Fraud Detection Analysis

Brice Howe

DS 3001 Project

Terence Johnson

May 7, 2025

ABSTRACT

This project investigates the potential of machine learning techniques to detect fraudulent transactions using labeled, synthetic financial transaction data. The central question is whether transaction-level data can be used to distinguish fraudulent activity from legitimate transactions based on the characteristics of the transaction, user, merchant, timestamp, and device information. This issue is critical for financial institutions and payment platforms aiming to reduce fraud-related losses without compromising the user experience for legitimate customers. I conducted a supervised learning analysis, treating fraud detection as a binary classification problem. The dataset contained individual transaction records labeled as either fraudulent or non-fraudulent, encompassing both categorical and numeric features. The methodology followed a structured plan, beginning with logistic regression as a baseline model, and expanding to decision trees and random forests for enhanced predictive power and interpretability. After preprocessing and stratified sampling, I trained each model and evaluated performance using metrics that account for class imbalance, including F1 score, precision, recall, and ROC AUC. The results indicate that the random forest model achieved the best balance between false positives and false negatives, with high F1 scores and improved recall over simpler models. The study also addresses challenges such as imbalanced class distribution, feature redundancy, and the synthetic nature of the data. I explore these limitations as opportunities for further development and stress the importance of combining rigorous evaluation with domain-informed feature engineering in real-world fraud prevention systems.

INTRODUCTION

Digital payment systems have become a foundational element of modern commerce, offering convenience, speed, and scalability for consumers and merchants. As businesses increasingly adopt cashless infrastructures and consumers rely more heavily on card-not-present and mobile transactions, the volume and complexity of payment data have surged. However, this shift toward digitization has also created fertile ground for fraudulent activity. Fraudsters now exploit vulnerabilities in remote authentication, device spoofing, and social engineering schemes, making traditional rule-based detection systems increasingly inadequate.

The scale of the problem is gradually increasing. According to the 2025 Credit Card Fraud Report, “63% of U.S. credit card holders have been victimized by fraud, and 51% have experienced fraud multiple times” [1]. Globally, credit card fraud is estimated to cost businesses over $43 billion by 2026, with the number projected to rise each year[2]. As fraudsters adopt increasingly sophisticated and adaptive tactics, the need for robust, real-time fraud detection systems, especially those that can scale with large volumes of data and adapt to changing behaviors, has become urgent.

This project responds to that need by applying supervised machine learning techniques to a synthetic dataset of individual financial transactions, labeled as either fraudulent or legitimate. The central research question is: Can transaction data accurately distinguish fraudulent transactions from legitimate ones, and what patterns direct this distinction? Specifically, I focus on attributes commonly captured by digital payment systems, such as transaction amount, merchant category, card type, transaction location, device type, and authentication method. These

features are typically available in real-time and do not include sensitive personal identifiers, making them especially useful for building practical and privacy-preserving detection systems.

The task is framed as a binary classification problem, where each transaction is categorized as either fraud (1) or not fraud (0). The analysis begins with logistic regression to establish a baseline performance level due to its interpretability and transparency. I then implement more sophisticated models, decision trees and random forests, to capture complex, non-linear relationships and variable interactions that logistic regression may miss. Decision trees provide human-readable logic that can be valuable in operational settings, while random forests offer improved generalization by averaging over multiple decision paths and reducing overfitting.

A major challenge in fraud detection is class imbalance as fraudulent transactions comprise a minority of the dataset. This imbalance can skew models toward favoring the majority class, resulting in high accuracy but poor fraud detection. To address this, I use stratified sampling, resampling techniques, and evaluation metrics suited to imbalanced data. Rather than relying solely on accuracy, the F1 score is emphasized, which balances precision and recall, and I also report metrics such as ROC AUC, confusion matrices, and precision-recall curves. These tools help visualize and quantify trade-offs between false positives (flagging legitimate transactions) and false negatives (missing actual fraud).

By identifying patterns in transaction features that are most predictive of fraud, this project contributes to the broader effort to build adaptable, data-driven fraud prevention systems. By understanding the patterns that differentiate fraudulent transactions from legitimate ones, we can develop models that not only improve detection rates but also adapt to the ever-evolving landscape of financial fraud.

DATA

This study uses a publicly available synthetic dataset obtained from Kaggle, titled “Fraud Detection Transactions Dataset”. The dataset simulates real-world financial transactions and is designed to support the development and evaluation of fraud detection models. Each row in the dataset corresponds to a single transaction, labeled as either fraudulent (1) or legitimate (0), making it well-suited for binary classification tasks.

The dataset contains 50,000 transaction records and includes a wide range of features that reflect various attributes of a typical digital transaction. These features can be grouped into several categories:

- **Transaction Characteristics:** Includes Transaction_ID, Transaction_Amount, Transaction_Type, Account_Balance, Authentication_Method, Daily_Transaction_Count, Transaction_Distance, Risk_Score, Is_Weekend Avg_Transaction_Amount_7d, Failed_Transaction_Count_7d, and Timestamp
- **Merchant Information:** Includes Merchant_Category and Location
- **Customer Information:** Includes User_ID, Card_Type, Card_Age Authentication_Method, Previous_Fraudulent_Activity, and Device_Type

These features are largely anonymized and do not contain personally identifiable information, which makes the dataset appropriate for testing privacy-respecting machine learning workflows. All transactions are timestamped, and many features are categorical in nature, requiring appropriate encoding strategies.

Upon loading the dataset, I performed an initial inspection of the variable types. Numerical variables such as Amount and Card_Age were retained as continuous features, while categorical features such as Merchant_Category, Transaction_Type, and Authentication_Method were marked for one-hot encoding. The dataset exhibited no immediate signs of data corruption or formatting issues, and there was no missing data.

To better understand the distribution of key variables and identify potential outliers or irregularities, I then conducted an exploratory data analysis. Descriptive statistics revealed a high degree of skewness in the Transaction_Amount variable, an expected pattern in financial data due to only the occasional presence of large purchases. Although log transformation was considered, I ultimately relied on tree-based models, which are typically robust to skewed input distributions. The presence of a slight class imbalance was also confirmed as fraudulent transactions made up a smaller proportion of the total, at around 30% of all cases. This finding underscored the need to use evaluation metrics that account for imbalance, such as precision, recall, and F1 score, rather than overall accuracy alone.

Visual inspection of the data provided further insights. Histograms (Appendix A) revealed the distribution of all numeric variables in the dataset while count plots were generated to assess the prevalence of categorical features.

These insights guided the preprocessing strategy and allowed for the engineering of a feature matrix that captured relevant patterns without introducing unnecessary noise Specifically, I identified the most predictive and well-distributed features for modeling, and noted which variables required encoding, scaling, or additional engineering.

METHODS

The central methodological approach of this study is to treat fraud detection as a supervised machine learning problem. Specifically, this task is framed as a binary classification problem in which each financial transaction is labeled as either fraudulent (1) or legitimate (0). The objective is to train classification models capable of identifying patterns in transaction features that accurately distinguish between these two classes. This approach mirrors real-world use cases in fraud detection systems where incoming transactions must be classified in real time based on non-personal behavioral features.

To model this problem, I implement and compare three machine learning algorithms: logistic regression, decision trees, and random forests. Logistic regression serves as a benchmark model due to its simplicity and interpretability. It offers a probabilistic framework for predicting binary outcomes and allows for straightforward evaluation of variable coefficients and significance. However, it assumes linear relationships between predictors and the log-odds of the outcome, which may be insufficient to capture the complexity of fraudulent behavior.

To account for potential non-linearity and feature interactions, I incorporate a decision tree classifier. Decision trees are powerful for their ability to split on feature thresholds and provide human-readable rules. However, they are prone to overfitting, especially in small or noisy datasets. To address this, random forest is used, an ensemble method that aggregates the predictions of many decision trees trained on bootstrapped samples. Random forests improve predictive stability and reduce variance, making them well-suited for high-dimensional and noisy data environments like fraud detection.

The dataset was randomly split into training and testing subsets using an 80/20 train-test split, ensuring that both subsets retained the original class proportions through stratified sampling. This stratification is critical given the severe imbalance between fraudulent and legitimate transactions; without it, the test set could contain too few fraudulent examples to effectively evaluate performance. The training set was used to fit all models and optimize hyperparameters through cross-validation, while the test set was strictly held out for final model evaluation.

Success in this context was measured primarily by the F1 score, which balances precision and recall. This is particularly important in fraud detection where both false positives (flagging legitimate transactions as fraud) and false negatives (missing actual fraud) carry serious costs. While high precision limits unnecessary friction for customers, high recall ensures that actual fraudulent activities are caught. In addition to F1 score, I also report precision, recall, accuracy, and ROC AUC to provide a well-rounded evaluation. Confusion matrices, ROC curves, and precision-recall plots are used to visualize trade-offs between different kinds of prediction errors.

Several challenges were anticipated in model development. First, the class imbalance poses a risk of biased models that default to the majority class. To address this, I explored class weighting and oversampling techniques to ensure minority class representation. Second, as the dataset is synthetic, it may contain patterns that do not generalize well to real-world fraud scenarios. To mitigate this, I emphasized robust validation strategies and limited model complexity to reduce overfitting. Third, feature redundancy or irrelevant variables could introduce noise, so I monitored correlation among features and considered regularization or feature selection methods to improve model robustness.

Feature engineering played a critical role in model preparation. Categorical variables were one-hot encoded to enable their use in standard classifiers. Numerical features were assessed for skewness and considered for scaling where appropriate.

Finally, model performance and insights were presented through both quantitative summaries and visual aids. Confusion matrices and metric tables provide detailed performance breakdowns, while feature importance visualizations from tree-based models highlight the most influential predictors. A comparison table was used to synthesize results across models, emphasizing where each algorithm excels or underperforms.

RESULTS

Among the three classifiers, the random forest model achieved the best overall performance. All models were assessed using F1 score, precision, recall, accuracy, and ROC AUC, with a particular emphasis on the F1 score due to its relevance in imbalanced binary classification. Logistic regression, while interpretable, struggled to capture nonlinear interactions and produced lower recall. The decision tree model improved slightly but remained prone to overfitting. In contrast, the random forest provided a strong balance between false positives and false negatives, demonstrating more robust generalization.

The table found in Appendix D summarizes each model's performance on the hold-out test set across four key metrics. These results confirm that the Random Forest classifier outperformed both logistic regression and decision tree models in all four categories. It achieved the highest F1 score (0.867), indicating a strong balance between precision and recall, and significantly improved overall accuracy compared to the baseline. The improvements in both recall and precision highlight the model's ability to detect a larger proportion of fraudulent transactions

while keeping false positives low. In contrast, while logistic regression offered interpretability, it lagged behind in both recall and overall predictive power. The decision tree model performed slightly better, but still fell short of the ensemble-based robustness of the Random Forest model.

The confusion matrix for the random forest classifier on the 12,500-transaction test set showed 8,483 true negatives, 2,478 true positives, 1,539 false negatives, and 0 false positives (Appendix E). This outcome suggests that the model is highly conservative in issuing fraud flags. While it avoids disrupting legitimate customer experiences (no false positives), it does so at the cost of missing a substantial number of actual fraud cases, a trade-off that reflects a preference for precision over recall.

The precision-recall curve offered further insight into this trade-off. The curve started with near-perfect precision (~1.0) at low recall levels and gradually declined as recall increased (Appendix F). The average precision (AP) was 0.80, a strong result indicating the model's ability to identify fraud while minimizing false alarms. However, the curve's shape revealed the classic tension in fraud detection: achieving higher recall comes at the expense of a sharp drop in precision, often yielding diminishing returns.

The Receiver Operating Characteristic (ROC) curve also confirmed the model's overall strength, with an area under the curve (AUC) of 0.81 (Appendix G). The steep initial rise indicated the model could correctly identify a large proportion of fraudulent transactions with relatively few false positives. However, the middle section of the curve flattened, reflecting reduced marginal gain in fraud identification as the false positive rate increases. This points to a ceiling in model sensitivity, even for a high-performing classifier like random forest.

A deep dive into the feature importances generated by the random forest model yielded critical insights into which variables most influenced classification outcomes (Appendix H). The most dominant feature was Failed_Transaction_Count_7d, accounting for over 35% of the model's decision-making weight. This result suggests that a spike in failed transaction attempts in the preceding week is a strong behavioral signal for fraud.

Other influential features included Transaction_Distance, Avg_Transaction_Amount_7d, and Account_Balance, each contributing roughly 1% of total model importance. These features capture contextual and behavioral anomalies, such as sudden geographic shifts in spending, unusual spending volume in a short period, or low balances that precede account compromise.

In contrast, most categorical variables, such as Card_Type, Device_Type, Authentication_Method, and Merchant_Category, contributed less than 0.2% individually to the model's decisions. This disparity suggests that while categorical descriptors may provide marginal signal, dynamic, temporal, and quantitative behavior patterns carry far more predictive power in fraud detection contexts.

The results, overall, demonstrate that random forests offer a practical balance between interpretability and predictive power in fraud detection tasks. The model was especially adept at identifying high-confidence fraud cases with low false alarm rates. However, its conservative nature came at the cost of lower recall, highlighting a central dilemma in financial fraud detection systems: optimizing for user experience vs. comprehensive risk identification. These findings lay the foundation for further exploration into cost-sensitive learning and threshold adjustment in future work.

CONCLUSION

This study set out to examine whether machine learning models could effectively detect fraudulent transactions using structured, transaction-level data. The motivation is rooted in a rapidly evolving financial landscape where digital payments are the norm, and fraudsters constantly refine their tactics to exploit vulnerabilities. Traditional rule-based systems have shown limitations in scalability and adaptability, which opens the door for data-driven approaches that learn patterns dynamically and improve over time. Through this project, I sought to evaluate the practical viability and comparative strengths of three commonly used classification models, logistic regression, decision trees, and random forests, on a labeled dataset of financial transactions.

The findings demonstrate that machine learning models can indeed provide meaningful predictive performance in the context of fraud detection, especially when supported by careful feature engineering, thoughtful handling of class imbalance, and appropriate evaluation strategies. Among the models tested, the random forest classifier emerged as the most effective. It achieved superior scores across all metrics, F1 score, precision, recall, and accuracy, offering a strong balance between identifying fraudulent transactions and minimizing disruptions to legitimate customers.

A deeper analysis of feature importance underscored the value of behavioral transaction data. Variables such as Failed_Transaction_Count_7d, Transaction_Distance, and Avg_Transaction_Amount_7d stood out as the most predictive, suggesting that fraud is often detectable through patterns of behavior rather than through fixed characteristics such as merchant category or device type. This has important implications for fraud prevention strategies. Rather

than building systems that rely heavily on static rules or categorical identifiers, institutions may benefit from tracking dynamic indicators that reflect changes in behavior over time.

However, despite its strengths, the random forest model exhibited a conservative bias. It produced no false positives, meaning it never incorrectly flagged a legitimate transaction as fraud, but this came at the cost of 1,539 false negatives, fraud cases that went undetected. This trade-off reflects a broader dilemma in fraud detection, how to balance the competing risks of customer dissatisfaction and financial loss. Depending on institutional priorities, organizations may opt to tune the model threshold to allow for more aggressive fraud detection, even if it means accepting a higher false positive rate. Future work could also explore cost-sensitive learning, which explicitly incorporates the relative costs of false positives and false negatives into model training.

The study faced several limitations that should be acknowledged. First and foremost, the dataset was synthetic. While this allowed for controlled experimentation and simplified access, it may not fully capture the complexity and nuance of real-world financial fraud. In a real deployment, data drift, adversarial behavior, and more complex fraud tactics could challenge the model's effectiveness. Second, while categorical and numeric features were effectively handled through one-hot encoding and basic scaling, more sophisticated representation learning may improve performance on high-cardinality variables or non-linear interactions. Third, even though stratified sampling helped preserve class balance during training and testing, class imbalance remains a persistent obstacle in fraud detection systems, requiring ongoing experimentation with sampling, reweighting, and anomaly detection techniques.

Looking ahead, several promising avenues exist for extending this work. One direction is the integration of temporal modeling, such as using time series-aware architectures to capture the evolution of user behavior over time. Another path is exploring real-time detection pipelines, where inference must be both fast and explainable in production environments. Incorporating external data sources, such as geolocation intelligence, user device fingerprints, or prior fraud history, may also enhance model performance and contextual awareness. Finally, further research could investigate the ethical and operational implications of deploying automated fraud detection systems, particularly in relation to fairness, transparency, and the potential for biased outcomes.

This project provides strong evidence that machine learning can be a valuable tool in combating financial fraud. While no model is flawless, the random forest approach showcased here offers a practical, scalable solution that balances precision with recall. By focusing on behavioral indicators and embracing robust modeling strategies, this work contributes to the growing field of algorithmic fraud detection and offers a roadmap for future innovations that are not only technically sound but also operationally and ethically responsible.

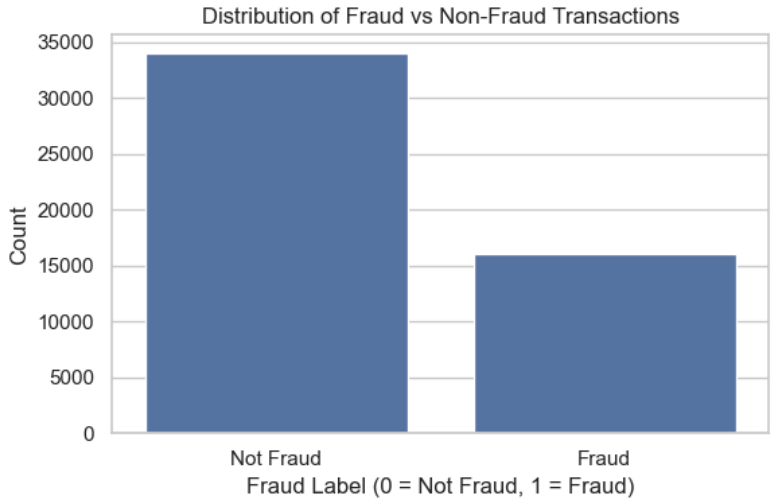
REFERENCES

[1] Cruz, Brett. “62 Million Americans Experienced Credit Card Fraud Last Year.” *Security.Org*, 27 Jan. 2025, www.security.org/digital-safety/credit-card-fraud-report/.

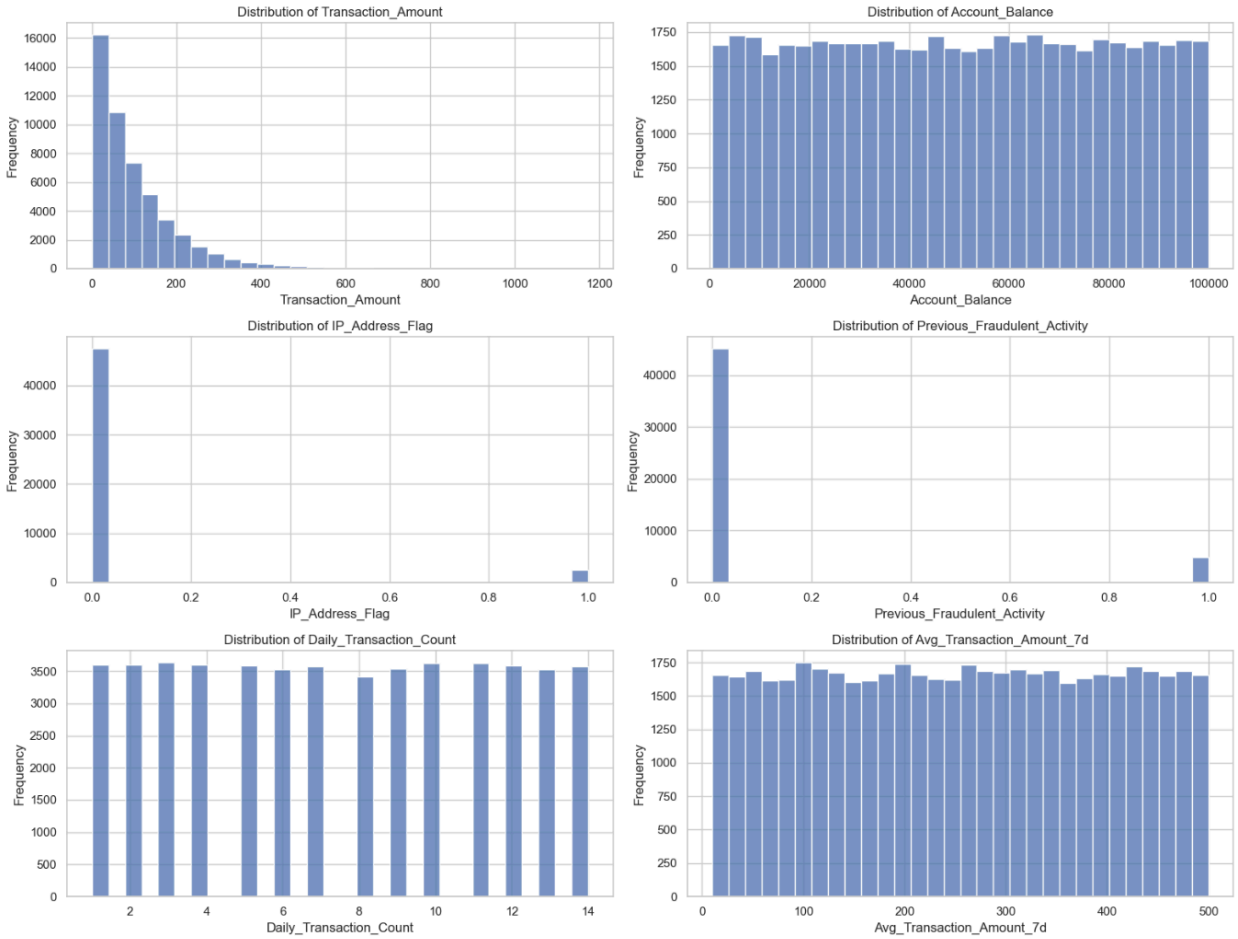
[2] Rej, Matt. “Credit Card Fraud Statistics (2025).” *Merchant Cost Consulting*, 9 Dec. 2024, merchantcostconsulting.com/lower-credit-card-processing-fees/credit-card-fraud-statistics/.

Data Source:
https://www.kaggle.com/datasets/samayashar/fraud-detection-transactions-dataset/data

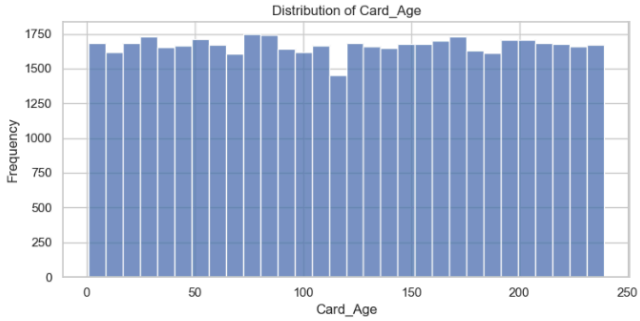
Appendix A.

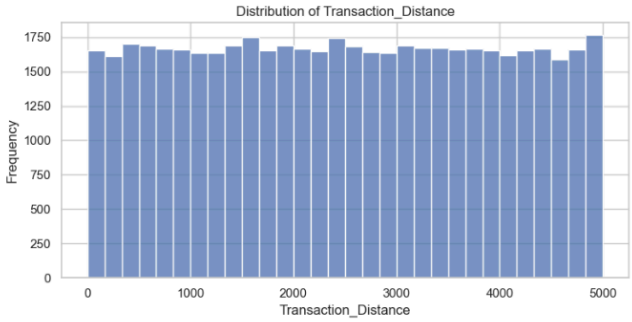


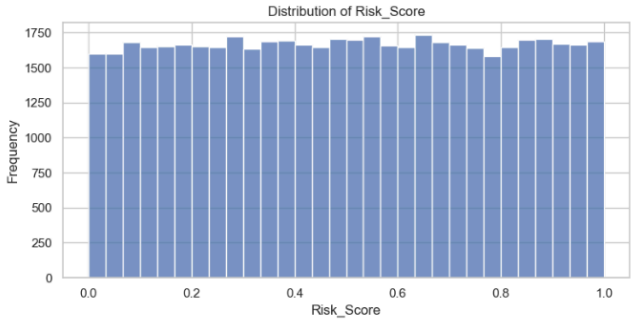
Appendix B.

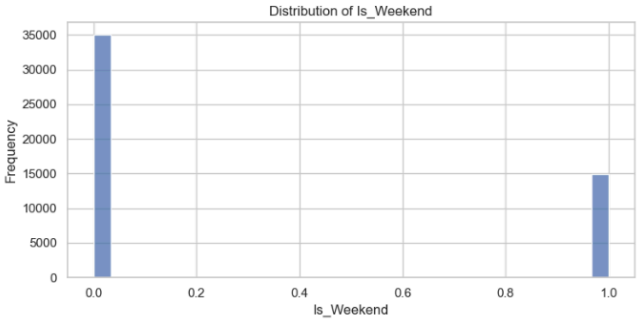



Distribution of Failed_Transaction_Count_7d
Frequency
Failed_Transaction_Count_7d

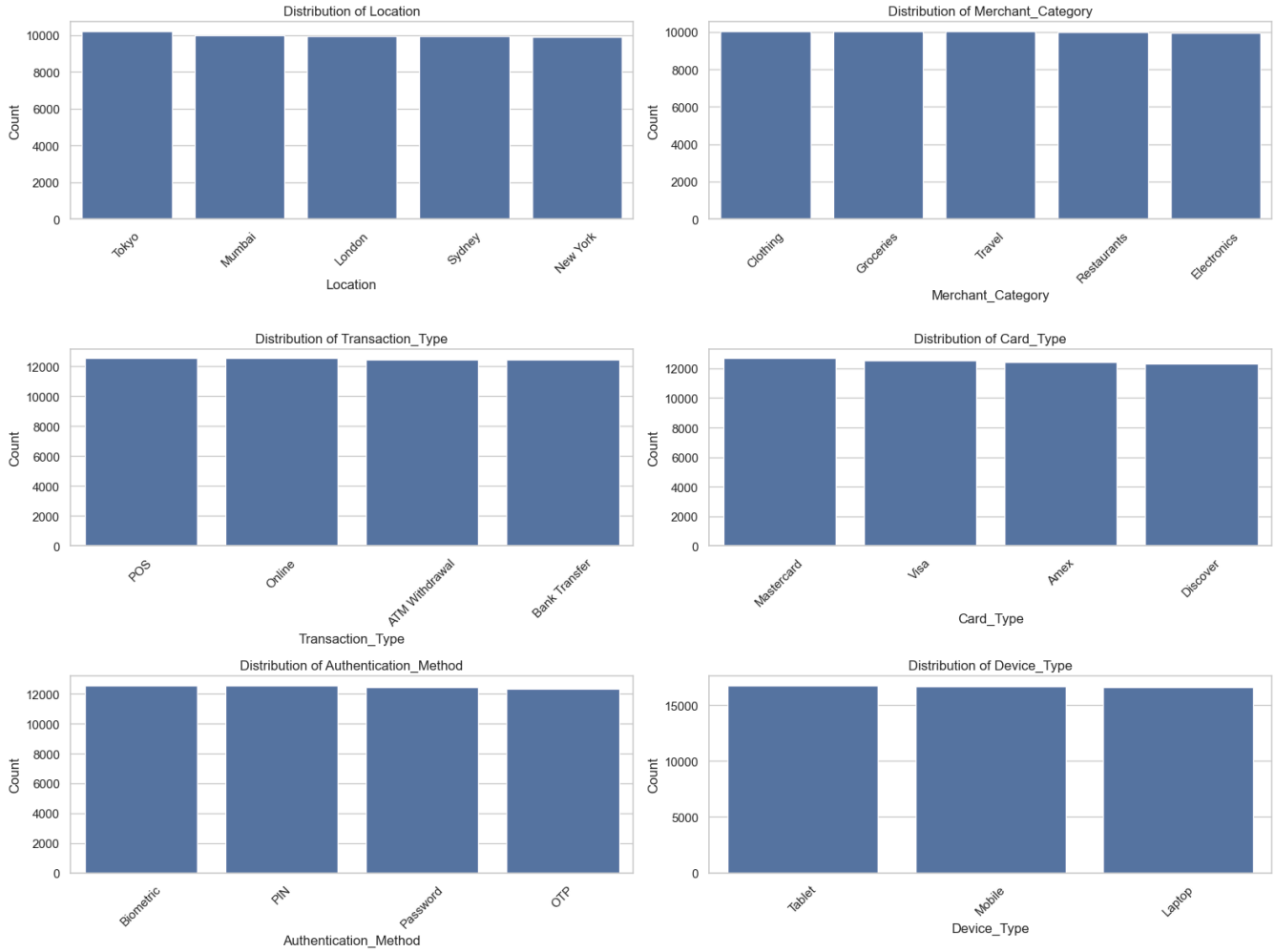

Distribution of Card_Age
Frequency
Card_Age


Distribution of Transaction_Distance
Frequency
Transaction_Distance


Distribution of Risk_Score
Frequency
Risk_Score


Distribution of Is_Weekend
Frequency
Is_Weekend

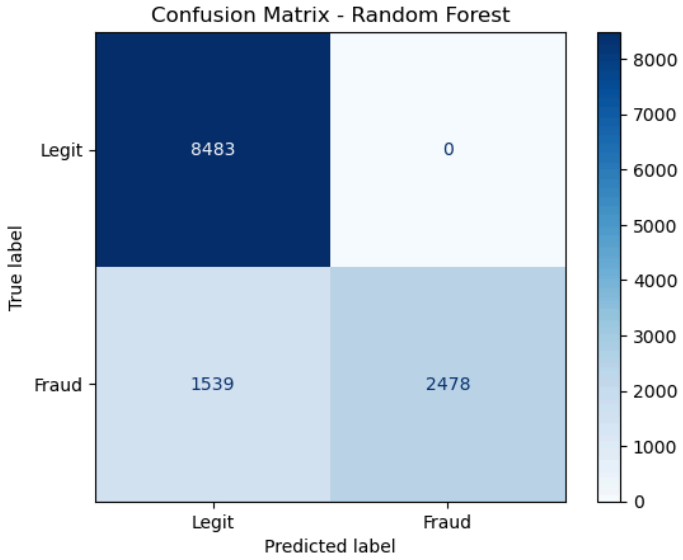
Appendix C.



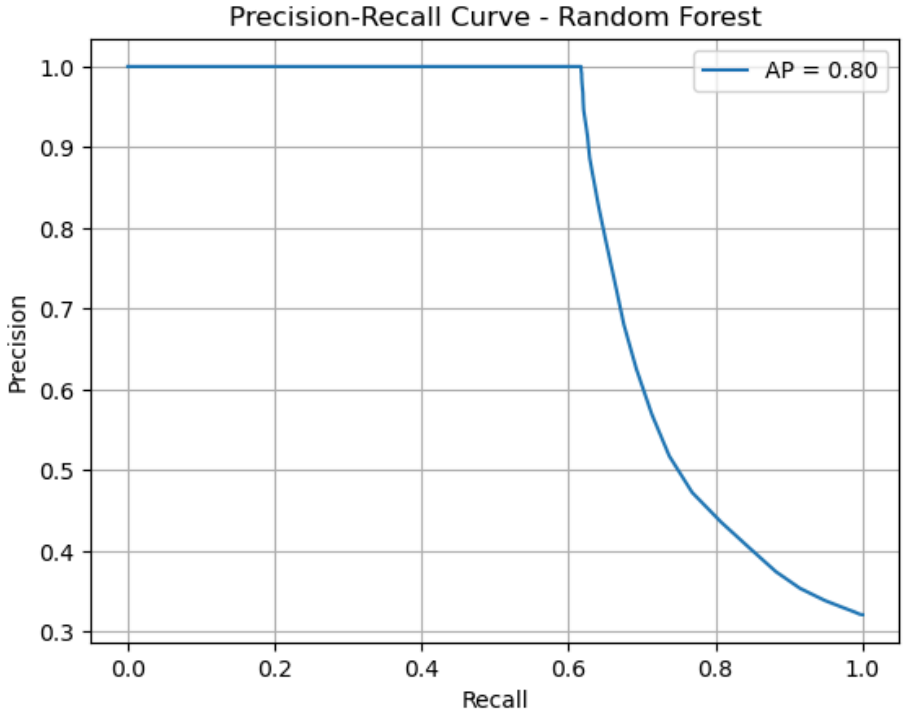
Appendix D.

	F1 Score	Precision	Recall	Accuracy
Logistic Regression	0.738713	0.754594	0.73224	0.73224
Decision Tree	0.802992	0.801932	0.80496	0.80496
Random Forest	0.867413	0.895787	0.87688	0.87688

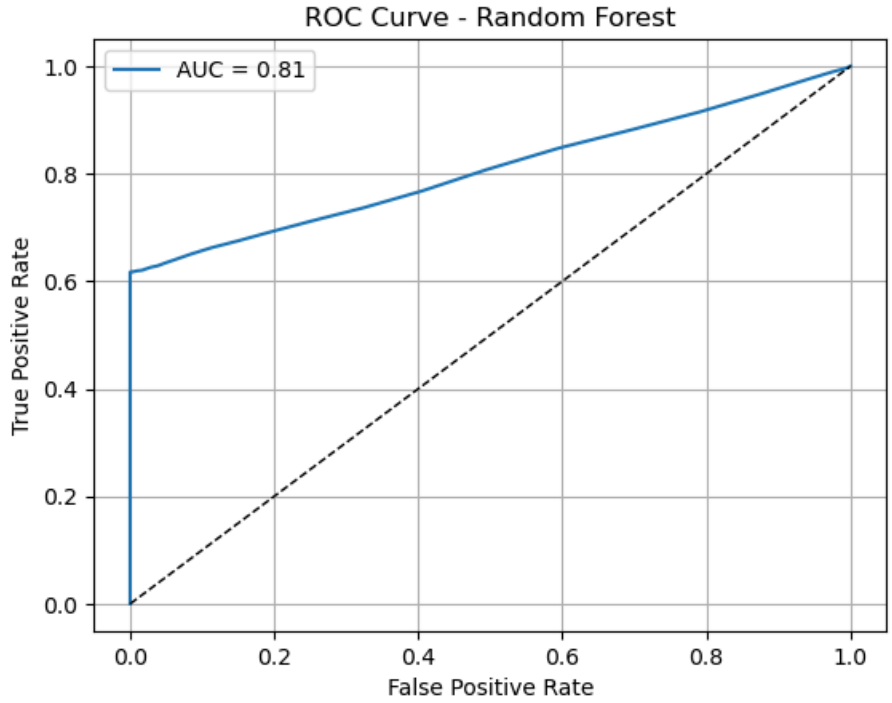
Appendix E.



Appendix F.



Appendix G.



Appendix H.

	Feature	Importance
6	num__Failed_Transaction_Count_7d	0.351892
8	num__Transaction_Distance	0.011586
5	num__Avg_Transaction_Amount_7d	0.011554
1	num__Account_Balance	0.011396
0	num__Transaction_Amount	0.011336
7	num__Card_Age	0.010696
4	num__Daily_Transaction_Count	0.007744
9	num__Is_Weekend	0.002336
82545	cat__Device_Type_Mobile	0.002010
82544	cat__Device_Type_Laptop	0.001983
82564	cat__Authentication_Method_Password	0.001959
82557	cat__Card_Type_Amex	0.001958
82563	cat__Authentication_Method_PIN	0.001927
82546	cat__Device_Type_Tablet	0.001910
82559	cat__Card_Type_Mastercard	0.001896
82560	cat__Card_Type_Visa	0.001882
82561	cat__Authentication_Method_Biometric	0.001879
82551	cat__Location_Tokyo	0.001865
46362	cat__Transaction_Type_Bank Transfer	0.001865
82556	cat__Merchant_Category_Travel	0.001857