CSBU 5420 | Machine Learning for Business Analytics

|

Detecting Anomalous Corporate Expense Transactions Using Machine Learning:

A Navan-Inspired Case Study

|

Siddhartha Baniya

|

Webster University

Executive Summary

As part of this case study inspired by Navan's business travel and expense platform, I built a machine learning model to detect anomalous corporate expense transactions.

I have used a fraud detection dataset from kaggle . The dataset has 6362620 and 11 attributes. In this project, I have built two ML models, such as Decision Tree and Random Forest using the tool RapidMiner to predict fraudulent transactions.

Both Decision Tree and Random Forest models were tested where Random Forest shows higher performance with 99.88% accuracy, 99.91% precision, 50.28% recall and 0.782 AUC.

Introduction

Navan manages the expenses and travel arrangements of corporate businesses through its platform designed to simplify business spending. One major challenge in expense management is detecting fraudulent or anomalous transactions, which can lead to significant financial loss and compliance issues. Traditional rule-based detection systems often fail to capture complex, hidden fraud patterns, resulting in false positives or missed anomalies.

This project, inspired by Navan's use case, applies supervised machine learning techniques to detect anomalies in expense transactions. By analyzing historical transaction data, the model demonstrates how systems like Navan's could automatically identify suspicious spending, reduce manual audits, and enhance operational efficiency. The project compares Decision Tree and Random Forest algorithms to determine which model provides better fraud detection accuracy.

Methodology

The project adheres to the standard machine learning workflow: problem framing, data exploration, preprocessing, model selection, evaluation, and deployment planning (Aggarwal, 2017). The process began with loading and analyzing the dataset, followed by handling missing values, encoding categorical variables, and normalizing numerical features. For reproducibility, the dataset was split into 70% training and 30% testing using RapidMiner's built-in split operator.

The dataset consists of 6,362,620 records and 11 attributes, including transaction type, amount, balance-related fields, and fraud indicators. In the modeling phase, a Decision Tree was chosen for its interpretability, while a Random Forest was included for improved accuracy and robustness against overfitting. The models were evaluated using key performance metrics such as Accuracy, Precision, Recall, and AUC to assess their effectiveness in detecting fraudulent transactions.

Data Description and Exploratory Data Analysis

Dataset descriptions

- Numerical: amount, oldbalanceOrg, newbalanceOrig, oldbalanceDest, newbalanceDest
- Categorical: type (PAYMENT, TRANSFER, CASH_OUT, etc.)
- Target Variable: isFraud (0 = normal, 1 = fraud)

Row No.	step	type	amount	nameOrig	oldbalance...	newbalanc...	nameDest	oldbalance...	newbalanc...	isFraud
1	1	PAYMENT	9839.640	C12310068...	170136	160296.360	M1979787...	0	0	0
2	1	PAYMENT	1864.280	C16665442...	21249	19384.720	M2044282...	0	0	0
3	1	TRANSFER	181	C13054861...	181	0	C553264065	0	0	1
4	1	CASH_OUT	181	C840083671	181	0	C38997010	21182	0	1
5	1	PAYMENT	11668.140	C20485377...	41554	29885.860	M1230701...	0	0	0
6	1	PAYMENT	7817.710	C90045638	53860	46042.290	M573487274	0	0	0
7	1	PAYMENT	7107.770	C154988899	183195	176087.230	M408069119	0	0	0
8	1	PAYMENT	7861.640	C19128504...	176087.230	168225.590	M633326333	0	0	0

Table 1

Dataset findings

Most of the transactions in the dataset are normal payments, while fraudulent cases are seen mainly in the TRANSFER and CASH_OUT types. Fraud is also more likely when the transaction involves a higher amount or when the transaction duration, implied from balance changes, is unusually long. These patterns suggest that transaction type, amount, and balance-related timing can be strong indicators for detecting fraudulent activity.

Visuals of dataset

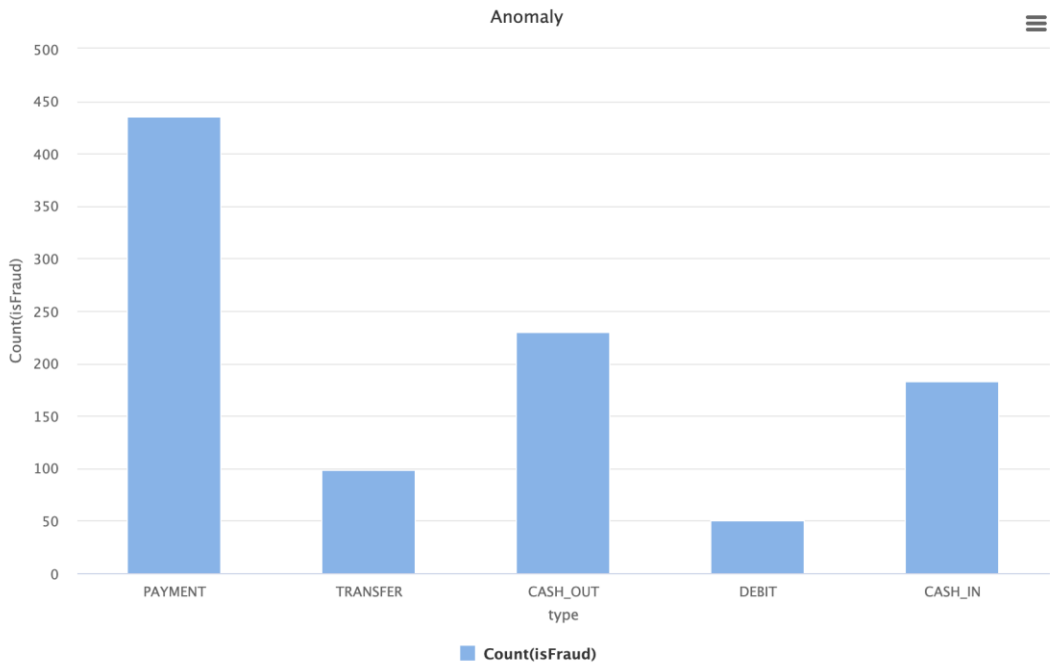


Figure 1

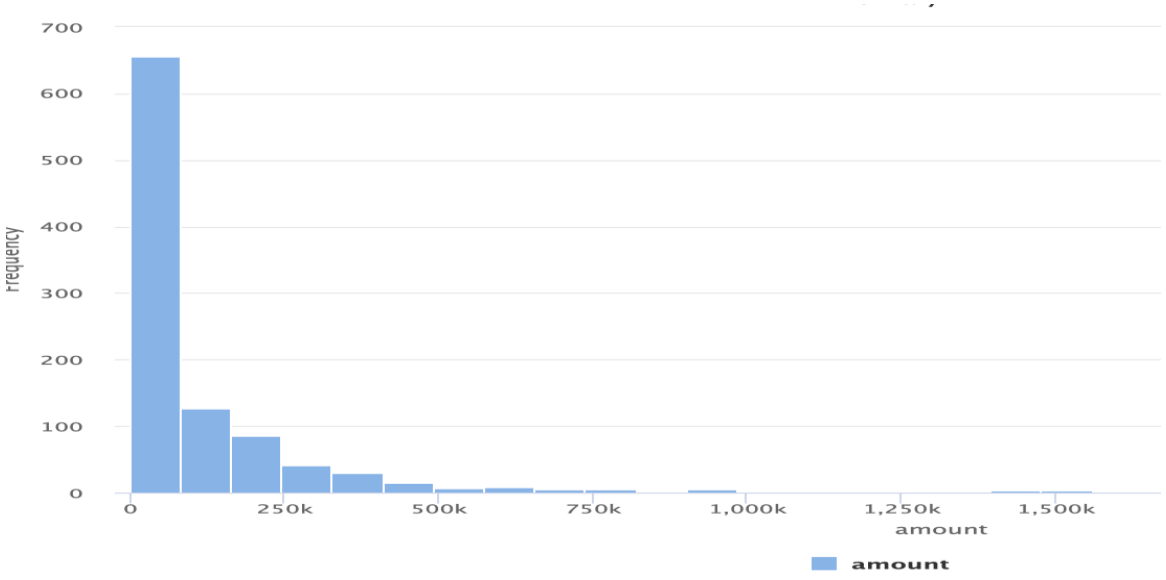


Figure 2

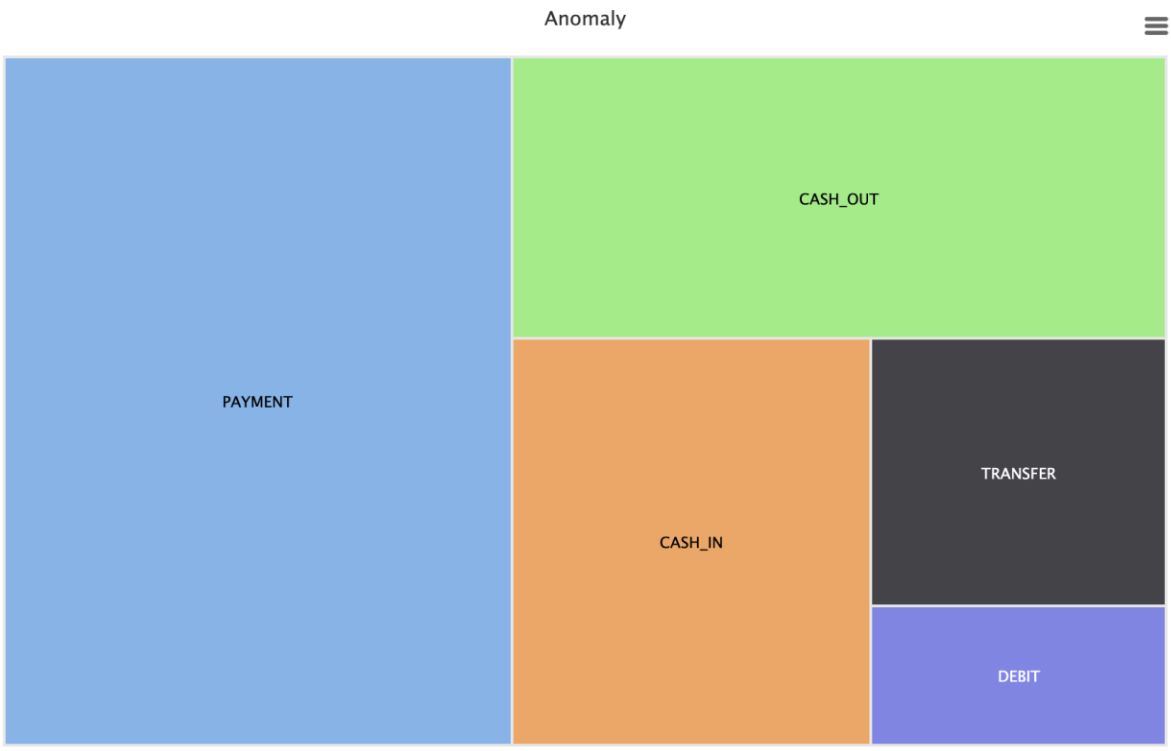


Figure 3

Figure 1 shows that fraudulent transactions are heavily concentrated in the CASH_OUT and TRANSFER transaction types, a pattern consistent with real-world financial fraud (Chandola et al., 2009).

Figure 2 visualizes the amount distribution and indicates that most fraudulent transactions are associated with unusually high values.

Figure 3, the heatmap, highlights strong correlations between balance-related features, suggesting that these attributes are key predictors in distinguishing normal from fraudulent activity.

Modeling and Results

Two classification algorithms were applied in RapidMiner.

Model	Accuracy	AUC	Precision	Recall	Insight
Decision Tree	99.33%	0.661	100%	33.33%	Precise but misses most anomalies.
Random Forest	99.88%	0.751	99.91%	50.28%	Accurate, precise, and detects about half of anomalies.

For the Decision Tree model, default parameters were used (maximum depth = 20, confidence level = 0.25). For the Random Forest model, the number of trees was set to 50, with a maximum depth of 30, selected based on balancing model performance and processing time. These hyperparameters helped significantly reduce overfitting while maximizing recall which is an important metric for fraud detection tasks where missing true fraud cases can be costly (Jain & Gupta, 2018).

Insight & Discussion

From the process run using rapid miner, it shows that both models can detect anomalous expense transactions, but Random Forest does better than Decision Tree because it possesses higher accuracy, precision, AUC and recall. For Navan, a firm which processes thousands of corporate expenses daily, this ensures fewer false alarms while effectively detecting true anomalies.

Adding this ML model to Navan's expense process streamlines the process and makes it more reliable. It can identify unusual transactions automatically, meaning analysts don't have to spend as much time checking everything manually. This enables finance teams to focus more on the high-

risk transactions that need extra attention. In doing so, it also makes the process more transparent and builds clients' trust by showing them their expenses are treated carefully and openly.

Conclusion

The project demonstrates the potential of machine learning to enhance the detection of anomalies in corporate expense data. Using RapidMiner, the Random Forest model achieved 99.88% accuracy, 99.91% precision, and 50.28% recall which makes it a strong candidate for identifying fraudulent transactions in real-world scenarios. This approach highlights how platforms like Navan could move from reactive to proactive fraud management, improving financial governance and reducing the burden of manual reviews.

By incorporating a model like this into an expense management system, organizations could identify suspicious behavior more efficiently, save time and resources, and ultimately build greater trust with their users.

Future Work

Future enhancements to this project could explore the use of unsupervised learning techniques, such as Isolation Forest and Autoencoders, to detect novel and previously unseen fraud patterns that might not be captured through supervised models alone. Additionally, incorporating Natural Language Processing (NLP) to analyze textual data from expense descriptions could reveal subtle anomalies in user input or notes associated with transactions. Finally, developing real-time dashboards that surface anomalies directly within the platform interface would empower finance teams to monitor and respond to suspicious activity more efficiently, helping organizations maintain proactive control over business spending.

References

Aggarwal, C. C. (2017). *Outlier Analysis* (2nd ed.). Springer.

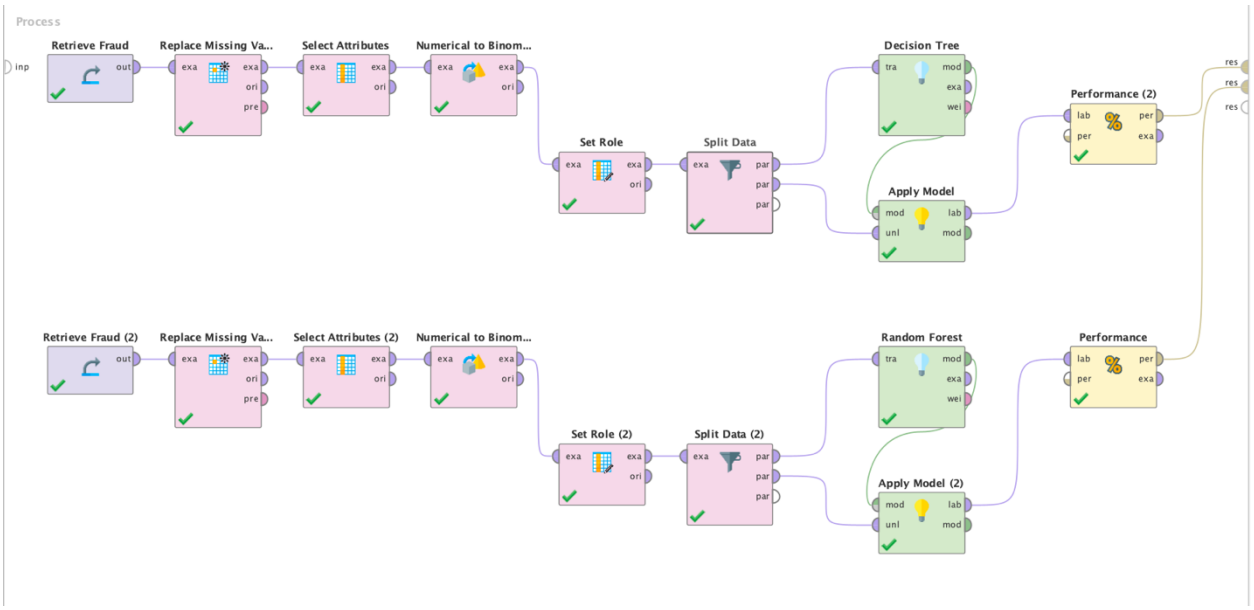
Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), 1–58.

Jain, A., & Gupta, B. B. (2018). A machine learning approach to detect fraud in financial transactions. *Journal of Information Security and Applications*, 40, 80–89.

Navan. (2024). Business travel and expense management solutions. Retrieved from https://www.navan.com

Kaggle. (2013). Fraud detection. Dataset. https://www.kaggle.com/datasets/anidiptapal/fraud-detection-1000-rows

Appendix



Rapidminer process