

Identifying Bank Fraud with Anomaly Detection and Classification Algorithms

23 May 2024

Benedict Brunker

36106 - Machine Learning Algorithms and Applications
Master of Data Science and Innovation
University of Technology of Sydney

Table of Contents

1. Business Understanding	4
A. Use Case and Background	4
B. Key Objectives	4
C. Stakeholder Requirements	4
2. Data Understanding	6
Data Sources	6
Data Collection Methods	6
Data Limitations	7
Exploratory Data Analysis (EDA)	7
Key Insights	10
3. Data Preparation	11
Key Steps	11
4. Modelling	13
a. Machine Learning Algorithms	13
b. Rationale for Algorithm Selection	13
c. Parameter Tuning and Model Selection	14
d. Advanced Feature Engineering	14
e. Modelling	15
5. Evaluation	17
a. Evaluation Metrics	17
b. Results and Analysis	17
c. Key Insights	18

d.	Possible Improvements	19
e.	Business Impact and Benefits	19
f.	Data Privacy and Ethical Concerns	19
i.	Data Privacy	19
ii.	Ethical Concerns	19
iii.	Impacts on Aboriginal and Torres Strait Islander Peoples	20
6.	Conclusion	21
a.	Key Outcomes and Insights	21
b.	Meeting Stakeholder Requirements	21
c.	Future Work, Recommendations, and Next Steps	21
7.	References	23

Executive Summary

This project applies machine learning techniques to enhance the business operations of a bank by identifying fraudulent transactions. A bank has gathered transactional data from its database of customers for the past four years. Although rich in transactional data, the bank lacked expertise in data science and machine learning. We explore this data and propose machine learning applications that bring direct value to the business and its customers. We generate predictive models and analytical insights that could drive better decision-making, improve customer satisfaction, and increase operational efficiency.

The project resulted in a classification model that successfully identified fraudulent transactions, enhancing a bank's ability to prevent financial crimes. We engineered new features with anomaly detectors and K-Means clustering such that fraudulent transactions can be identified with these features through a simple Logistic Regression model.

1. Business Understanding

A. Use Case and Background

The Australian Competition and Consumer Commission (ACCC) last year revealed that “Australians lost a record \$3.1 billion to scams in 2022 ... an 80 per cent increase on total losses recorded in 2021” (ACCC, 2023). The existing literature on this use case is surprisingly thin, perhaps because the enormous value of effective fraud-detection algorithms makes them a jealously guarded industrial secret. However, our research revealed promising results in the use of Logistic Regression, K-Nearest Neighbours, Decision Tree, and Random Forest algorithms (Ranjan et al., 2022; Roy & Prabhakaran, 2021). We tried to replicate and improve on these results, and also innovated with our own approaches, developing Machine Learning Algorithms (MLAs) to detect fraudulent activities in real-time, drawing on vast reserves of both transaction and anonymized customer data. We hope this work will improve banks’ fraud detection capabilities to prevent losses and maintain customer trust. Our fraud-detection algorithms can flag suspicious transactions in real-time and initialize security protocols.

B. Key Objectives

The key objectives of the project are to leverage machine learning algorithms to address the problems outlined above, providing tangible benefits to the bank and its customers.

C. Stakeholder Requirements

1. Customers

- *Requirement* Customer deposits secured against malicious actors.
- *Project Aim* Prevent the defrauding of our customers through fraud-detection algorithms.

2. Compliance Team

- *Requirement* Effective mechanisms to detect and prevent fraudulent transactions and potential money laundering activities.
- *Project Aim* Implement classification models and anomaly detection techniques to accurately identify fraudulent activities and unusual patterns that may indicate money laundering.



3. Marketing Team

- *Requirement* Improving our reputation as a secure bank.
- *Project Aim* Improve our security reputation by deploying state-of-the-art fraud detection.

4. Customer Support Team

- *Requirements* Ability to detect and address abnormal spending patterns promptly, and to alert customers at risk of theft.
- *Project Aim* Develop anomaly detection models to identify and respond to irregular spending behaviours, providing proactive support to customers, and fraud-detection algorithms to flag suspicious transactions in real time.



2. Data Understanding

The dataset consists of customer and transactional data collected by the bank over four years from the beginning of 2019 to end of 2022. When harnessed effectively, this data was crucial for understanding customer behaviours, identifying fraudulent activities, segmenting customers, and detecting anomalies in spending patterns.

Data Sources

1. Customer Data:

- *File* customers.csv
- *Content* Demographic and personal information. 1000 unique customers are represented.
- *Fields* Social Security Numbers (ssn), credit card numbers (cc_num), first and last names, gender, address details (street, city, state, zip), geographical coordinates (lat, long), city population (city_pop), job title, date of birth (dob), and account number (acct_num).

2. Transaction Data

- *Files* transactions_0.csv - transactions_131.csv
- *Content* Detailed records of customer transactions.
- *Fields* Credit card number (cc_num), account number (acct_num), transaction number (trans_num), timestamp (unix_time), transaction category (category), amount (amt), fraud indicator (is_fraud), merchant name (merchant), and merchant geographical coordinates (merch_lat, merch_long).

Data Collection Methods

The data was collected through the bank's transactional systems and customer databases and shared with us for this project.

Data Limitations

1. *Data Quality Issues* Large datasets needed to be handled with care to avoid corruption and proper interpretation. When handled properly the data was sound. Problems early in the project tended to be due to corruption by improper handling.
2. *Privacy Concerns:* Ensuring data privacy and compliance with regulatory standards, especially concerning sensitive customer information. We have not used any feature that may identify that customer in our models. Most sensitive information was dropped in the early stages of the project.

Exploratory Data Analysis (EDA)

1. Summary Statistics:

- *Customer Age* 'dob' ranges from 1927-07-30 to 2007-08-20. Stipulating data collected in 2023, this was a plausible range which did not suggest out-of-bounds issues.
- *Transaction Time* When converted to a *datetime* object, unix time ranges from 2018-12-31 13:00:19 to 2022-12-31 12:59:42. The range again does not suggest out-of-bounds issues.
- *AMT* stands for Alternative Minimum Tax liability, and is a proxy for transaction volume. The minimum AMT is \$1 and max is \$41,300.53, median is \$44.49.

2. Fraudulent Transactions

There are 5034 cases of labelled fraudulent transactions, which represents only ~0.12% of all observations. This presents a classification problem with an extreme class imbalance.

The class balance is inconsistent across years. *Table 1* displays the percentage of fraudulent to total transactions by year. We can see that the fraud rate in 2022 for example is more than three times higher the 2021 rate. The distribution of fraud in the sample data therefore accords with the more general population trend (ACCC, 2023). We therefore selected data from the fourth financial quarter (Q4) of 2022, as the most relevant for deployment.

Table 1	
<i>Fraud Rate by Year</i>	
Year	~ % Fraud
2019	0.095
2020	0.069
2021	0.059
2022	0.19

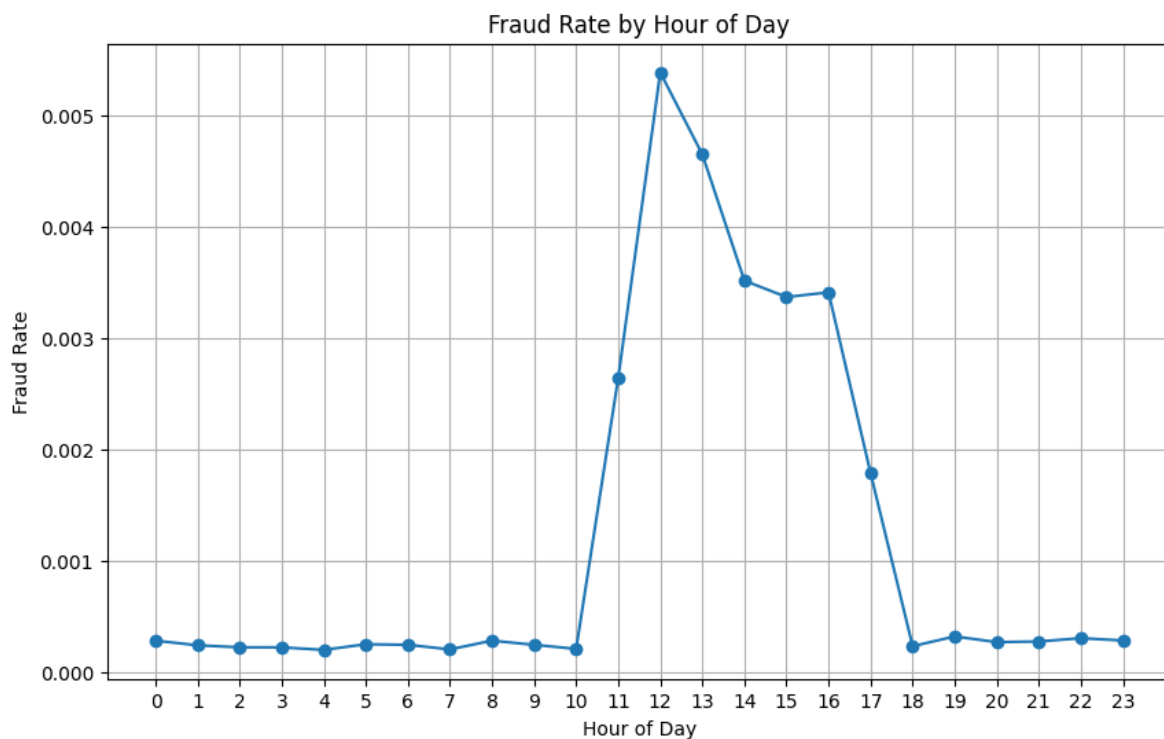
By using 'AMT' (alternative minimum tax liability), we estimate that fraud cost the bank and the 1000 customers represented in the data at least \$3m at the most conservative estimate.

Only a few key variables are relevant to the problem of fraud-detection, and most are liable to introduce noise. Careful EDA is therefore crucial for feature selection, with the added benefit of speeding up execution time for fitting complex models with many observations.

We ran a simple version of Luhn's checksum algorithm on all 'cc_num' values in the data. This algorithm determines the validity of credit cards by computing a relatively simple mathematical formula (Joshua et al., 2023). All values in 'cc_num' passed checksum, which means that fraudulent transactions in the data represent illegitimate use of legitimate cards, not use of illegitimate cards. 'cc_num' is therefore only useful as a customer identifier that presents the least data privacy issues.

The most significant variable by far is the *hour of day* at which a transaction takes place, as visualized in *Figure 1*.

Figure 1



The relationship between fraud and day of the week ('weekday') is less stark, but we could observe promising variance in fraud-rate over the weekly cycle (*Figure 2*).

The *category* of purchase also appears to be crucial in detecting fraud (*Figure 3*).

Figure 3

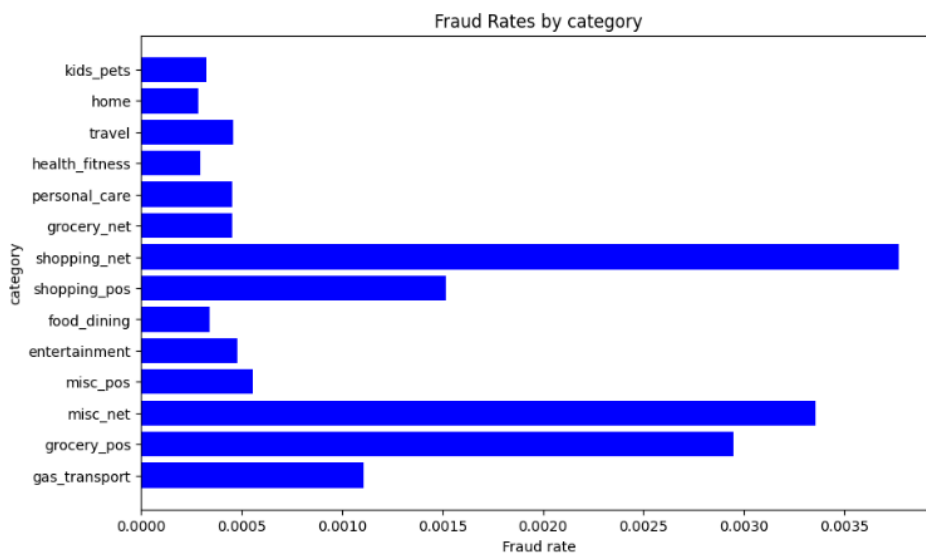
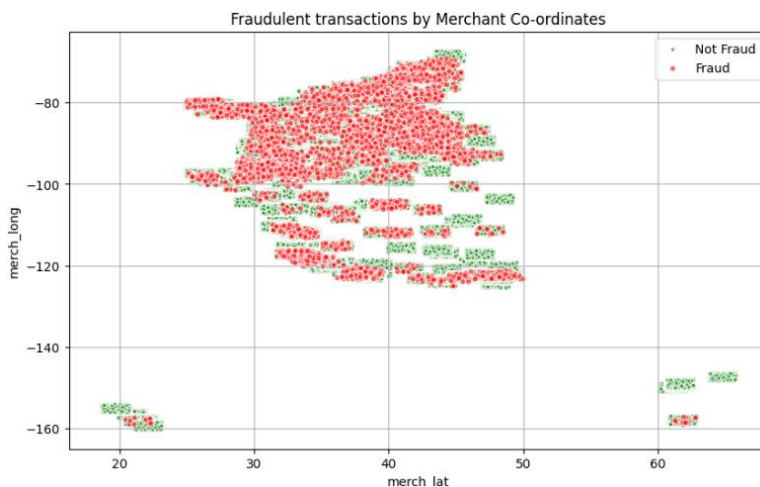
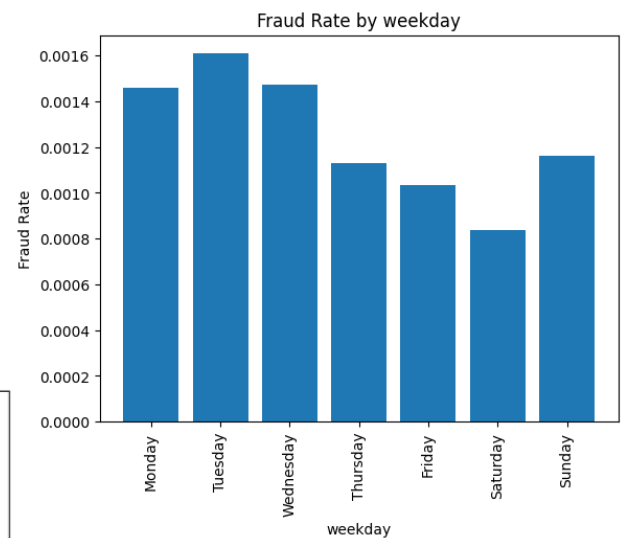


Figure 4



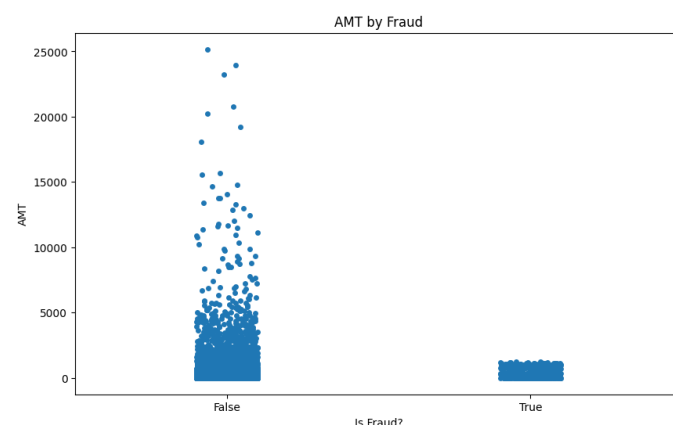
The relationship between fraud and transaction location ('merch_lat' and 'merch_long') is less clear, but we could observe clustering of fraud in certain areas, and its absence in others (*Figure 4*).

Figure 2



Though one might expect higher rates of fraud for larger transactions, EDA revealed the opposite: fraudsters seem to limit themselves to transactions in the low-mid range (*Figure 5*). This suggests that large spending anomalies do not correspond to identified fraud.

Figure 5



We could observe a very slight difference in the defrauding of female over male customers, which might be sufficient to justify the inclusion of 'gender' as a feature for more sophisticated MLAs (*Figure 6*).

Key Insights

1. *Customer Age Distribution*

Most customers are in the 25-50 age range.

2. *Transaction Amounts*

Generally follow a right-skewed distribution, with a small number of high-value transactions.

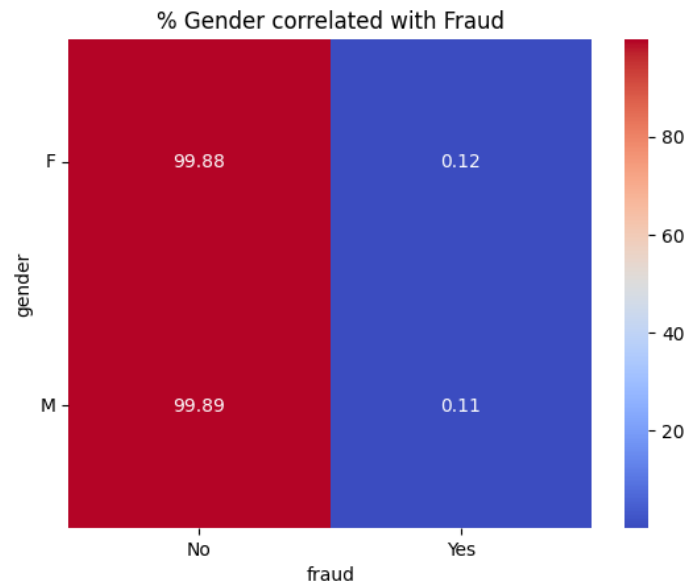
3. *Fraud*

- Occurs almost exclusively for small-medium volume transactions.
- Appears clustered in specific areas and absent in others.
- Occurs more frequently in certain purchase categories.
- Occurs more frequently at night and on certain days of the week.
- Is perpetrated repeatedly against a small group of customers.

4. *Spending Behaviour*

Varies significantly across different geographical locations and demographic segments.

Figure 6



3. Data Preparation

The data preparation phase is crucial for ensuring that the datasets are clean, consistent, and suitable for modelling.

Key Steps

1. Data Cleaning

- *Handling Split Rows* Identified and corrected rows where data was split across multiple columns due to delimiter issues.
- *Malformed Dates* Addressed malformed date of birth (**dob**) entries by validating and correcting formats.

2. Preprocessing

- *Numeric Conversion* Converted **cc_num** to numerical values to run Checksum validation and to assess whether any observations had different cc_num for same acct_num and other customer information (no such cases).
- *Date Conversion* Converted unix_time values to datetime objects for time-series analysis, and to extract new features like the **weekday** and **hour** when a transaction occurred. Converting **dob** to datetime to calculate customer **age**.
- *Dummy variables* One-hot encoded categorical variables like **job** and **category** to facilitate machine-learning.
- *Frequency Tabling* The **merchant** variable contained too many unique values to be represented as dummies, so we transformed the variable to a number representing its frequency in the data, primarily for use in anomaly detection.

3. Feature Engineering:

- *Monthly Spending* Aggregated transaction amounts to compute the total monthly spending per customer. We also engineered a variable for the **cumulative spend** ('cum_amt') of each individual customer across the whole dataset, and a variable for time **elapsed** since last spend for a given customer.
- *Distance* between co-ordinates of the customer's home bank and of the merchant who processed their transaction.

4. Outliers

- *Detection and Treatment* Identified outliers in transaction amounts using statistical methods (e.g., Z-score) and decided on removing of those where absolute Z-score exceeds 3 to normalize the data distribution and reduce potential skewness and variance.

5. Imbalanced Data

- *Fraud Detection* The class imbalance in the **is_fraud** field is addressed by:
 - **Randomly undersampling** the majority class (legitimate transactions) to balance the dataset before fitting models. This strategy has the additional advantage of reducing the size of the training data and speeding up fit-times for more computationally expensive models like Support Vector Machines (SVMs) and Neural Networks (NNs).
 - **Stratifying** data-splits by class.
 - **Tuning** models with the 'balanced' mode for *class_weight* where available, which weights the minority class in inverse proportion to its distribution in the training data.
 - **Evaluating** and tuning models with a focus on improving models' *sensitivity* (or *Recall*) to cases of fraud.



4. Modelling

The modelling phase involved selecting and implementing machine learning algorithms to address the specific business use cases identified earlier. The choice of algorithms was guided by the nature of the data and the objectives of each use case. We also engaged in extensive parameter tuning and model selection to optimize the performance of the models.

a. Machine Learning Algorithms

1. **Classification Algorithms:** To identify binary outcomes, such as fraudulent transactions.
 - *Logistic Regressors* Selected as the most sparse, robust and simple models, with the minimum variance between training and testing performance.
2. **Clustering Algorithms** were used to engineer new features, such as geographic proximity of transactions, transaction types and spending patterns.
 - *K-means clustering* Used for its efficiency, scalability and ability to create clear and distinct clusters.
3. **Anomaly Detection Algorithms** were also used to engineer a new 'anomaly score' feature, which turned out to be highly predictive of fraud.
 - *Isolation forest* This method was used to score the anomalousness of transactions, which could be used as a feature in prediction.

b. Rationale for Algorithm Selection

- **Predictive Accuracy:** Algorithms were selected based on their potential to provide accurate predictions and classifications.
- **Interpretability:** Models like Logistic Regression were chosen for their ease of interpretation, aiding in understanding the underlying patterns.
- **Scalability:** Algorithms capable of handling large datasets efficiently were prioritised.
- **Robustness:** Ensemble methods and deep learning models were selected to improve robustness and handle complex data patterns.

c. Parameter Tuning and Model Selection

- **Grid Search:** Employed to systematically explore combinations of hyperparameters for algorithms.
- **Cross-Validation:** Used to evaluate model performance and prevent overfitting by assessing the model on multiple subsets of the data.
- **Evaluation Metrics:** Recall, Precision, and F-Scores for classification; and Silhouette Score for clustering were used to guide the selection of the best-performing models.

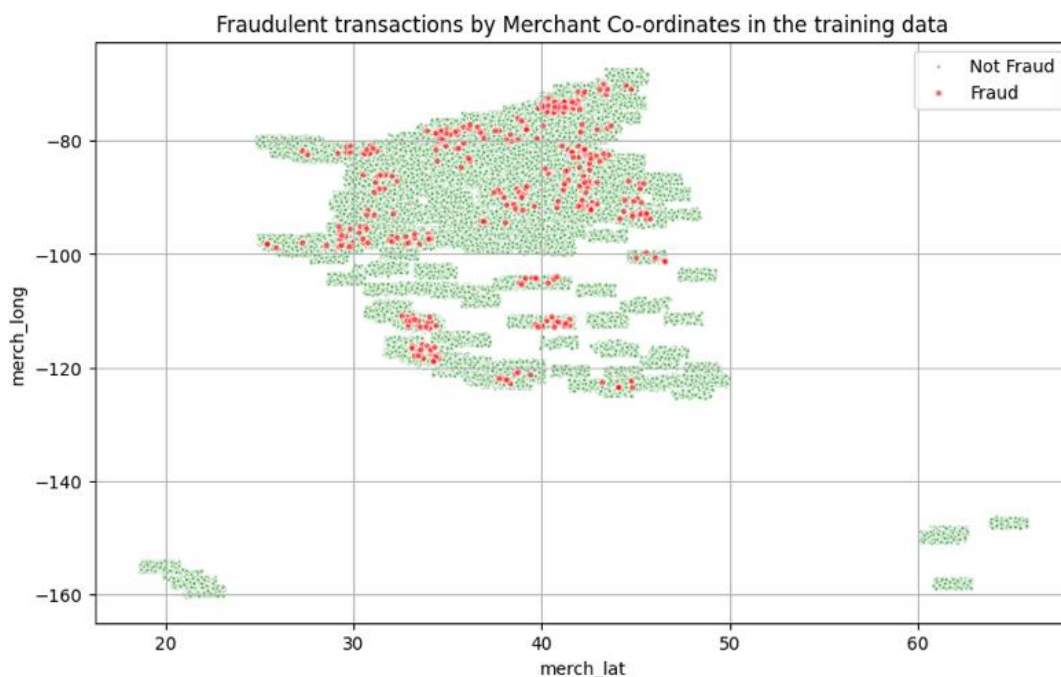
d. Advanced Feature Engineering

We selected only data from Q4 2022 as the most up-to-date and therefore relevant to detecting fraud in future. We **split** the data into training (60%), validation (20%) and test sets (20%), stratifying the split to keep the class balance roughly even.

After splitting, we were able to **engineer new features based only on the training data**. Engineering these features before splitting would have risked contaminating the training data with information from the validation and test sets.

We implemented the K-Means algorithm on merchant co-ordinate data ('merch_lat' and 'merch_long') corresponding to fraudulent transactions, creating a new **transaction zone** variable. We selected the number of clusters K by visually analysing fraud hotspots (*Figure 7*), along with the relative reduction in *inertia* (sum of squared errors between samples and *centroids*, or cluster-centres) and relative improvement in *silhouette score* (a measure of cluster distinctiveness) as new clusters were added.

Figure 7



We identified ten of these hotspots and used the K-Means model to assign all samples from all three sets to these clusters. We inverted the values of the clusters so that larger numbers represented more concentrated regions of fraud. We then calculated the distance from each sample to the cluster centre ('hotspot_distance').

We then fitted our anomaly detectors to the training data, and assessed how well they identified labelled fraud. We excluded features for which EDA had revealed a lack of correlation between anomalies and fraud, like 'AMT'. Our Local Outlier Factor (LOF) model, tuned for novelty detection, marked 86% of labelled fraud as anomalous, and 99% of normal transactions as normal (Figure 9). We used this model to score samples on all three sets ('lof_nov_score'). We took the same approach with our Isolation Forest, which marked 89% of fraud as anomalous, and stored these scores as `ifr_score`.

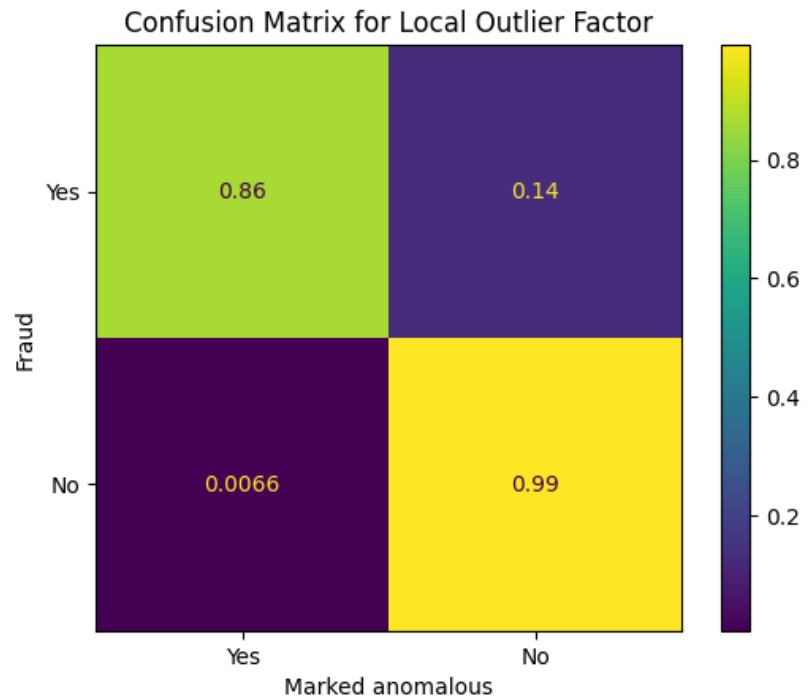


Figure 8

Along with these new variables, we engineered 'age', 'gender', 'amt' and dummies for *weekday* and *category* into feature-set *X*.

e. Modelling

We experimented with a range of MLAs, but the most effective was a relatively simple Logistic Regression model with the following parameters:

```
penalty='l2',  
C=1.0,  
class_weight='balanced',  
random_state=40
```


We selected this model for testing, among others, due to its strong training and validation results, and especially because its validation results improved over training, a sign of robustness (see *Figures 9 and 10* below).

Figure 9

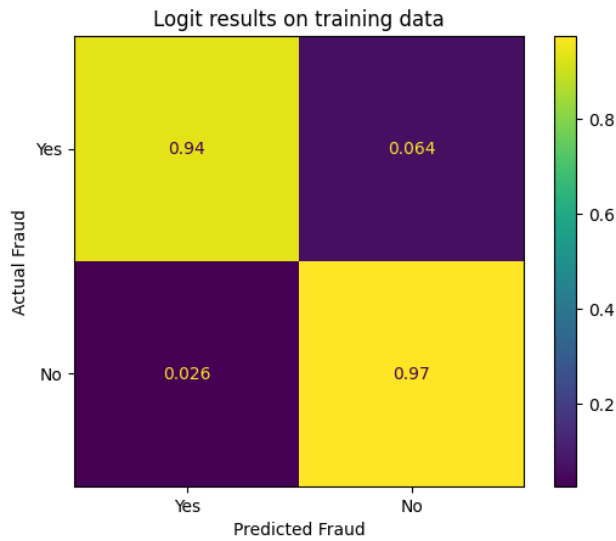
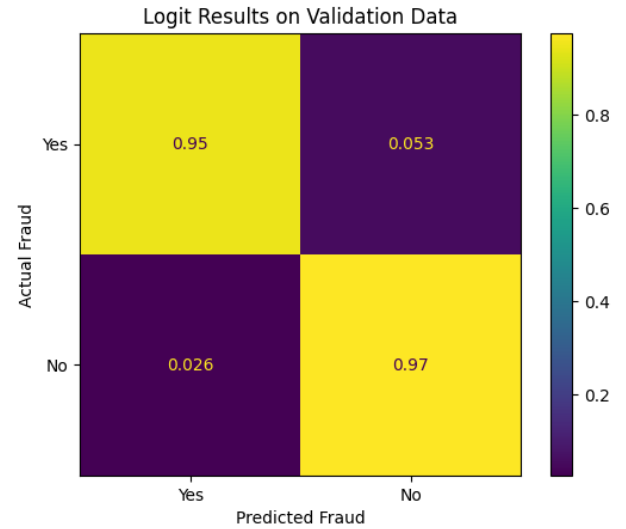


Figure 10





5. Evaluation

a. Evaluation Metrics

- *Accuracy Score* Measures correct predictions to total predictions but is misleading for imbalanced classes, as it may result from always predicting no fraud, which is not useful.
- *Recall Score* Indicates the percentage of actual fraud correctly identified, a crucial measure of the overall success of the model in identifying fraud.
- *Precision Score* Represents the percentage of predicted fraud that is actually fraud. In this case Precision measures how well our model avoids raising *false alarms* and therefore unnecessary costs and inconveniences for customers and staff.
- *F1 Score* Balances recall and precision, considering both missed fraud and false alarms. It is the most important metric for our model, since Recall needs to be tempered by Precision. Given the huge volume of transactions that the bank needs to process daily, even a small percentage of False Positives would represent an enormous number of false alarms.

b. Results and Analysis

Table 2 below summarizes the results of our best performing model on unseen test data (key results in bold).

Table 2			
Test Results for Logistic Regressor Model			
	Precision	Recall	F1-Score
<i>Not Fraud</i>	1.00	0.97	0.99
<i>Fraud</i>	0.05	0.96	0.09
Accuracy			0.97
Macro avg.	0.52	0.97	0.54
Weighted avg.	1.00	0.97	0.99

Figure 10 displays the model's confusion matrix, normalized over true classes.

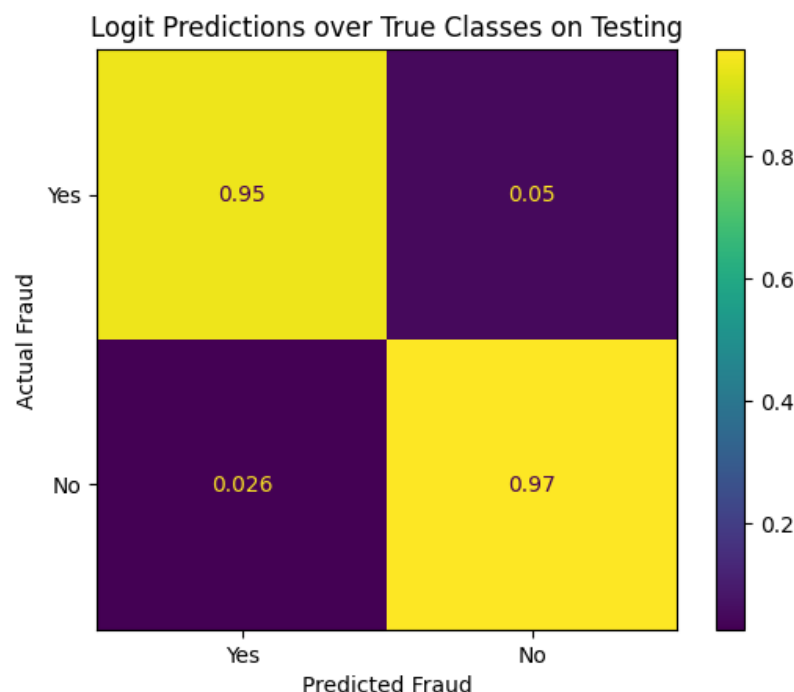


Figure 9

c. Key Insights

Sparser, simpler models like Logistic Regressors actually performed better than more complex linear models like Neural Networks or Support Vector Machines, and tree-based models like Decision Trees or Random Forest. This was likely due to:

- The relatively small number of consistently relevant variables.
- The **two-stage modelling approach** we took, which integrated K-Means and Anomaly Detection algorithms into feature engineering. Much of the relevant information was already condensed into these features. Although useful, these features may have overfitted more complex models to the training data.
- The **heavy class imbalance** in the data meant that more complex algorithms found it too easy to find patterns relating features to legitimate transactions. The additional bias introduced by a simpler model was much preferable to the additional variance of more complex ones. In experimenting with Neural Networks for example, we had to use random undersampling to weight fraud, which reduced the size of the training data and affected model performance, whereas logit models performed better using all the training data with a 'balanced' *class_weight* hyperparameter.

d. Possible Improvements

Although logit models performed best, Neural Networks were also promising. However, we lacked the experience needed to optimize the performance of these models on highly imbalanced data. We also lacked the time or computational resources to fully optimize their performance through automated hyperparameter tuning. Future work on this problem should pursue this approach further.

e. Business Impact and Benefits

- **Impact** The testing data contained 103,464 transactions. The model correctly identified 134 cases of fraud (95%) and missed 7. If deployed on these transactions it would have raised 2732 false alarms (2.6% of legitimate purchases). It would have validated over 100,000 legitimate transactions. We believe this figures would justify deployment.
- **Benefit** Increased security and trust in the bank's services, leading to lower financial losses and improved customer trust.

f. Data Privacy and Ethical Concerns

i. Data Privacy

- *Implications* The project involved handling sensitive customer information, necessitating strict adherence to data privacy regulations such as the General Data Protection Regulation (GDPR).
- *Steps Taken* Ensured data anonymisation by removing fields that can be used to identify a customer before modelling. For example, variables providing information like *Social Security Numbers, names, account numbers* and *addresses* were dropped in the earliest stage of the project.

ii. Ethical Concerns

- *Model Deployment* Addressed potential biases in the models to ensure fair and unbiased treatment of all customers.
- *Negative Impacts* Assessed and mitigated risks of false positives/negatives, particularly in fraud detection, to balance the importance of sensitive fraud-detection with the inconveniences imposed on customers by false alarms.



iii. Impacts on Aboriginal and Torres Strait Islander Peoples

Indigenous Australians' financial losses from **scams and fraud** in 2022 increased by 5.3% from the previous year to \$5.1 million, and "indigenous Australians reported higher median losses (\$1,075) to classified scams compared with all reporters (\$900)" (ACCC, 2023, p.22). We hope our fraud detection models will help to mitigate the impact of fraud on indigenous Australians who are disproportionately impacted, given that "35% of First Nations people had ... incomes in the bottom 20% of the ... distribution for all Australians" (Australian Institute of Health and Welfare, 2023).



6. Conclusion

a. Key Outcomes and Insights

The project successfully leveraged machine learning techniques to address key business challenges faced by the bank.

- **Outcome:** Implemented effective classification models that enhanced the detection of fraudulent transactions.
- **Insights:** Importance of EDA for feature selection to reduce noise and handle large amounts of data. Effectiveness of anomaly detection as a form of feature engineering, and K-Means clustering in using geographic data. Surprising effectiveness of logistic regression using this preprocessing approach.

b. Meeting Stakeholder Requirements

The project achieved its primary goals and met the stakeholders' requirements effectively. The structured application of machine learning algorithms led to significant improvements in the bank's operational capabilities:

- **Customer Satisfaction:** Enhanced by providing predictive insights and proactive support.
- **Security:** Improved through effective fraud detection, protecting the bank and its customers.
- **Operational Efficiency:** Boosted through early detection of fraudulent transactions, allowing timely interventions.


c. Future Work, Recommendations, and Next Steps

1. Model Enhancements:

- Further refine and optimise the models using additional features and more sophisticated techniques.
- Continuously monitor and update the models to adapt to changing patterns and new data.

2. Integration and Deployment:

- Integrate the developed models into the bank's existing systems for real-time predictions and decision-making.

- 
- Deploy the models in a production environment with regular monitoring and maintenance to ensure sustained performance.

3. **Data Enrichment:**

- Incorporate additional data sources, such as social media or transaction metadata, to enrich the dataset and improve model accuracy.
- Explore the use of external data for better customer profiling and segmentation.

4. **Ethical and Privacy Considerations:**

- Continue to prioritise data privacy and ethical considerations in all aspects of model development and deployment.
- Implement strict data governance policies to protect sensitive customer information and comply with regulatory standards.

5. **Collaborative Efforts:**

- Maintain and enhance collaborative efforts with stakeholders to ensure continuous improvement and alignment with business goals.
- Foster a culture of continuous learning and adaptation to keep pace with advancements in machine learning and data science.

In conclusion, the project demonstrated the transformative potential of machine learning in enhancing business operations and customer satisfaction. By addressing key challenges and leveraging data-driven insights, the bank is well-positioned to achieve sustained growth and innovation.



References

- Australian Competition and Consumer Commission. (2023, April). *Targeting scams: Report of the ACCC on scams activity 2022*.
<https://www.accc.gov.au/system/files/Targeting%20scams%202022.pdf>.
- Australian Institute of Health and Welfare. (2023). Income and finance of First Nations people. Retrieved from <https://www.aihw.gov.au/reports/australias-welfare/indigenous-income-and-finance>.
- Joshua, I. O., Idris, A. O., Adebisi, A. A., Kadri, A. F., & Precious, A. E. (2023). Application of Modular Algorithm for Payment Card Number Validation on Mobile Devices Using LUHN. *2023 International Conference on Science, Engineering and Business for Sustainable Development Goals (SEB-SDG)*, 1, 1–8. <https://doi.org/10.1109/SEB-SDG57117.2023.10124487>.
- Ranjan, P., Santhosh, K., Kumar, A., & Kumar, S. (2022). Fraud Detection on Bank Payments Using Machine Learning. *2022 International Conference for Advancement in Technology (ICONAT)*, 1–4. <https://doi.org/10.1109/ICONAT53423.2022.9726104>.
- Roy, N. C., & Prabhakaran, S. (2023). Insider employee-led cyber fraud (IECF) in Indian banks: from identification to sustainable mitigation planning. *Behaviour & Information Technology*, 43(5), 876–906. <https://doi.org/10.1080/0144929X.2023.2191748>.