Toronto Metropolitan University

Master of Science in Data Science and Analytics
MRP - Results

Optimizing Supervised Machine Learning for Enhanced Transaction Fraud Detection Accuracy

Submitted to:

MRP Supervisor – Dr. Shengkun Xie
MRP Second Reader – Dr. Ceni Babaoglu

Submitted by:

Nguyen Duy Anh Luong (Student ID – 500968520)

July 28, 2025

Abstract

In this final stage of the project, the focus is on refining the best-performing models and comparing the results with real-world standards, which is the continuation of the Methodology and Experiments stage. Two approaches were tested: a tuned XGBoost model on its own and an ensemble that combined XGBoost with a Deep Neural Network (DNN). After optimization, both models were evaluated using key metrics like precision, recall, F1-score, ROC-AUC, and PR-AUC. The results were also compared to the winning solution from the original Kaggle competition to assess competitiveness. Beyond that, this stage also offers practical insights for applying fraud detection models in real-world industry, highlighting the balance between accuracy, speed, and the need to keep false positives under control.

Data Preparation

This stage begins by loading the training and validation datasets from preprocessed CSV files: X_train.csv and y_train.csv for training, and X_val.csv and y_val.csv for validation. These datasets contain cleaned and transformed features prepared in a prior stage, which included merging the original transaction and identity data, as well as applying necessary pre-processing such as encoding, imputation, and feature extraction.

XGBoost Classifier Standalone

Objective

This section focuses on training and testing an improved standalone XGBoost model using the full labeled training dataset. From earlier experiments, XGBoost stood out as one of the best performance model due to its solid balance of recall, precision, and overall accuracy. In this section, the model is further tuned with optimized hyperparameters and better preprocessing to boost its performance. Its effectiveness is then evaluated using key metrics like recall, precision, F1-score, ROC-AUC, and PR-AUC to see how well it detects fraud while limiting false positives. Finally, the results are compared to the winning Kaggle solution to assess how competitive the model is in a real-world setting.

Trained XGBoost Model

XGB Classifier with advanced hyperparameters:

- n_estimators = 1200
- max_depth = 9
- learning_rate = 0.015
- scale_pos_weight = 27.58
- subsample = 0.92
- colsample_bytree = 0.92

- gamma = 3
- min_child_weight = 4
- reg_alpha = 4
- reg_lambda = 8

Generated Predictions:

- Used model.predict_proba(X_val)[:, 1] to compute fraud probabilities
- Output: 506,691 probability scores
- Range: values between 0 and 1

Validation Performance:

- Recall (Fraud class): 0.8285
- Precision (Fraud class): 0.4148
- F1-score (Fraud class): 0.5528
- ROC-AUC Score: 0.9628
- PR-AUC Score: 0.7676

Confusion Matrix:

	Predicted: Non-Fraud	**Predicted: Fraud**
Actual: Non-Fraud	109,145	4,830
Actual: Fraud	709	3,424

Table 1: Enhanced XGB Classifier - Confusion Matrix

Interpretation:

The enhanced XGBoost model, tuned with advanced hyperparameters, performs strongly on the validation set, especially when it comes to spotting fraudulent transactions. With a ROC-AUC of 0.9628 and a PR-AUC of 0.7676, it shows an excellent ability to distinguish fraud from non-fraud, even with the heavy class imbalance in the data.

Additionally, the model was able to catch over 82% of actual fraud cases, with a recall of 0.8285, a crucial factor in real-world fraud detection, where missing fraud can be costly. Its precision, 0.4148, shows that about 41% of the transactions it flagged were truly fraudulent, which is a reasonable trade-off given the data imbalance and the priority on catching fraud. The resulting F1-score of 0.5528 shows the model strikes a fair balance between identifying fraud and keeping false alarms under control.

The confusion matrix provides further insight:

- Out of all fraud cases, 3,424 were correctly detected, while 709 were missed.
- Among non-fraud cases, 4,830 were incorrectly flagged as fraud, a manageable number given the scale of the data (113,975 total non-fraud cases).

Overall, these results suggest that the model is highly effective at prioritizing potentially fraudulent transactions, offering strong support for real-world fraud detection systems.

Curve Analysis:

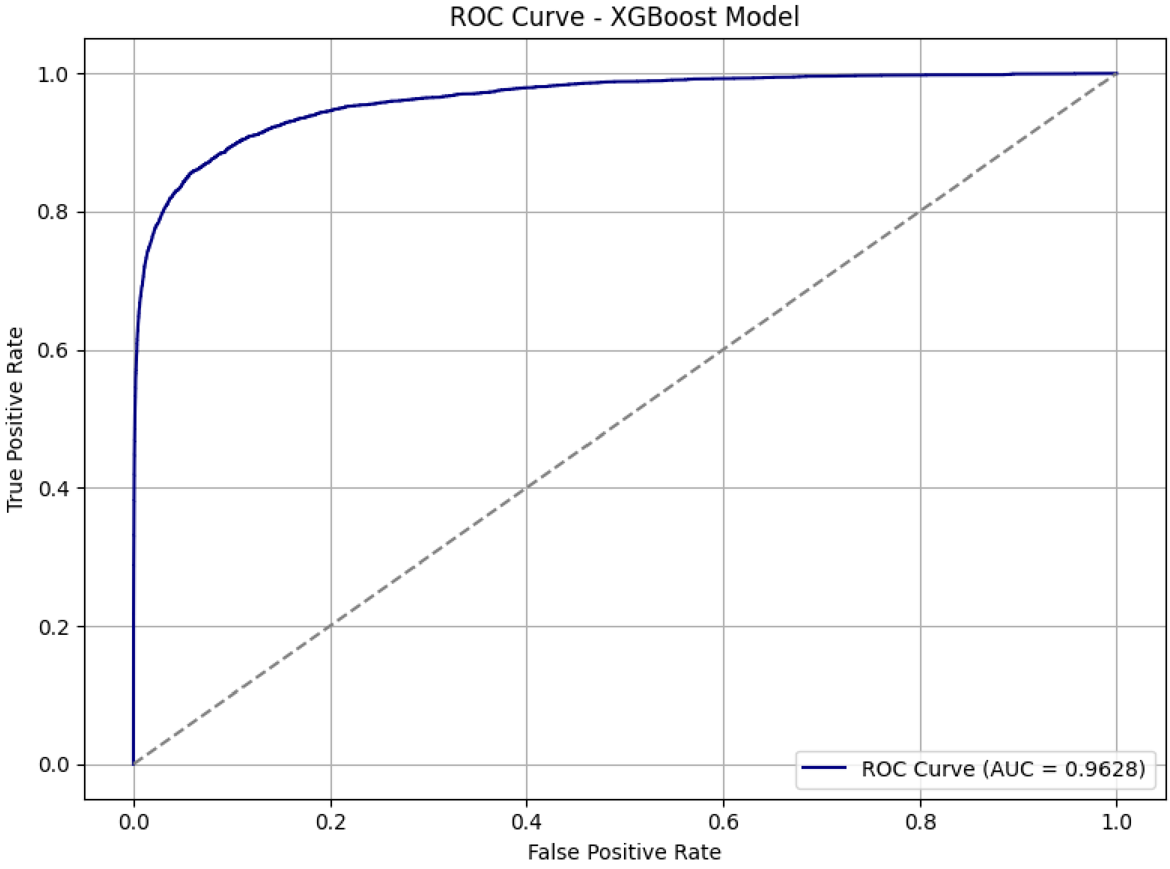

Figure 1: ROC Curve - Enhanced XGBoost

ROC Curve: The ROC curve shows that the XGBoost model does a great job separating fraudulent transactions from legitimate ones, with a strong ROC-AUC score of 0.9628. The curve's sharp rise toward the top-left corner highlights that the model achieves high recall while keeping false positives under control. This balance is important in fraud detection, by catching as much fraud as possible without overload investigators with unnecessary alerts.

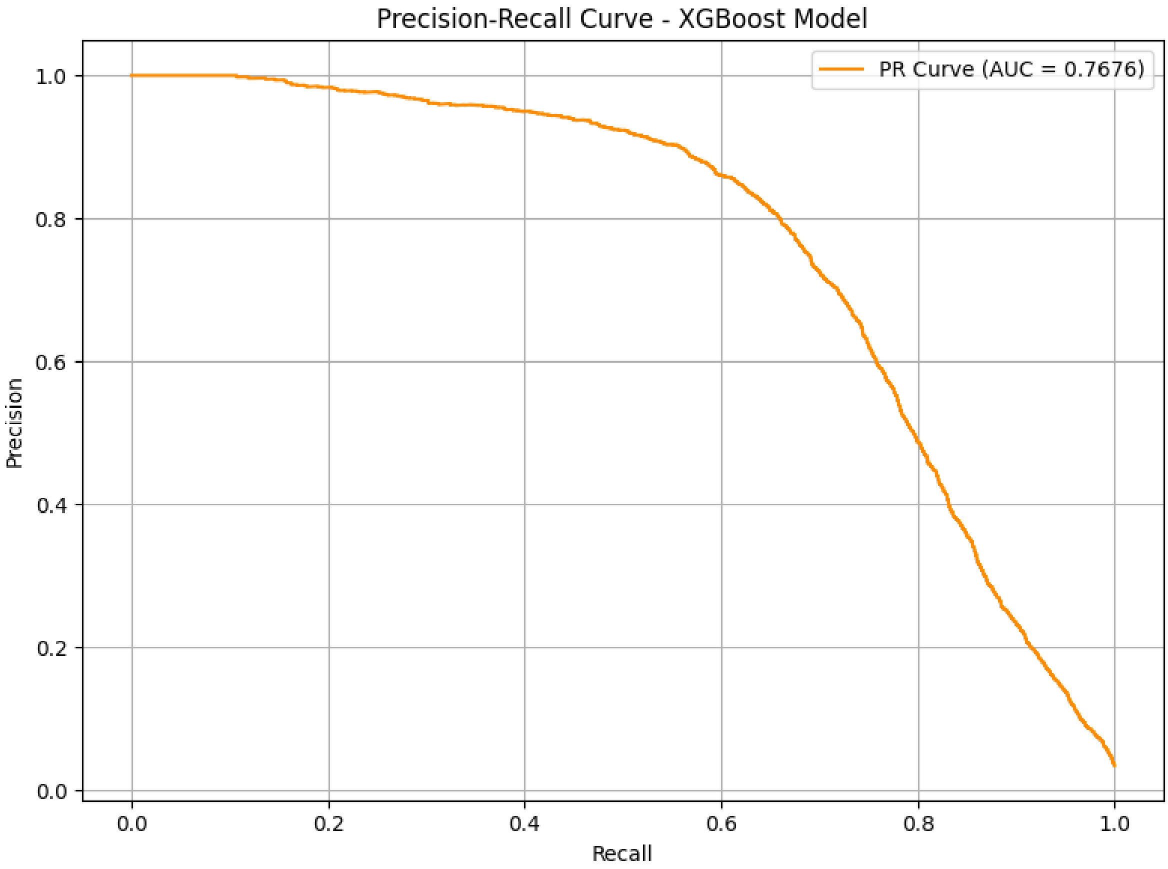

Figure 2: Precision-Recall Curve - Enhanced XGBoost

Precision-Recall Curve: The Precision-Recall curve shows the XGBoost model maintains solid performance when balancing false positives and missed fraud cases, with a PR-AUC of 0.7676. Precision stays relatively high across a wide range of recall values, meaning the model can catch a large number of fraudulent transactions.

Model Comparison - MRP Enhanced XGBoost vs. 1st Place XGBoost

Metric	MRP XGBoost	1st Place XGBoost (Public)	1st Place XGBoost (Private)
ROC-AUC	0.9628	0.9602	0.9324

Table 2: ROC-AUC comparison between MRP XGBoost and the 1st place model

Overall, the XGBoost model developed in this MRP achieved a higher ROC-AUC on the validation set compared to the 1st place team's standalone XGBoost model, which recorded a ROC-AUC of 0.9602 on the public leaderboard and 0.9324 on the private leaderboard. This result highlights the strong and competitive performance of the proposed approach.

XGBoost + DNN Ensemble

Objective

This section focuses on building a hybrid fraud detection model that combines the strengths of both XGBoost and a Deep Neural Network (DNN). XGBoost is great at handling structured, tabular data and picking up clear decision patterns, while the DNN can learn more abstract, high-level representations that might not be obvious to tree-based models. By combining their predictions in an ensemble, the goal is to improve overall performance, especially in detecting hard-to-catch fraud cases. The model is evaluated using the same key metrics as before (recall, precision, F1-score, ROC-AUC, and PR-AUC), and results are compared to both the standalone XGBoost model and the winning solution from the Kaggle competition to see if this combined approach offers a better result.

Trained XGB + DNN Model

XGBoost Submodel:

- Used the same upgraded tuned parameters as in the standalone model
- Trained directly on raw (non-scaled) tabular features
- Output: Probability scores from predict_proba(X_val)[:, 1]

DNN Submodel:

- Model Architecture:
  - Dense(1024) → BatchNorm → Dropout(0.4)
  - Dense(512) → BatchNorm → Dropout(0.3)
  - Dense(128) → BatchNorm → Dropout(0.2)
  - Output: Dense(1, activation=`'sigmoid'`)
- Loss: Custom binary focal loss ($\gamma = 2.0$, $\alpha = 0.25$)
- Optimizer: Adam (learning rate = 0.0005)
- Used class weights and early stopping
- Trained on scaled input features

Ensemble Strategy:

- Combined model outputs using a weighted average:
  - ensemble_probs = 0.6 * xgb_probs + 0.4 * dnn_probs
- Threshold optimized using precision-recall curve to maximize F1-score

Validation Performance:

- Recall (Fraud class): 0.6927
- Precision (Fraud class): 0.7818
- F1-score (Fraud class): 0.7346
- ROC-AUC Score: 0.9629
- PR-AUC Score: 0.7898

Confusion Matrix:

	Predicted: Non-Fraud	**Predicted: Fraud**
Actual: Non-Fraud	113,155	820
Actual: Fraud	1,243	2,890

Table 3: XGBoost + DNN Ensemble - Confusion Matrix

Interpretation:

The XGBoost + DNN ensemble performs strongly at spotting fraudulent transactions. On the validation set, it achieved a ROC-AUC of 0.9629, showing it can clearly distinguish fraud from non-fraud. Its PR-AUC of 0.7898 also highlights how well it handles the tricky balance between precision and recall, which is especially important for highly imbalanced data.

For fraud cases specifically, the model caught about 69% of actual fraud (recall = 0.6927) while keeping most flagged cases accurate (precision = 0.7818). Its F1-score of 0.7346 shows it strikes a good balance between catching fraud and avoiding too many false alarms.

The confusion matrix confirms this performance:

- True Positives (2,890): Correctly identified fraud cases.
- False Negatives (1,243): Fraud cases missed by the model.
- False Positives (820): Legitimate transactions incorrectly flagged as fraud.
- True Negatives (113,155): Correctly identified non-fraudulent transactions.

Overall, the ensemble model not only outperforms the individual models in ROC-AUC but also provides a stronger balance between catching fraud and avoiding false alarms. This balance makes it more practical for real-world fraud detection, where both missed fraud and unnecessary alerts can be costly.

Curve Analysis:

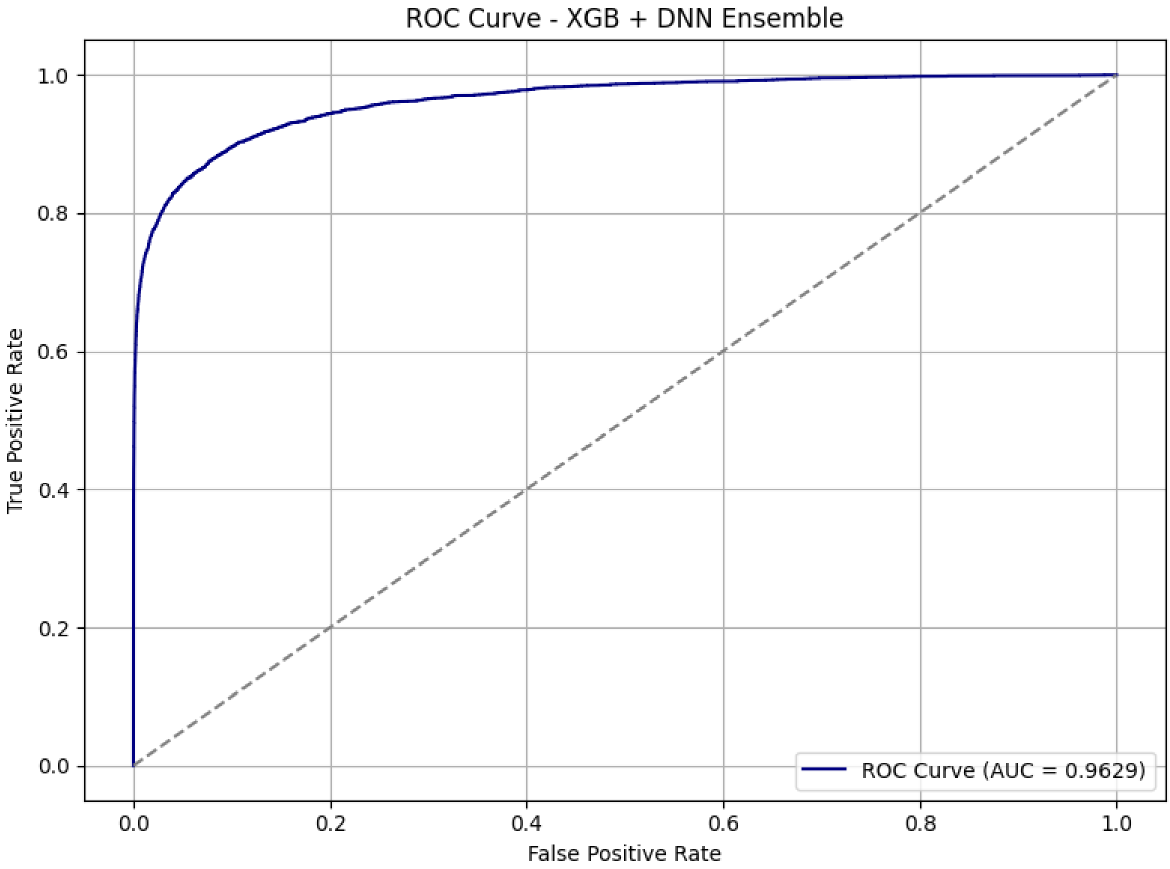

Figure 3: ROC Curve - XGB + DNN Ensemble

ROC Curve: The ROC curve for the XGBoost + DNN ensemble shows a strong ability to separate fraudulent from legitimate transactions, with an AUC of 0.9629. The curve rises steeply toward the top-left corner, indicating the ensemble maintains high true positive rates while keeping false positives low, an important balance for fraud detection systems.

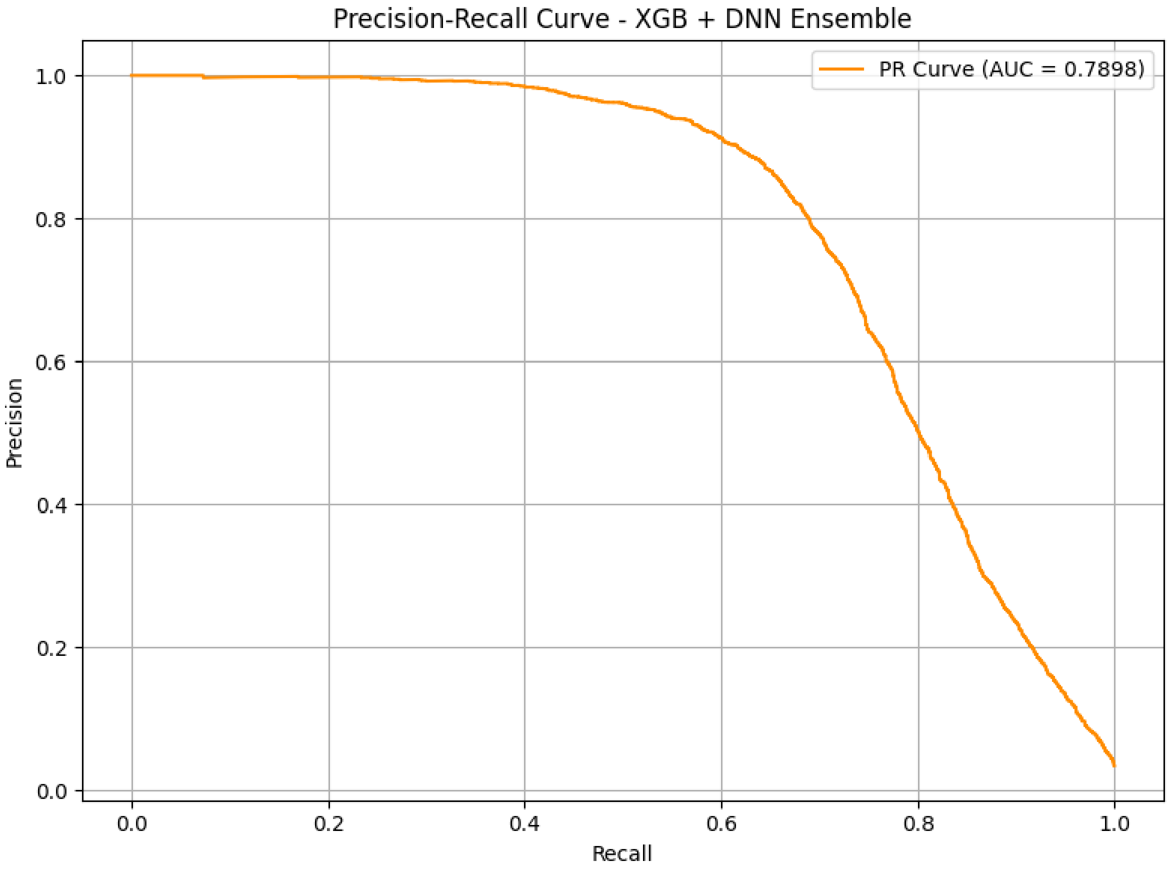

Figure 4: Precision-Recall Curve - XGB + DNN Ensemble

Precision-Recall Curve: The Precision-Recall curve for the XGBoost + DNN ensemble highlights its ability to balance precision and recall across different thresholds. With a PR-AUC of 0.7898, the model maintains strong precision even as recall increases, making it effective at catching fraud without distracting the system with false positives.

Model Comparison - MRP XGBoost + DNN Ensemble vs. 1st Place XGBoost

Metric	MRP XGB + DNN (4 Layers)	1st Place XGBoost (Public)	1st Place XGBoost (Private)
ROC-AUC	0.9629	0.9602	0.9324

Table 4: ROC-AUC comparison between MRP XGBoost + DNN Ensemble and the 1st place model

Overall, the ensemble model achieved a validation ROC-AUC of 0.9629, outperforming the 1st place team's standalone XGBoost model, which scored 0.9602 on the public leaderboard and 0.9324 on the private leaderboard. This improvement demonstrates a better balance between recall and precision, which is an essential feature in real-world fraud detection, where reducing false positives and identifying fraudulent activities are equally important.

Conclusion

This final stage of the project demonstrated how optimized machine learning methods, such as a tuned XGBoost model and a hybrid XGBoost + Deep Neural Network (DNN) ensemble, can significantly improve transaction fraud detection. Both models achieved strong results on the validation set, with ROC-AUC scores of 0.9628 for the standalone XGBoost and 0.9629 for the ensemble. The ensemble, however, provided a higher PR-AUC of 0.7898 and a stronger balance between precision (0.7818) and recall (0.6927), making it more effective at minimizing both missed fraud cases and false positives.

Compared to the original Kaggle competition's 1st place XGBoost solution (ROC-AUC 0.9602 public, 0.9324 private), both models achieved better performance, highlighting the benefits of advanced tuning and ensemble strategies. The findings highlight that combining tree-based models with deep learning techniques can enhance detection accuracy, providing a reliable approach for real-world fraud prevention systems. Future improvements could explore dynamic thresholding, additional feature engineering, or alternative ensemble strategies to further optimize results.

Insights and Recommendations for Real-World Deployment

Applying the XGBoost and XGBoost + DNN ensemble models into production requires more than just strong validation metrics. To ensure practical success in real-world fraud detection, the following recommendations are worth to consider:

- **Threshold Optimization:** Adjust decision thresholds based on business needs, balancing fraud detection with acceptable false positive rates.
- **Model Monitoring and Retraining:** Continuously track performance metrics and detect data changes. Retrain models regularly to adapt to evolving fraud tactics.
- **Explainability and Compliance:** Use feature importance (for XGBoost) and tools such as SHAP or LIME (for DNN) to provide transparency, ensuring regulatory and operational trust.
- **A/B Testing and ROI Measurement:** Any production launch should be implemented with controlled A/B testing to quantify impacts on fraud loss reduction, customer satisfaction, and operational workload.

By following these recommendations, organizations can maximize the value of machine learning-based fraud detection systems, ensuring not only strong detection performance, but also operational efficiency, scalability, and adaptability to evolving fraud tactics.

GitHub Repository

All codes and visualizations for this Results stage are available at the following GitHub repository:
Link: `https://github.com/nguyenduyanhluong/TMU-MRP-2025/tree/main/results`