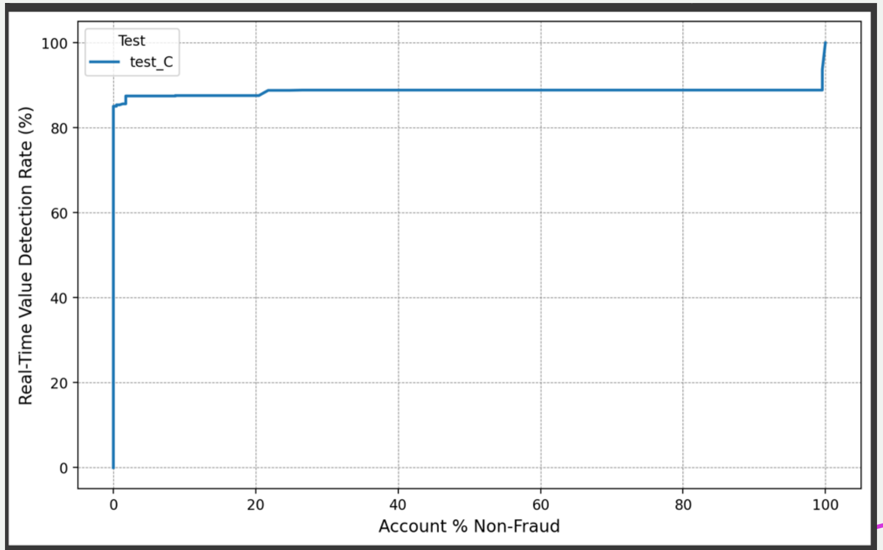


Final performance plots



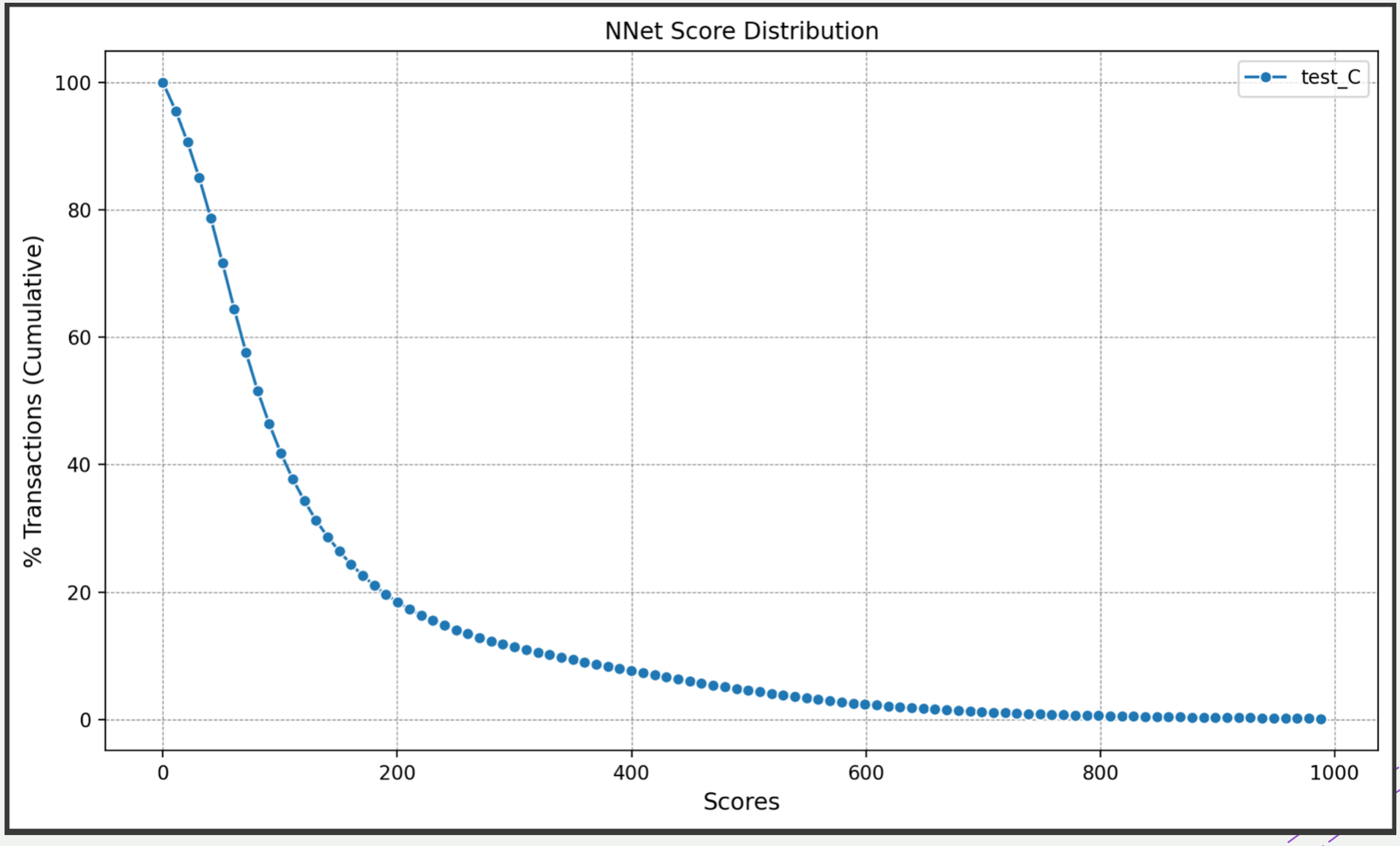






- TDR vs %NF (ROC)
- TVDR vs %NF (Dollar Weighted ROC)
- ADR vs A%NF
- TVDR vs A%NF

Score Distribution Plot



Conclusion

- **Key Takeaways:**
- The neural network achieves high accuracy and AUC in detecting fraud.
- Good features significantly improve model performance.
- **Next Steps:**
- Test on larger and more diverse datasets.

Acknowledgements

- **We would like to thank:**
- Our mentors and instructors for guidance.
- Our team members for their contributions.
- Data sources and tools that made this project possible.