Fraud Detection in Digital Transactions: Stage 2 Report

Tom Delahaye, Gabriel Carlotti, Pierre Briand, Kentin Guillemot

November 28, 2024

Abstract

This report presents the second stage of a multi-phase project focused on fraud detection in digital payments. Building on the baseline models implemented in Stage 1, this stage leverages advanced ensemble methods such as Random Forest, XGBoost, and a stacking approach combining these methods. The results demonstrate significant improvements in recall, precision, and F2-score, providing a robust foundation for future stages.

Contents

1 Introduction

Digital payments have become a cornerstone of modern financial systems, offering speed and convenience to users worldwide. However, the increasing volume of digital transactions has also heightened exposure to fraud, presenting a critical challenge for financial institutions. Detecting fraudulent transactions with high accuracy is essential to minimize financial losses and maintain user trust.

In Stage 1, the project implemented two baseline models: Logistic Regression and K-Nearest Neighbors (KNN). Despite their simplicity, these models achieved commendable results:

- Logistic Regression demonstrated strong interpretability and scalability for large datasets, making it a reliable starting point.
- KNN delivered excellent classification performance, especially in terms of recall, proving its effectiveness for fraud detection.

While these models provided a robust foundation, certain challenges remained, particularly in terms of computational efficiency and the ability to handle non-linear relationships in the data. Stage 2 aims to address these limitations by introducing advanced ensemble methods (Random Forest, XGBoost, and Stacking), leveraging their strengths to further enhance precision and recall.

—

2 Improvement Assumptions

2.1 Strengths and Limitations of Stage 1 Models

Stage 1 models, despite their strengths, revealed some areas for improvement:

- **Strengths:**
  - **Logistic Regression:** Provided solid results with a recall-focused approach, ensuring that most fraudulent transactions were identified. Its efficiency in training on large datasets was a key advantage.
  - **KNN:** Achieved excellent classification metrics, particularly for recall, proving effective in detecting fraud despite its simplicity.
- **Limitations:**
  - Logistic Regression struggled to capture non-linear relationships in the data, which can limit its effectiveness in modeling complex fraud patterns.
  - KNN, while accurate, faced scalability challenges due to its computational cost for large datasets.
  - Both models showed sensitivity to synthetic data generated by SMOTE, which could potentially affect generalization to unseen data.

2.2 New Assumptions

Building on the strong performance of Logistic Regression and KNN, Stage 2 hypothesizes that ensemble methods will further enhance performance by addressing the limitations identified in Stage 1:

- **Improved Handling of Non-Linearity:** Tree-based models like Random Forest and XGBoost are inherently capable of capturing complex, non-linear relationships, which will improve classification accuracy.
- **Enhanced Precision and Recall:** By leveraging the complementary strengths of Random Forest and XGBoost, stacking will provide a better balance between precision and recall, minimizing both false positives and false negatives.
- **Scalability:** Ensemble methods are more computationally efficient for large datasets compared to KNN, while maintaining high classification performance.

3 Methodology

3.1 Data Preprocessing

- Outliers were handled using Z-scores, similar to Stage 1, to ensure consistent input quality.
- SMOTE was applied to balance the dataset, ensuring better representation of fraudulent transactions.
- The dataset was split into training and validation sets using stratified sampling to preserve class distributions.

3.2 Model Selection and Optimization

Random Forest:

- Hyperparameters (`n_estimators`, `max_depth`) were optimized using GridSearchCV.

XGBoost:

- Optuna was used for hyperparameter tuning, exploring parameters such as `eta` (learning rate) and `lambda` (regularization).
- Optuna was chosen for its faster runtime compared to GridSearchCV while achieving similar performance.

Stacking:

- The meta-model (logistic regression) was optimized with GridSearchCV, using recall as the guiding metric to prioritize minimizing false negatives.

3.3 Ensemble Learning

The stacking approach combined the strengths of Random Forest and XGBoost, with a logistic regression meta-model to integrate their predictions. This method demonstrated the best balance between precision and recall.

3.4 Monitoring Overfitting and Early Stopping

To ensure proper generalization and evaluate the risk of overfitting, learning curves were generated for both Random Forest and XGBoost. These curves provide a visual representation of the model's performance on both training and validation datasets as the training set size increases. Early stopping was implemented for XGBoost to mitigate overfitting and optimize training efficiency.

3.4.1 Random Forest

Learning Curve Analysis: The learning curve for Random Forest (Figure 1) demonstrates the following:

- **Training Performance:** The training score consistently reaches a near-perfect value of 1.0, indicating that the model perfectly fits the training data.
- **Validation Performance:** The validation score steadily improves as the training set size increases, converging close to the training score. This indicates that the model generalizes well without significant overfitting.
- **Conclusion:** The minimal gap between training and validation scores suggests that Random Forest effectively captures the underlying data patterns while avoiding overfitting. However, the near-perfect scores also indicate that the task may be relatively easy for this model or that the dataset is well-suited for Random Forest.

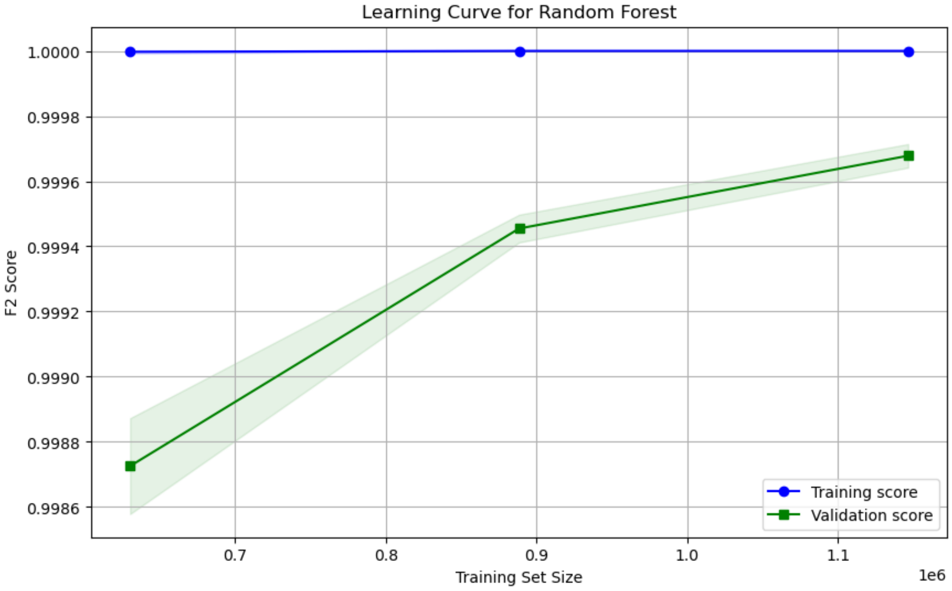

Figure 1: Learning Curve for Random Forest.

3.4.2 XGBoost

Learning Curve Analysis: The learning curve for XGBoost (Figure 2) exhibits the following characteristics:

- **Training Performance:** The training score is consistently high, reflecting the model's ability to fit the training data well.
- **Validation Performance:** The validation score starts slightly lower than the training score but quickly converges as the training set size increases. This suggests that XGBoost benefits significantly from larger training datasets.
- **Generalization:** The minimal gap between the training and validation scores indicates good generalization, with no significant signs of overfitting.

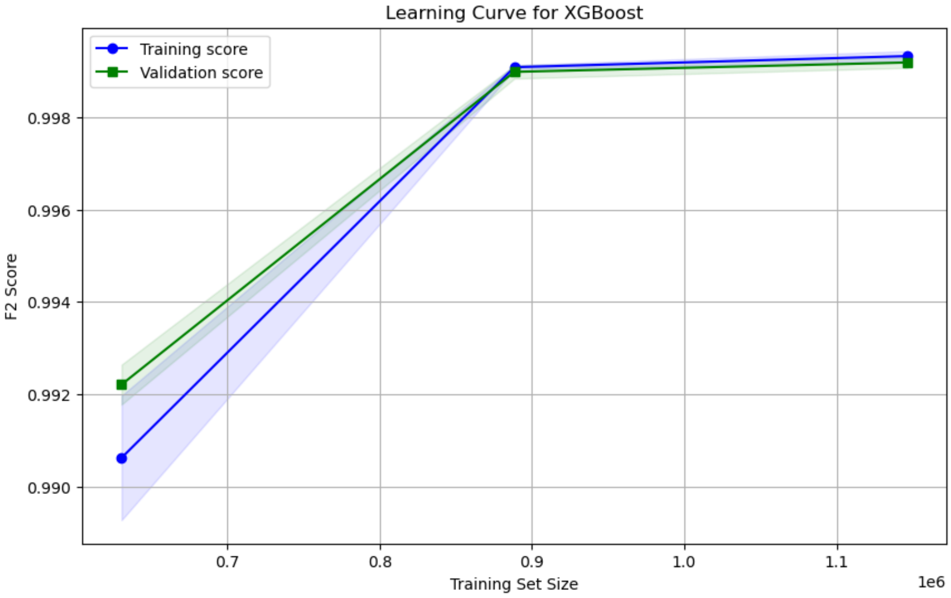

Figure 2: Learning Curve for XGBoost.

Conclusion: The learning curves for both Random Forest and XGBoost demonstrate strong and consistent performance, with minimal gaps between training and validation scores. This indicates that both models generalize well to unseen data. The application of early stopping for XGBoost further enhances its training efficiency while maintaining high performance, showcasing its suitability for computationally intensive tasks such as fraud detection.

4 Results and Evaluation

4.1 Performance Metrics

The evaluation of the models relied on several key performance metrics:

- **Precision:** Measures the proportion of correctly predicted fraudulent transactions among all predicted fraudulent transactions.
- **Recall:** Measures the proportion of actual fraudulent transactions correctly identified by the model, crucial for minimizing false negatives.
- **F2-Score:** A metric that prioritizes recall over precision to align with the goal of minimizing missed fraud cases.
- **False Positives (FP):** The number of non-fraudulent transactions incorrectly classified as fraudulent.
- **False Negatives (FN):** The number of fraudulent transactions missed by the model.

4.2 Model Comparison

The performance of Random Forest, XGBoost, and the Stacking model is summarized in Table 1. Random Forest and XGBoost were optimized using the F2-score as the guiding metric, consistent with Stage 1. However, for the Stacking model, the grid search for hyperparameter tuning prioritized recall. This decision aligns with the critical objective of fraud detection, which emphasizes identifying as many fraudulent transactions as possible, even at the expense of a slightly higher false positive rate.

Model	F2-Score	Recall	False Positives	False Negatives	Complexity
Random Forest	**0.99972**	0.99977	122	4	Low
XGBoost	0.99918	0.99760	358	42	Moderate
Stacking	**0.99980**	**0.99971**	117	5	High

Table 1: Performance comparison of Random Forest, XGBoost, and Stacking.

Random Forest: The Random Forest model achieved exceptional results, achieving the highest F2-score of **0.99972**. The confusion matrix in Figure 3 illustrates its performance:

- **True Positives (TP):** 17,477
- **True Negatives (TN):** 182,397
- **False Positives (FP):** 122
- **False Negatives (FN):** 4

This model demonstrated excellent generalization, with minimal false positives and false negatives. Its simplicity and low computational complexity make it ideal for practical applications.

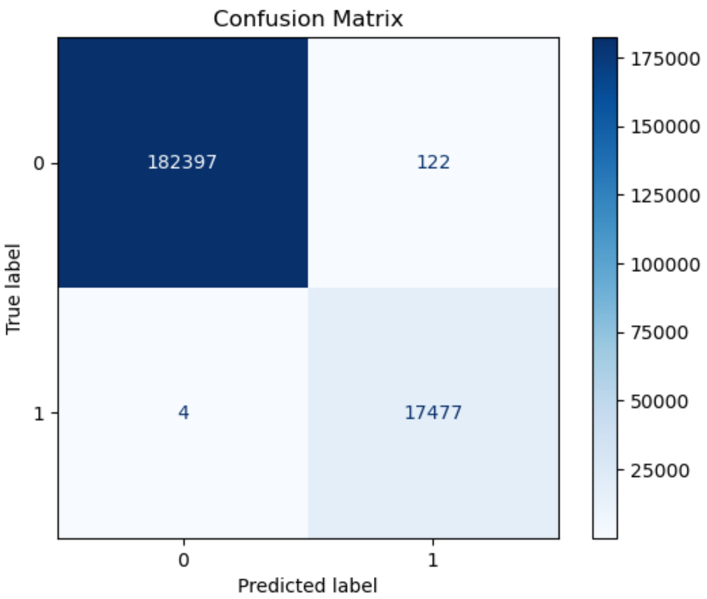

Figure 3: Confusion Matrix for Random Forest.

XGBoost: XGBoost, while slightly behind Random Forest, delivered robust results, achieving an F2-score of **0.99918**. The confusion matrix in Figure 4 highlights its performance:

- **True Positives (TP):** 17,439
- **True Negatives (TN):** 182,161
- **False Positives (FP):** 358
- **False Negatives (FN):** 42

Despite its higher false positive and negative rates compared to Random Forest, XGBoost's gradient boosting framework and regularization mechanisms ensured competitive results. However, its computational cost was higher due to its iterative nature and hyperparameter optimization.

Stacking: The Stacking model, leveraging the outputs of Random Forest and XGBoost, achieved the highest recall (**0.99971**) and an F2-score of **0.99980**. The confusion matrix in Figure 5 provides further insights:

- **True Positives (TP):** 17,476
- **True Negatives (TN):** 182,402
- **False Positives (FP):** 117
- **False Negatives (FN):** 5

By optimizing the meta-model using recall, the Stacking framework prioritized minimizing false negatives, aligning with the project's primary objective. While the model demonstrated exceptional performance, it required significant computational resources due to the training of multiple base models and the meta-model.

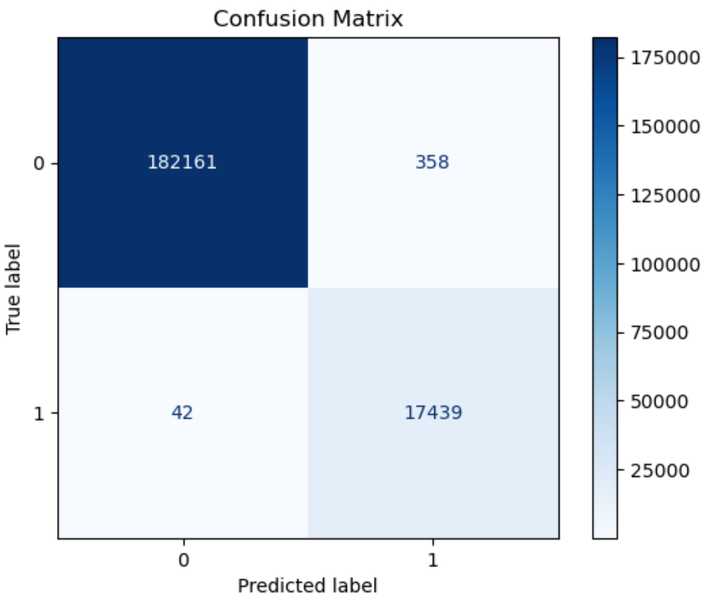

Figure 4: Confusion Matrix for XGBoost.

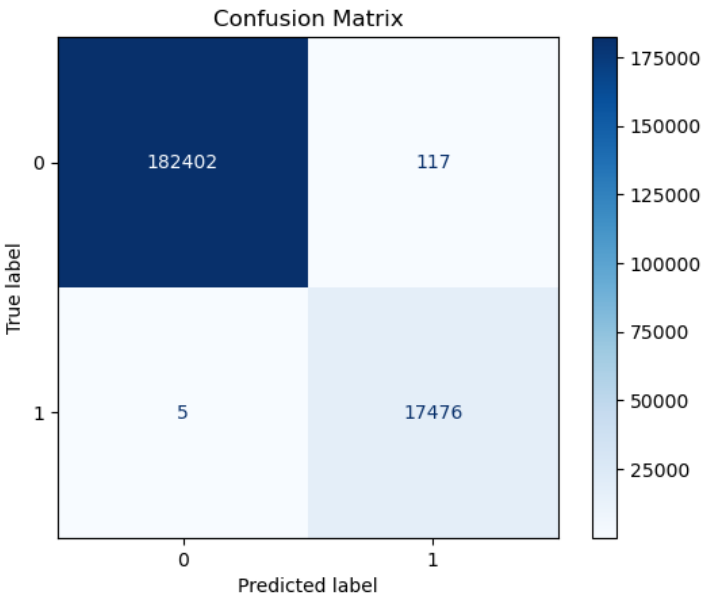

Figure 5: Confusion Matrix for Stacking Model.

4.3 Critical Analysis

Strengths:

- All models achieved exceptionally high F2-scores, with Random Forest and Stacking leading due to their superior balance between precision and recall.
- The Stacking model minimized false negatives while maintaining competitive false positive rates, making it particularly effective for fraud detection.
- Random Forest demonstrated simplicity and efficiency, achieving excellent results with minimal computational overhead.

Weaknesses:

- XGBoost exhibited higher false positive rates compared to Random Forest and Stacking, which could lead to increased manual review requirements in practice.
- The computational complexity of the Stacking model was significantly higher, which may limit its scalability for larger datasets or real-time applications.

4.4 Conclusion

The results highlight the strengths of ensemble methods in fraud detection. Random Forest emerged as the most balanced model, achieving the highest F2-score with minimal complexity. XGBoost demonstrated strong generalization but with slightly higher false positives. Finally, the Stacking model achieved the best overall performance, prioritizing recall to ensure minimal false negatives, making it the most suitable model for high-stakes fraud detection scenarios where missed fraudulent transactions have severe consequences.

5 Discussion and Conclusion

5.1 Summary

The second stage of this project achieved remarkable advancements in fraud detection compared to the baseline models from Stage 1. By introducing ensemble methods such as Random Forest, XGBoost, and Stacking, the limitations of Logistic Regression and KNN in handling non-linear relationships and large datasets were successfully addressed.

Key Highlights:

- **Random Forest:** Demonstrated exceptional performance with the highest F2-score of **0.99972**, a near-perfect balance between precision and recall, and minimal false positives and false negatives. Its simplicity and efficiency make it a strong candidate for deployment in real-world systems.

- **XGBoost:** Achieved an F2-score of **0.99918**, leveraging its ability to capture intricate data patterns through gradient boosting. However, its higher computational complexity and slightly elevated false positives suggest room for further optimization.

- **Stacking:** Outperformed individual models with an F2-score of **0.99980** and the highest recall (**0.99971**), effectively reducing false negatives to just 5 cases. By prioritizing recall in the hyperparameter tuning process, the stacking model aligned with the primary objective of fraud detection: minimizing missed fraudulent transactions.

The stacking approach, while computationally demanding, demonstrated the benefits of leveraging complementary models for enhanced prediction accuracy. This stage validated that ensemble learning methods are robust and scalable solutions for fraud detection in digital transactions.

Insights from Scoring Metrics: In this stage, the decision to maintain the F2-score as the guiding metric for Random Forest and XGBoost was consistent with Stage 1. However, for the stacking model, recall was prioritized during hyperparameter tuning. This shift reflects the critical nature of fraud detection, where minimizing false negatives is paramount to reduce potential financial and reputational risks.

Remaining Challenges: Despite these advancements, challenges remain:

- The stacking model's computational complexity requires careful consideration for scalability in production environments.

- XGBoost's higher false positive rate underscores the need for further refinement to optimize its performance.

- While Random Forest and Stacking achieved exceptional results, the potential for overfitting with synthetic data generated by SMOTE requires ongoing monitoring.

6 Discussion and Conclusion

6.1 Summary

The second stage of this project achieved remarkable advancements in fraud detection compared to the baseline models from Stage 1. By introducing ensemble methods such as Random Forest, XGBoost, and Stacking, the limitations of Logistic Regression and KNN in handling non-linear relationships and large datasets were successfully addressed.

Key Highlights:

- **Random Forest:** Demonstrated exceptional performance with the highest F2-score of **0.99972**, a near-perfect balance between precision and recall, and minimal false positives and false negatives. Its simplicity and efficiency make it a strong candidate for deployment in real-world systems.
- **XGBoost:** Achieved an F2-score of **0.99918**, leveraging its ability to capture intricate data patterns through gradient boosting. However, its higher computational complexity and slightly elevated false positives suggest room for further optimization.
- **Stacking:** Outperformed individual models with an F2-score of **0.99980** and the highest recall (**0.99971**), effectively reducing false negatives to just 5 cases. By prioritizing recall in the hyperparameter tuning process, the stacking model aligned with the primary objective of fraud detection: minimizing missed fraudulent transactions.

The stacking approach, while computationally demanding, demonstrated the benefits of leveraging complementary models for enhanced prediction accuracy. This stage validated that ensemble learning methods are robust and scalable solutions for fraud detection in digital transactions.

Insights from Scoring Metrics: In this stage, the decision to maintain the F2-score as the guiding metric for Random Forest and XGBoost was consistent with Stage 1. However, for the stacking model, recall was prioritized during hyperparameter tuning. This shift reflects the critical nature of fraud detection, where minimizing false negatives is paramount to reduce potential financial and reputational risks.

Remaining Challenges: Despite these advancements, challenges remain:

- The stacking model's computational complexity requires careful consideration for scalability in production environments.
- XGBoost's higher false positive rate underscores the need for further refinement to optimize its performance.
- While Random Forest and Stacking achieved exceptional results, the potential for overfitting with synthetic data generated by SMOTE requires ongoing monitoring.

6.2 Future Work: Stage 3 Preview

Building on the success of Stage 2, Stage 3 will focus on integrating advanced neural network architectures, specifically Multi-Layer Perceptrons (MLPs), to address the remaining challenges and further improve performance.

Key Objectives for Stage 3:

- **Capturing Complex Non-Linear Relationships:** Neural networks, with their ability to learn high-dimensional feature representations, will enable the modeling of more intricate patterns in transaction data.
- **Enhancing Precision and Recall:** By leveraging deep learning techniques, the project aims to surpass the precision and recall achieved by ensemble methods, particularly for edge cases in fraud detection.
- **Scalability Improvements:** Investigate lightweight neural network architectures and optimization techniques to ensure computational efficiency and scalability.
- **Automated Feature Engineering:** Neural networks' capability to learn features directly from raw data may reduce the need for manual preprocessing, streamlining the pipeline for future deployments.

Expected Benefits: The integration of neural networks is anticipated to unlock new possibilities, including:

- Greater flexibility in handling diverse datasets and transaction types.
- Reduced reliance on feature engineering and SMOTE-generated data by learning directly from imbalanced datasets.
- Improved robustness in fraud detection, even for rare or sophisticated fraud patterns.

Conclusion: For the next stage, we aim to achieve perfection by leveraging neural networks. This approach will allow us to explore more complex architectures, such as fully connected Multi-Layer Perceptrons (MLPs). By addressing the remaining limitations and incorporating deep learning techniques, Stage 3 promises to push the boundaries of performance and deliver a fraud detection system capable of handling the evolving landscape of digital transactions.