Report: How do transactional and behavioral factors affect the likelihood of a person being a fraud victim?

Group 3: Jiarong Guo, Suchakrey Nitisanon, Max Wong, Lillian Xu

APAN 5205: Applied Analytics Frameworks and Methods II

Instructor: Dr. Birol Emir

April 20th, 2025

Research Problem

With the rapid evolution of technology in this generation, transactions between businesses and customers have become more convenient. The digitization of financial services allows customers to use online and mobile payment systems. Digital banking and e-commerce platforms change the traditional transactional methods, allowing customers to transfer payments anytime and anywhere.

As people rely more on digital systems, cyber threats tend to become more vulnerable. Fraud tactics change rapidly, affecting millions and billions of people globally. There is a need to adopt better strategies to identify fraudulent transactions. Machine learning models and techniques can learn complex patterns and observe potential relationships and correlations between features. With the help of advanced data-analyzation techniques, developing a fraud detection system with better accuracy helps companies minimize financial losses and strengthen trust with customers.

Literature review

Financial fraud has emerged as a significant concern for businesses and consumers. The *2024 Global eCommerce Payments & Fraud Report* highlights that merchants worldwide are facing a wider variety of fraud attacks than ever before. The most common types include first-party misuse, account takeovers, loyalty fraud, and triangulation schemes, which all experienced an increase in the past year. The report shows that fraud affected merchants' annual eCommerce revenue. Merchants lost approximately 3% of their revenue (Global Fraud Report, 2024). Businesses are using artificial intelligence techniques for fraud detection to reduce the risk of fraud. They leverage machine learning models to improve the accuracy of identifying fraudulent transactions.

Chang et al. (2022) studied the performance of different machine learning algorithms to detect fraud in digital transactions. This study compared five algorithms, including logistic regression, decision tree, k-nearest neighbors, random forest, and autoencoder. The results showed that random forest and logistic regression achieved the highest accuracy, with almost 99% accuracy. However, unsupervised learning models like autoencoders performed well in detecting anomalies without requiring labeled fraud cases. One limitation of this study is that machine learning models may have trouble keeping up with fast changes in fraud patterns. The fraud detection models need to be updated regularly in order to stay effective.

In addition, Awoyemi et al. (2017) examined the imbalanced data of credit card fraud using the naïve bayes, k-nearest neighbors, and logistic regression models. The results showed that k-nearest neighbors outperformed naïve bayes and logistic regression. It achieved an accuracy of 97.69% after combining both oversampling and undersampling techniques. However, the study showed a limitation where logistic regression displayed poor performance, with a 54.86% accuracy, when dealing with extreme data imbalance. While these studies lay out a strong foundation for machine learning-based fraud detection, they also highlight areas for further research. Current studies on fraud detection have limitations. Fraud detection models are trained based on historical data. Relationships between fraudulent transactions and contextual variables are neglected. Further research is needed to investigate how certain contextual variables, such as geographical and temporal factors, impact fraud probability. In addition, future studies should explore more efficient and scalable fraud detection models to ensure that high-performance fraud detection is accessible across various business scales.

Research Questions

This study investigates how different features in the dataset influence credit card fraud. By analyzing these factors, we aim to identify fraud patterns and improve detection methods.

How do transactional and behavioral factors affect the likelihood of a person being a fraud victim?

RQ1: Would certain types of merchants have a higher possibility of fraudulent transactions?

- H1: Some types of merchants are more likely to have fraudulent transactions than others.
- H0: The type of merchant does not affect the likelihood of a transaction being fraudulent.
- X: Merchant type (category).
- Y: Fraud occurrence (is_fraud).

RQ2: Would merchants in more populated cities be more likely to be fraudulent compared to those in less populated cities?

- H1: Merchants in cities with larger populations are more likely to have fraudulent transactions than those in smaller cities
- H0: The population size of a city does not affect how likely a merchant is to have fraudulent transactions
- X: Number of population (city_pop).
- Y: Fraud occurrence (is_fraud).

RQ3: Would the transaction hour of the day affect the probability of fraud in digital transactions?

- H1: Certain transaction hours of the day are more likely to have fraudulent transactions than others.
- H0: The transaction hour of the day does not affect the likelihood of a transaction being fraudulent.
- X: Transaction time (transaction_hour)
- Y: Fraud occurrence (is_fraud).

Data Description

The dataset contains simulated credit card transactions that happened between the years of 2019 and 2020. It has 693 different merchants and 999 credit card holders. The dataset consists of 23 columns and more than 1.8 million rows. Features in the dataset include transaction date and time, credit card number, merchant, type of merchant, transaction amount, transaction number, name of cardholder (first and last), age and gender of cardholder, billing address (street, city, state, zip, latitude, longitude), payment address, city population, cardholder occupation, cardholder date of birth, unix time, merchant latitude and longitude, and indicator of transaction legitimacy.

Due to confidentiality restrictions, companies typically do not release original feature names in real-world datasets. As a result, public credit card fraud datasets often anonymize variables. While this limits access to specific transactional details, the dataset still reflects realistic patterns and behaviors, making it highly practical for machine learning applications.

The dataset provides a reliable foundation for developing and testing predictive models. It supports the investigation of key factors this project aims to study: merchant type, location, and transaction hour of the day (see Figure 1-3). The dataset is well-suited for training fraud detection systems and evaluating model performance in a controlled but realistic environment.

Although the data is simulated, the insights and model framework developed through this research have real-world applications. Companies can adapt this model with necessary modification to identify fraud risk in their own transaction data and suspicious patterns in real time. This helps enhance fraud detection systems, reduce false positives, and prioritize high-risk cases for further review.

Data Preparation and Analytical Techniques

To prepare the dataset, we first combined the two raw datasets using R. We removed columns that were not useful for the analysis and observed missing data in the dataset. The missing data followed an arbitrary pattern and were scattered across several features, such as age and transaction_day. We impute missing data using the mice package. We verified and dropped duplicated records, keeping only unique transactions in our final dataset. Additionally, we adjusted data types carefully to ensure data consistency. For example, dates were converted from character to date format, and categorical variables were set as factors. For the merchant variable, we removed the prefix "fraud_" to clean the data. For data manipulation, we created a new column that indicates the cardholder's age at the time of each transaction and adjusted the zip code format by adding missing leading zeros. Lastly, we noted that the dataset was highly imbalanced, with fraudulent transactions comprising approximately 0.5% of all transactions. This imbalance was taken into account during model training and evaluation by selecting appropriate metrics and applying techniques to ensure fair performance assessment.

As the first step before predictive modeling, clustering analysis was performed to identify natural groupings or patterns of the credit card transactions without using "is_fraud". Specifically, we utilized k-means and model-based clustering to complete our clustering analysis. For k-means clustering, we repetitively tested from a two-cluster solution to an eight-cluster solution in all eight scenarios to identify the best grouping structures and assess how fraud patterns changed in each scenario and segmentation. For each scenario, clusters were created as a percentage of transactions in that cluster, with 0 representing non-fraudulent transactions and 1 representing fraudulent transactions. This is a suitable technique for our research because it captures underlying transactional structures that may not be visible through raw features alone,

making our models more reliable and insightful. Besides k-means clustering, model-based further validated the segmentation by estimating clusters and calculating the likelihood of each transaction belonging to a given cluster in a probabilistic manner. As a type of unsupervised learning, model-based clustering does not assume that data is divided equally, making it effective for our research because it mimics the real-world and complex financial transactions. In the no clustering scenario, we trained models directly on the original dataset using standard predictive modeling techniques without segmentation, serving as a comparison benchmark. In the no clustering scenario, we trained models directly on the original dataset using standard predictive modeling techniques without segmentation, serving as a benchmark. Overall, clustering helped us understand and uncover structures and features regarding fraudulent transactions in our dataset.

After we understood the patterns and groupings of the transactions, we applied supervised predictive models, including Logistic Regression, Random Forest, Neural Network, and Deep Learning models to predict fraudulent transactions. We started off with Logistic Regression as a baseline model, which offered coefficients and predictive power as the results. Then, we used Random Forest in all scenarios by leveraging an ensemble of decision trees, giving us better accuracy and clear insights into different variables and their impact on fraudulent transactions. To further our analysis, we built Neural Network models consisting of one hidden layer of 32 neurons to capture nonlinear relationships. Finally, our choice of Deep Learning model with two hidden layers, 32 neurons for the first layer and 16 neurons for the second layer, was able to deliver the highest accuracy. These models were evaluated using performance metrics such as accuracy, sensitivity, precision, recall, AUC, and confusion matrix. In addition,

we chose these models to balance between understanding the results and making accurate predictions, which made them a great fit for the complexity and size of our dataset.

Results

This study was designed to support fraud detection strategy by testing how different features (merchant category, city population, and transaction hour) affect fraud prediction accuracy. We built 9 scenarios using a combination of Clustering (K-means or Model-based) and Forecasting models (Logistic Regression, Random Forest, Neural Network, and Deep Learning). The first 6 scenarios were used to test each research question by excluding some variables. The last 3 scenarios used all variables to see if this gives better results.

We applied both K-means and Model-based clustering (2-8 clusters) across all scenarios to group transactions based on similar behaviors. After clustering, we calculated the percentage of fraud in each cluster to see if clustering could separate high and low risk groups. The results are shown in Table 1-8 in the Appendix.

In most scenarios, fraud rates varied clearly across clusters. Notably, in Scenario 3 (using populated city, city_pop) and Scenario 5 (using transaction hour, transaction_hour), K-means with 8 clusters shows one high-risk cluster with a fraud rate over 35%, while other clusters have a fraud rate below 0.5%. This shows that even when the type of merchant (category) was excluded, population or time-based patterns alone were strong enough to isolate the fraud-prone groups.

Each model was evaluated using the five metrics: accuracy (overall prediction performance), sensitivity (how well fraud cases were detected), precision (how many alerts were

detected), AUC (model's ability to separate fraud from non-fraud), and confusion matrix (true/false positives and negatives).

For this particular business use case, sensitivity is the most important. It helps reduce missed fraud cases and supports better decision-making. However, low precision can cause many false alarms, which is also a problem. We need to balance sensitivity and precision.

For Table 9 in the Appendix, the results show that clustering before modeling helps improve performance. Scenario 9 used all features with no clustering, but its precision (0.1259) and sensitivity (0.9820) are lower than Scenarios 7 (with k-means clustering) and 8 (with model-based clustering), which achieved 0.9805 sensitivity and 0.1837 precision, and 0.9830 sensitivity and 0.1734 precision, respectively.

Another key insight is that Random Forest and Deep Learning outperformed others in most scenarios. These two models appear most frequently as top-performing models with strong sensitivity and precision compared to Linear Regression and Simple Neural Network.

The best overall result came from scenario 8. It included all features and used Model-Based Clustering along with the Random Forest model. This combination gave the highest sensitivity (0.9830) and highest precision (0.1734), making it the most suitable model for business use and real-time fraud detection.

Conclusion and Limitations

The purpose of this project is to identify the most effective method to detect fraudulent transactions with the overarching goal of minimizing financial losses and protecting customers. In this context, it is important to have high sensitivity and a low false negative rate. A low precision rate is acceptable as a trade-off. Therefore, fraudulent transactions should not be

missed. Results show that random forest models can achieve maximum sensitivity. Clustering improves detecting fraudulent transactions. Model-based clustering yields better results than k-means clustering (Table 9 in the Appendix). From RQ1 regarding merchant type, we learned that merchant type is useful, especially with Random Forest, but not strong enough alone. Adding behavioral or location based features helps reduce financial risk. As for RQ2 related to populated cities, city population alone shows some signal, but not enough to prevent high-value frauds. Precision needs improvement to reduce business loss. For RQ3 about the likelihood of fraud, transaction hours, a time-based pattern, help with sensitivity but are not reliable alone for business use. High-value frauds still slip through. Lastly, we examined an extended question where all features mentioned above happen together. Using all features significantly improves fraud detection. Clustering adds value (scenarios 7 and 8), but scenario 9 shows that it is a strong model without clustering before performing prediction.

Based on the results, the configuration and model from scenario 8 best fit the project's overall goal (Figure 4 in the Appendix). This scenario has the lowest overall loss of $1638 compared to other scenarios (Table 9 in the Appendix). It detects the most fraudulent transactions and minimizes the cost of loss. We did not perform hyperparameter tuning for any of the machine learning models due to computational power constraints. All models were run using their default settings. As a result, the performance metrics may not reflect the best possible results that these models can achieve.

To effectively minimize financial loss and build customer trust, we recommend implementing a fraud detection approach that combines high-sensitivity models such as Random Forest models with two-factor authentication (2FA) for credit card transactions with high risk. While Random Forest models significantly reduce the chance of missing fraud, our analysis

shows that 2FA can immediately block unauthorized access as a safeguard. Moreover, integrating behavioral and location-based features, including transaction hour and merchant type, into our model can further enhance detection accuracy and improve impact on business use cases. While some demographic variables like city population show limited value, their interactions with other features may still provide support for fraud identification. Going forward, decision-makers should invest in continuous model updates and feature engineering to adapt to the fast-changing environment of fraud detection and provide further support for the customers.

References

Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (n.d.). *Credit card fraud detection using Machine Learning Techniques: A Comparative Analysis | IEEE conference publication | IEEE xplore*. IEEE. https://ieeexplore.ieee.org/document/8123782

Chang, V., Doan, L. M. T., Stefano, A. D., Sun, Z. L., & Fortino, G. (2022, May). *Digital Payment Fraud Detection Methods in digital ages and industry 4.0*. Computers and Electrical Engineering. https://www.sciencedirect.com/science/article/abs/pii/S0045790622000465

Visa Acceptance Solutions (n.d.). *2024 Global Fraud and Payments Report*. https://www.visaacceptance.com/en-us/insights/fraud-report.html?gad_source=1&gclid=EAIaIQobChMIn5mChKTuiwMVCjAIBR01vjt3EAAYAiAAEgILO_D_BwE

Shenoy, K. (2020, August 5). *Credit Card Transactions Fraud Detection Dataset*. Kaggle. https://www.kaggle.com/datasets/kartik2112/fraud-detection

Appendix

Figure 1. Fraudulent by Merchant Type (Category)

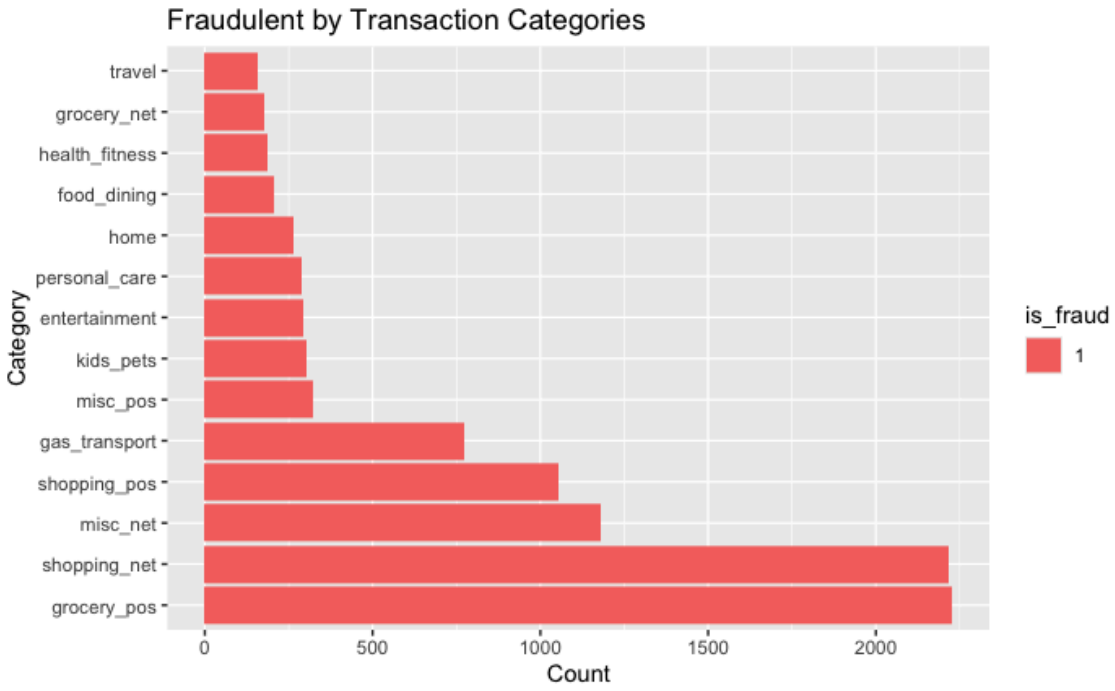

Figure 2. Fraudulent by Hour

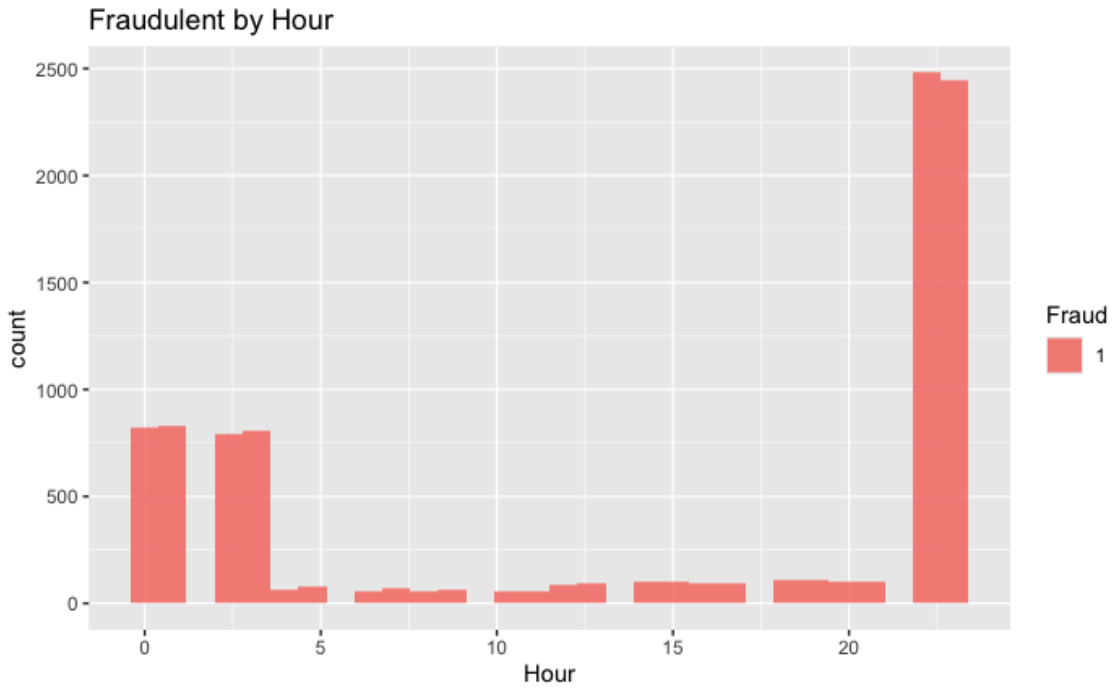

Figure 3. Fraudulent by City Population

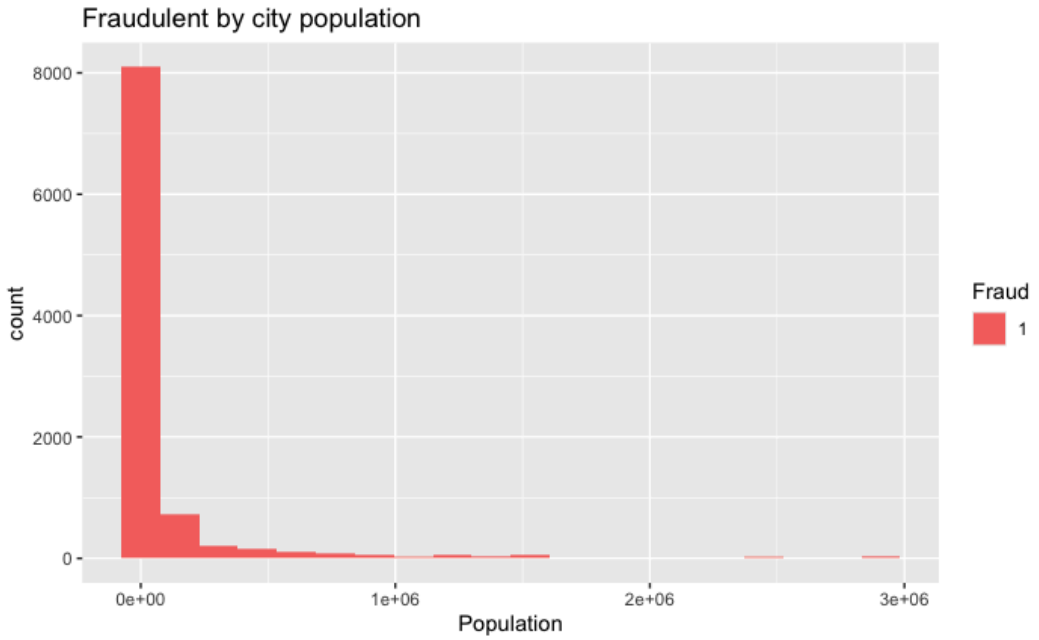

Figure 4. Sensitivity of Scenario 1-9

There is no clustering model for Scenario 9.

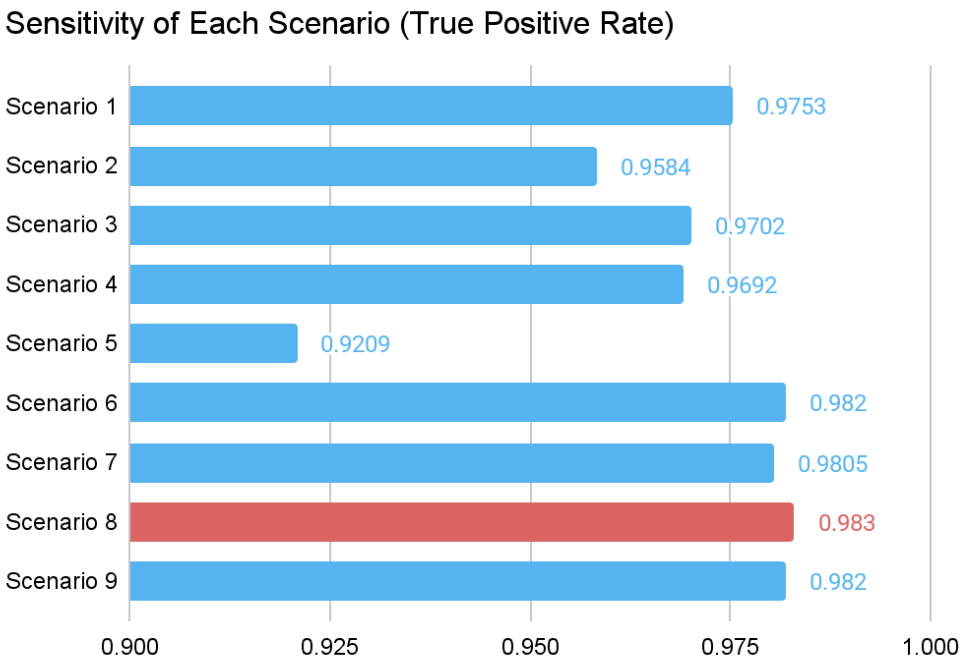

Table 1. Scenario 1 Clustering Results

Cluster	%Not Fraud	%Fraud
1	99.57	0.43
2	99.61	0.39
3	98.70	1.30
4	99.54	0.46
5	99.47	0.53
6	99.73	0.27
7	99.34	0.66
8	99.62	0.38

Table 2. Scenario 2 Clustering Results

Cluster	%Not Fraud	%Fraud
1	99.81	0.19
2	99.85	0.15
3	99.23	0.77
4	99.99	0.01
5	99.80	0.20
6	99.74	0.26
7	99.59	0.41
8	99.85	0.15

Table 3. Scenario 3 Clustering Results

Cluster	%Not Fraud	%Fraud
1	99.65	0.35
2	99.61	0.39
3	64.51	35.49
4	99.56	0.44
5	99.72	0.28
6	99.63	0.37
7	99.71	0.29
8	99.76	0.24

Table 4. Scenario 4 Clustering Results

Cluster	%Not Fraud	%Fraud
1	99.81	0.19
2	99.85	0.15
3	99.23	0.77
4	99.99	0.01
5	99.80	0.20
6	99.74	0.26
7	99.59	0.41
8	99.85	0.15

Table 5. Scenario 5 Clustering Results

Cluster	%Not Fraud	%Fraud
1	99.64	0.36
2	99.61	0.39
3	64.21	35.79
4	99.61	0.39
5	99.72	0.28
6	99.63	0.37
7	99.72	0.28
8	99.77	0.23

Table 6. Scenario 6 Clustering Results

Cluster	%Not Fraud	%Fraud
1	99.81	0.19
2	99.85	0.15
3	99.23	0.77
4	99.99	0.01
5	99.80	0.20
6	99.74	0.26
7	99.59	0.41
8	99.85	0.15

Table 7. Scenario 7 Clustering Results

Cluster	%Not Fraud	%Fraud
1	99.54	0.46
2	99.85	0.15
3	99.70	0.30
4	99.78	0.22
5	99.47	0.53
6	99.73	0.27
7	99.35	0.65
8	99.78	0.22

Table 8. Scenario 8 Clustering Results

Cluster	%Not Fraud	%Fraud
1	99.81	0.19
2	99.85	0.15
3	99.24	0.76
4	99.99	0.01
5	99.81	0.19
6	99.78	0.22
7	99.59	0.41
8	99.85	0.15

Table 9. Summary of Scenario Results (Highest Sensitivity Model In Each Scenario)

Scenario	Key Features	Clustering Model	Predictive Model	Sensitivity	Precision	Loss Amount	Insight
1	All except city_pop & transaction_hour	K-means	Deep Learning	0.9753	0.0737	$4673	High sensitivity but very low precision leads to high false alerts. Loss is moderate. Model catches fraud but wastes team effort.
2	All except city_pop & transaction_hour	Model-based	Random Forest	0.9584	0.1295	$3588	Better balance: improved precision and reduced loss. A practical choice with fewer false positives.
3	All except category & transaction_hour	K-means	Deep Learning	0.9702	0.0744	$3720	Good sensitivity, but low precision means more false alerts. Loss is relatively low, but may still burden fraud teams
4	All except category & transaction_hour	Model-based	Deep Learning	0.9692	0.0865	$8283	High sensitivity but low precision, leading to the highest loss. Model is weak at stopping large fraud cases despite catching many.
5	All except category & city_pop	K-means	Random Forest	0.9209	0.0927	$14905	Lowest precision and highest loss. transaction_hour alone is not enough. Not recommended for business use.
6	All except category & city_pop	Model-based	Deep Learning	0.9820	0.0742	$2603	Strong sensitivity but weak precision.
7	All features included	K-means	Random Forest	0.9805	0.1837	$2933	Strong option: highest precision , strong sensitivity, and low loss.
8	**All features included**	**Model-based**	**Random Forest**	**0.9830**	**0.1734**	**$1638**	**Best all-around: second-lowest loss with high precision and sensitivity. Slightly better than Scenario 7 in minimizing financial impact.**
9	All features included	-	Deep Learning	0.9820	0.1259	808.01	No clustering, but still solid results. Lowest loss and decent precision. Shows clustering improves performance, but not critical in every case.