Assignment 3

ML Group Project

May 24, 2024

Group 3

Student Last Name	Student First Name	Student ID	Group Allocation
Guo	Sophia	14157150	Student A
Montoya Mora	Santiago	24898381	Student B
Kandikattu	Vishal	25413586	Student C
Roman	Valeria	24896716	Student D

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

Table of Contents

Executive Summary

The You&Me Bank ("The Bank") has engaged our Data Science Team (DST) to enhance the banking experience for its customers. The Bank has accumulated transactional data over the past three years and has provided DST with this data alongside basic customer information.

DTS's task is to leverage the provided data to address key business challenges in four main areas:

- Cash Outflow Prediction: Utilizing customer's transaction histories to forecast future cash outflows across various categories.
- Anomaly Detection: Improving customer financial health and strengthening bank relationships by detecting unusually high spending and transactions.
- Fraud Detection: Increasing the security of customer financial accounts by proactively identifying fraudulent transactions.
- Customer Segmentation: Developing a model to categorize customers with similar spending behaviours.

DST employed various machine learning methodologies and conducted experiments to assist The Bank in understanding and predicting customer behaviour in the areas mentioned above. Following the deployment of these machine learning models into production environment, significant benefits have been realized:

For Customers:

- The ability to predict their future cash outflows.
- Alerts concerning potential fraud and anomalous transactions.

For The Bank:

- Enhanced customer loyalty and retention through the introduction of advanced features.
- Improved customer segmentation for targeted marketing initiatives.

These outcomes contribute to a more personalized and secure banking experience, benefiting customers and The Bank.

1. Business Understanding

a. Business Use Cases

The team focuses on developing features for predicting cash outflows, detecting fraudulent activities in real-time and anomalies, also... These projects address direct customer needs for better financial management tools and more robust security measures, thus providing significant value to customers and the bank.

Cash Outflow Prediction

The Bank's Customer Experience team wants to introduce a new feature in the online banking system so that customers can see their future cash outflow for different spending categories. The initiative was motivated by high interest rates and economic inflation; customers want to monitor their spending to cope with the challenging market.

After conducting an initial survey, The Bank understands that its customers use its spending history to make its budget. They would go through the bank statements and manually calculate cash flow to predict their future spending.

The Customer Experience team spots this gap where The Bank can provide better service by introducing a new feature that can automate this process by using customer's history transaction data and Machine Learning algorithms to predict the customer's cash outflow. This feature will also help The Bank monitor its liquidity position.

Prediction of fraudulent transactions

Furthermore, the Bank's Customer Experience team plans to implement a machine learning system that classifies real-time transactions to distinguish potential fraud. This crucial initiative enables prompt interventions on suspicious transactions, enhancing security and boosting customer confidence.

However, deploying this system poses several challenges, including handling the vast volume and rapid pace of transactions, addressing data imbalances affecting prediction accuracy, and continuously adapting to evolving fraudulent tactics. Additionally, minimising false positives is critical to avoid customer dissatisfaction and protect the Bank's reputation.

Machine Learning algorithms are adept at meeting these challenges. They process large data sets efficiently and identify complex fraud patterns. By utilising advanced techniques, these algorithms not only automate the detection process but also improve the system's adaptability through ongoing learning. This adaptability is essential for maintaining high predictive accuracy and operational efficiency in the dynamic financial environment, affirming the indispensable role of machine learning in effective fraud prevention.

Anomaly detection

The bank's Customer Experience team is implementing an advanced machine learning system designed to monitor and analyse financial transactions over $200. This approach focuses on high-value transactions to maximise the system's impact in identifying anomalies that could indicate potential fraud and significant errors. By concentrating efforts on these transactions, the bank seeks to protect its customers' financial assets and ensure that monitoring is used efficiently.

The main challenges for this implementation include handling the daily volume of transactions the bank processes and the speed with which these transactions must be evaluated. Adjustment is required to avoid false positives, which could lead to customer dissatisfaction and damage the bank's reputation.

Machine learning algorithms, especially those based on techniques such as Isolation Forest, are ideal for addressing these challenges. These algorithms not only process large volumes of data with high efficiency but can also identify complex and subtle patterns that indicate anomalous behaviour. By incorporating these systems, the bank not only enhances its ability to respond to financial threats but also reinforces the confidence of its customers by demonstrating an ongoing commitment to security and technological innovation in protecting their financial interests.

Clustering

The bank Customer Experience team uses machine learning to categorize customers by spending habits, allowing targeted marketing. By identifying groups like high spenders or budget customers, they can create customized offers, making marketing more effective and increasing customer engagement.

Challenges include:

- Combining and cleaning the data to ensure it’s accurate.
- Choosing the suitable algorithms for grouping customers and making sure the results are reliable.

The bank Customer Experience team must follow data privacy laws and integrate insights into current systems to maximize customer data responsibly.

This project boosts customer loyalty with targeted offers, helps upsell products, and increases earnings. It also improves fraud detection by spotting unusual spending patterns, enhancing customer safety and the overall customer experience.

b. Key Objectives

The project's main goal is to improve The Bank's abilities using advanced machine learning tools in various areas. This includes making models for spotting fraud, predicting customer spending, and enhancing marketing efforts. The aim is to be more efficient, accurate, and secure while following all rules about data and security.

Key people in this project are:

- DST (Data Science Team): They make and use machine learning models for the bank's systems.

- The Bank’s Legal Team: They make sure everything is legal and follows data rules.

- Banking Data Team: They help handle data safely.

- Banking Platform Team: They help put new tools into the bank's systems.

- Fraud and Marketing Teams: They work with DST to make sure the tools help them do their jobs better.

The project helps these groups by making custom tools and getting legal approval before using them. It also makes sure data is handled safely and that new tools fit well with the bank's systems. By working closely with teams like Fraud and Marketing, DST makes tools that help the bank work better and keep customers happy.

2. Data Understanding

DSA was given 2 datasets via secured cloud file transfer to work on first – Customers and Transactions.

In the one customer file:

- There are 1000 unique customer records.
- There are 12 data attributes excluding the customer and account identifiers.
- Among the 12 data attributes:
    - There are 8 attributes are specific to customers provided: First name, last name, street address, zip code, latitude, longitude, job, date of birth.
    - There are 4 attributes not specific to certain customers but apply to more generic customers' profile: gender, city, state, city population.
- The customers are well distributed in terms of customers profiling.
- Data collected through customer applications and transaction histories.
- The dataset includes only 1000 entries, which might not be representative of the entire customer base.
- Certain data points, like 'dob' (date of birth), don't change over time, but others, like 'job' or 'address', might become outdated if not regularly updated.

This dataset has detailed info on customers, useful for better marketing, service delivery, and spotting fraud.

In the 132 transactions files:

- There are ~5m records of transactions in total.
- There are 7 data attributes excluding the customer, account and transaction identifiers.
- Among the 7 data attributes:
    - There are 2 pre-categorised attributes:
        - category – indicating 14 different categories that the transaction belongs to.
        - is_faud – an indicator of fraud transaction.
    - There are 5 raw attributes related to transaction time, amount, merchandise name and location.

3. Data Preparation

Transactions

- Combine multiple Transaction Datasets into 1 Dataset.
- Remove duplicate Transactions and NA records.
- Use non-fraud transaction only for Regression and Clustering modelling.
- Transform the Unix Time to standard datetime format.
- Refine existing transaction categories and encode categorical features as an integer array.
- Used statistical methods to detect and handle outliers in numerical columns to prevent them from skewing the model.
- Scaled numerical features to a standard range to ensure uniformity and prevent certain features from dominating the model.
- Selected relevant features using correlation analysis or feature importance to improve model efficiency and reduce noise.

Customers

- Calculate customers' age base on Date of Birth
- Encode categorical features – gender, residency state as integer arrays.

4. Modelling

a. Regression Modelling

To predict the customers' future cash outflows / spendings, regression models are suitable for this problem statement. To assess which specific Machine Learning Algorithm is more fit for purpose, below analysis and observations have been taken into consideration and Extra Tree Regressor were chosen for this business use case.

- Aggregate transaction amount to each Month End date to see the high-level trend:

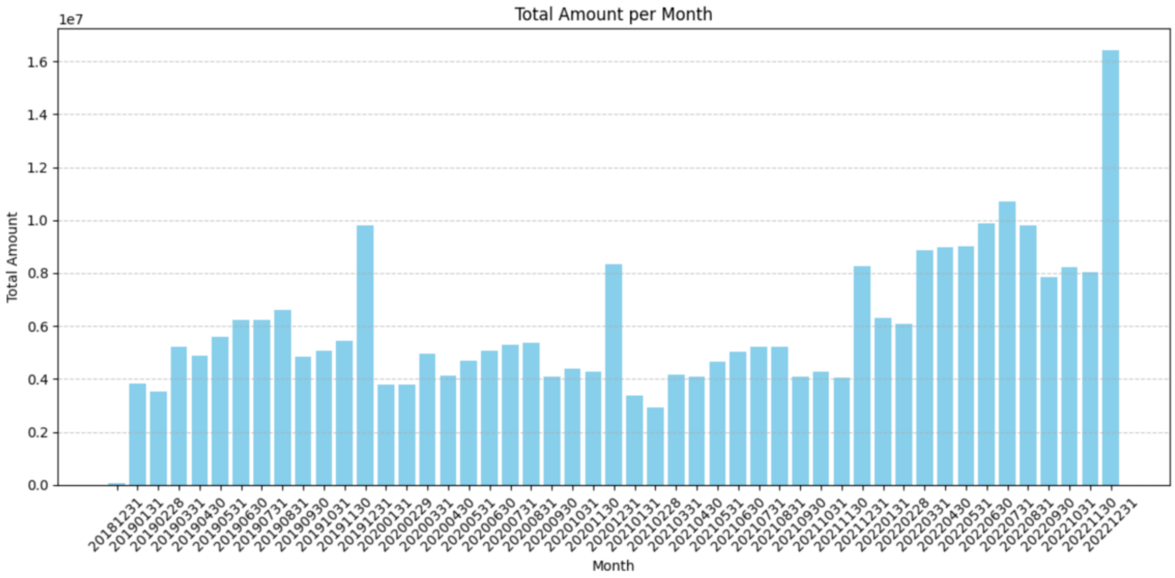


Figure 1*. Total Transaction Amount per Month*

  - The high-level bar,**figure 1** shows that the aggregated transaction amount and the timeline does not demonstrate a linear relationship.
  - There are significant data missing in Dec 2018 – exclude those transaction data to improve model accuracy and reduce noises.
  - Seasonality is observed in spending amounts for different category as per **figure 2** , which is a factor that should be considered in Regression modelling: a new attribute – Season has been created based on month.

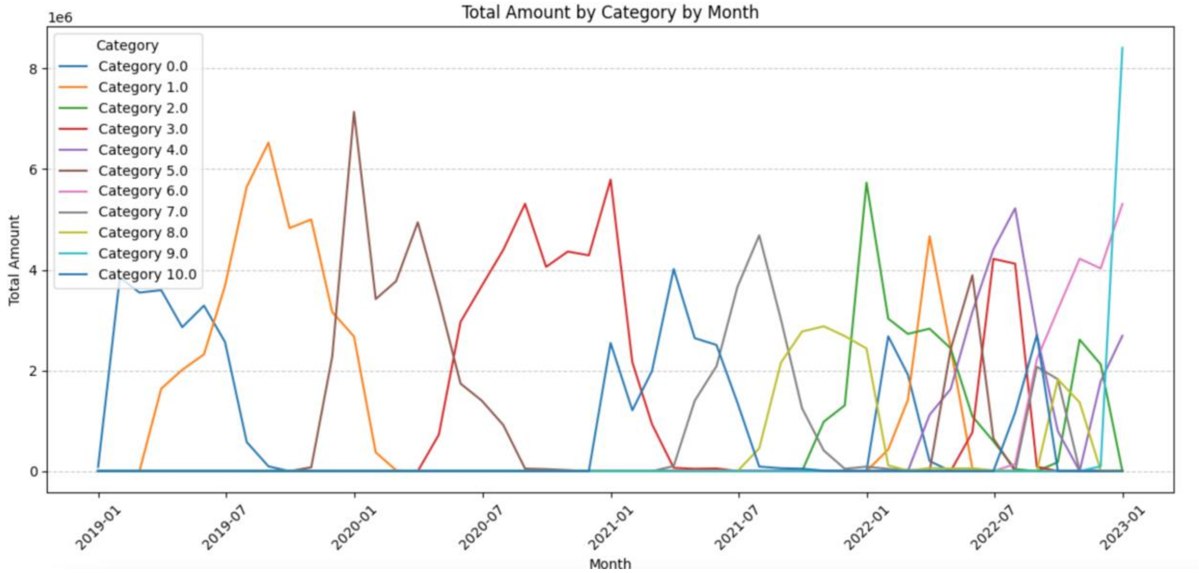
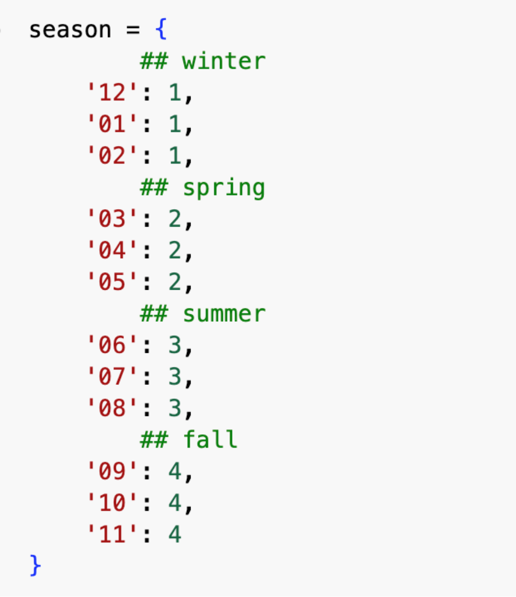


Figure 2. Total Transaction Amount by Category per Month



```
season = {
        ## winter
    '12': 1,
    '01': 1,
    '02': 1,
        ## spring
    '03': 2,
    '04': 2,
    '05': 2,
        ## summer
    '06': 3,
    '07': 3,
    '08': 3,
        ## fall
    '09': 4,
    '10': 4,
    '11': 4
}
```

- High Level customer profiling
    - Customers are well distributed across different ages, states, cities, genders and jobs.
    - Join the Transaction Dataset to Customers Dataset for the model to learn the customers' profile impact on the spending habits.

Gender	Count of Records
F	512
M	488

Table 1. Customer records for each gender

Data Attribute	Unique Values Count
state	51
city population	776
Age (calculated based on DoB)	978
city	736
job	507

Table 2. Unique Values for each data attribute

- Model parameter finetuning
  - Use the default parameters of Extra Tree Regressor first and feature_importances_ attribute to assess the features that has most importance to the prediction figures.
  - Pass the features to the next step for further parameter refinement:
    - No transformation: city_pop
    - Encoding transformation: State, category
    - Datetime transformation: end_of_month – derived from unix_time of transaction, season – derived from end_of_month, age – derived from customer's DoB
  - Use RMSE as the valuation tool to assess the model accuracy. Use different value combinations of parameters to try out the range where the accuracy is getting higher.
  - Use Cross Validation (Grid Search) to get the optimal combination of the parameters.

b. XGBoost Algorithm

The XGBoost algorithm, known for its efficiency in handling large and complex datasets, was used to identify fraudulent transactions. This algorithm is ideal for classification tasks such as fraud detection because it effectively manages unbalanced data and highlights the importance of different features influencing the predictions. The model's effectiveness is demonstrated by its focus on transaction amount as the most significant feature, as shown in **Figure 3**; this indicates that higher transaction values are closely scrutinized, aligning with the common understanding that more substantial amounts are often more susceptible to fraud, the database for this analysis was specifically filtered to include data from the last three months, ensuring the relevance and timeliness of the findings.

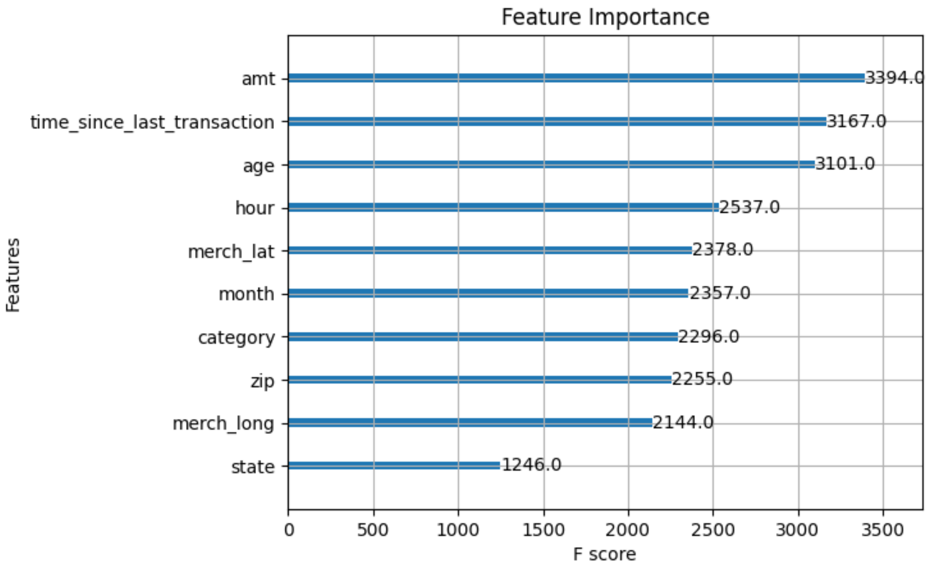


Figure 3. Feature importance XGBoost.

For the XGBoost model used in fraud detection, feature engineering involved several vital techniques:

- Encoding categorical variables for compatibility with the model.
- Scaling features to aid in convergence.
- Managing missing values through imputation to optimize performance.

The model was trained using data resampling with SMOTE to correct class imbalances and cross-validation to ensure robust performance across unseen data.

Additionally, a 'unix_time' column was created to extract 'hour,' 'day_of_week,' and 'month' features, after which 'unix_time' was removed from the model, unique identifiers and customer-specific values were excluded from the model since they do not enhance its predictive capability. The 'dob' was converted to calculate an 'age' column, which replaced the original 'dob' data to streamline the feature set. These steps were crucial for preparing the data to train the XGBoost model for fraud detection effectively.

On the other hand, parameter tuning in XGBoost involved adjusting several key hyperparameters to optimize performance, configured as shown in **Table 3**.

Hyperparameter	Description	Range of Values	Value
Max_depth	Controls the depth of the trees. Deeper trees can learn more detailed patterns but can lead to overfitting.	(3,5)	5
Min_child_weight	Helps decide whether to split at a node based on the weights of all observations that go to that node. Higher values prevent the model from learning overly specific patterns, reducing overfitting.	(1,5)	1
Gamma	Minimizes loss required to make further partition on a leaf node of the tree, which can help in pruning and controlling complexity.	(0.1, 0.5, 1.5)	0.1

Table 3. *Hyperparameters used to assess the model.*

Model selection was based on cross-validation scores focusing on recall, precision, and the area under the ROC curve to ensure that the model predicts accurately and minimizes false negatives, which are critical in fraud detection.

c. Isolation Forest

For the case of anomaly detection, the isolation forest algorithm was implemented; this algorithm was designed specifically for this purpose. It is a tree-based model that isolates anomalous points, creating partitions in data sets that identify patterns. It is a very efficient model for large datasets and where there is an imbalance in their categories.

Before the model, different pre-processing of essential data was performed to ensure better anomaly detection.

In the first instance, the transaction database was merged with the customer database to have relevant transaction and demographic information, after which it was filtered to use only the last six months of data because the use of more time implies more computing power and more training time, which the DST team does not have.

In terms of feature engineer:

New columns:

- Extracted temporal data (hour, day_of_week and month) from the unix_time column to have more detailed transaction information.
- The time between transactions of the same account was calculated to find anomalies in the time delta of the transactions.

Dropped columns:

- The unix_time and dob columns were removed.
- All identifier columns, such as transaction_id, cc_num, trans_num, etc., were removed to avoid overfitting the model.
- Removed columns with personal customer information, such as first and last name and specific address.
- Columns with a high degree of uniqueness, avoiding overfitting, long, lat, job, and merchant.
- The is_fraud column was eliminated because the information is not obtained instantly; given this, it is not helpful for this business case.

Following the featured engineer, the database was filtered, leaving only transactions greater than 200, as this is the focus of the business case.

Finally, the numeric variables were scaled using standard scaling while the categorical variables were given label encoding.

Since this is an unsupervised algorithm, the data were not separated between training and test. The configuration of this model is:

- **N_estimators**: 300, defining the number of trees; this value ensures a good model performance without greatly increasing the time.
- **Max_samples:** ‘auto’, this hyperparameter helps keep the model efficient and avoids overfitting since each tree is built from a subset of the data.
- **Contamination:** ‘auto’, defines the expected proportion of outliers in the dataset.

This setting ensures that the model can effectively handle large datasets and adapt to different amounts of outliers without manual intervention, which is expected since no label explains which data is outliers.

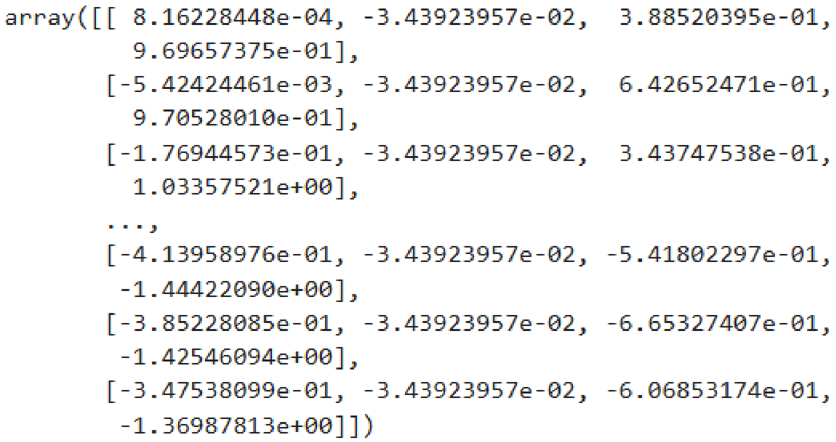
No tuning process was performed because this is an unsupervised algorithm with no optimisable performance metrics.

d. K-means clustering

The primary algorithm used for modelling to send customized marketing emails to groups of customers presenting similar spending behaviours is k-means clustering. This algorithm partitions the data into a specified number of clusters based on feature similarity, aiming to minimize the variance within each cluster.

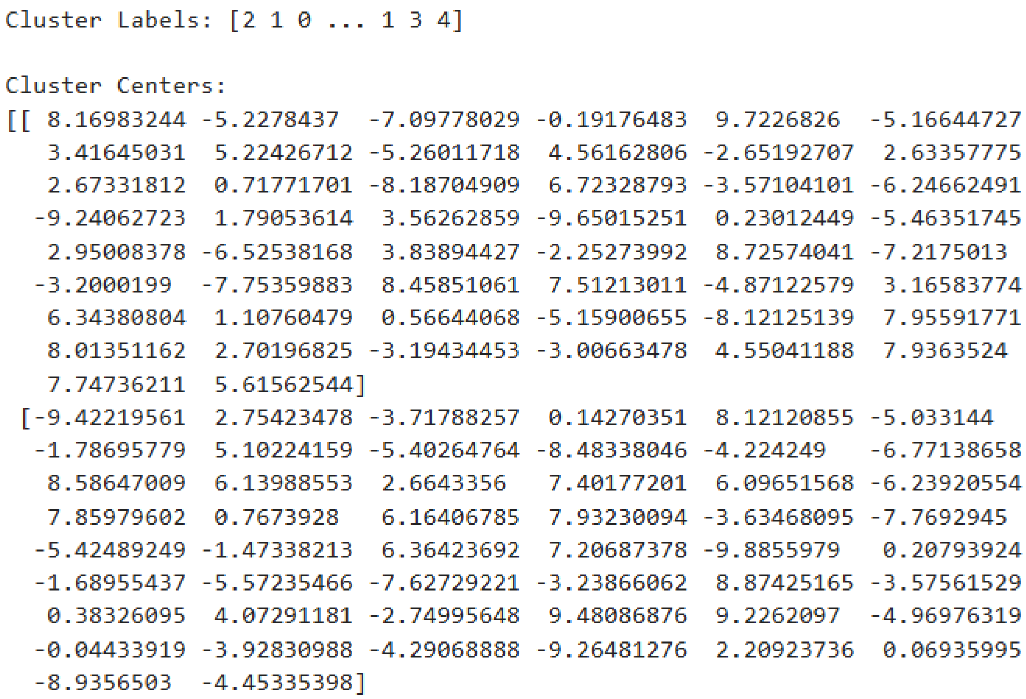
K-means clustering is a great tool for grouping data without needing someone to tell it how to group things. It's especially good at dividing up big sets of data, it's not too tricky to set up, and it works fast compared to other ways of grouping things. That's why we picked it for figuring out different types of customers based on how they spend money, which helps us plan marketing campaigns that target each type better.

The training process for k-means clustering involves a few simple steps. First, we pick some starting points called centroids to act as the centers of the clusters. Then, we assign each data point to the nearest centroid.



```
array([[ 8.16228448e-04, -3.43923957e-02,  3.88520395e-01,
         9.69657375e-01],
       [-5.42424461e-03, -3.43923957e-02,  6.42652471e-01,
         9.70528010e-01],
       [-1.76944573e-01, -3.43923957e-02,  3.43747538e-01,
         1.03357521e+00],
       ...,
       [-4.13958976e-01, -3.43923957e-02, -5.41802297e-01,
        -1.44422090e+00],
       [-3.85228085e-01, -3.43923957e-02, -6.65327407e-01,
        -1.42546094e+00],
       [-3.47538099e-01, -3.43923957e-02, -6.06853174e-01,
        -1.36987813e+00]])
```

After that, we update the centroids by calculating the average of all the points assigned to each one.



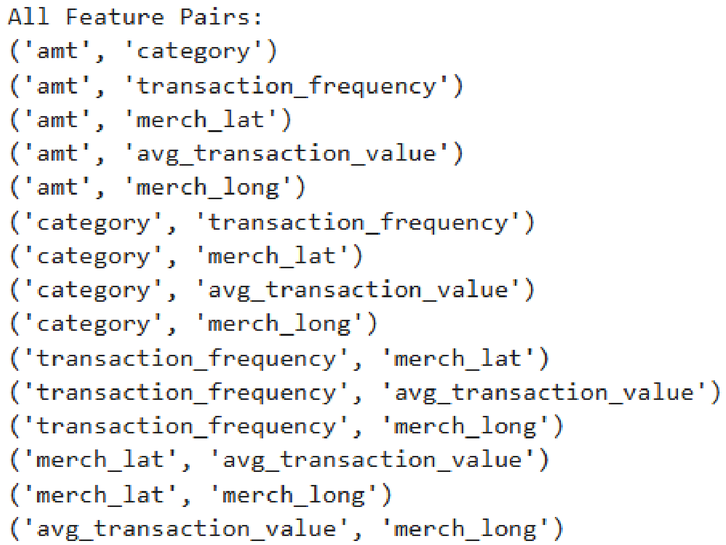
```
Cluster Labels: [2 1 0 ... 1 3 4]

Cluster Centers:
[[ 8.16983244 -5.2278437  -7.09778029 -0.19176483  9.7226826  -5.16644727
   3.41645031  5.22426712 -5.26011718  4.56162806 -2.65192707  2.63357775
   2.67331812  0.71771701 -8.18704909  6.72328793 -3.57104101 -6.24662491
  -9.24062723  1.79053614  3.56262859 -9.65015251  0.23012449 -5.46351745
   2.95008378 -6.52538168  3.83894427 -2.25273992  8.72574041 -7.2175013
  -3.2000199  -7.75359883  8.45851061  7.51213011 -4.87122579  3.16583774
   6.34380804  1.10760479  0.56644068 -5.15900655 -8.12125139  7.95591771
   8.01351162  2.70196825 -3.19434453 -3.00663478  4.55041188  7.9363524
   7.74736211  5.61562544]
 [-9.42219561  2.75423478 -3.71788257  0.14270351  8.12120855 -5.033144
  -1.78695779  5.10224159 -5.40264764 -8.48338046 -4.224249   -6.77138658
   8.58647009  6.13988553  2.6643356   7.40177201  6.09651568 -6.23920554
   7.85979602  0.7673928   6.16406785  7.93230094 -3.63468095 -7.7692945
  -5.42489249 -1.47338213  6.36423692  7.20687378 -9.8855979   0.20793924
  -1.68955437 -5.57235466 -7.62729221 -3.23866062  8.87425165 -3.57561529
   0.38326095  4.07291181 -2.74995648  9.48086876  9.2262097  -4.96976319
  -0.04433919 -3.92830988 -4.29068888 -9.26481276  2.20923736  0.06935995
  -8.9356503  -4.45335398]
```

We keep repeating these steps assigning points and updating centroids until the centroids don’t change much anymore, which means we’ve found the best clusters.

To figure out the best number of clusters (k), we used the Elbow Method and the Silhouette Score. For the Elbow Method, we made a plot showing the sum of squared distances (SSE) for different k values and looked for the point where the SSE stops decreasing quickly and starts to level off which is for 10 clusters. And the Silhouette Score helped us check how good the clusters were for different k values, with higher scores meaning better clusters. Using both methods together helped us choose the right number of clusters. Our Silhouette Score was 0.92.

To help the marketing team send customized marketing emails to groups of customers with similar spending behaviours using clustering, we need to ensure we have features that effectively capture the spending behaviour of customers. The best feature pairs are:



```
All Feature Pairs:
('amt', 'category')
('amt', 'transaction_frequency')
('amt', 'merch_lat')
('amt', 'avg_transaction_value')
('amt', 'merch_long')
('category', 'transaction_frequency')
('category', 'merch_lat')
('category', 'avg_transaction_value')
('category', 'merch_long')
('transaction_frequency', 'merch_lat')
('transaction_frequency', 'avg_transaction_value')
('transaction_frequency', 'merch_long')
('merch_lat', 'avg_transaction_value')
('merch_lat', 'merch_long')
('avg_transaction_value', 'merch_long')
```

Using these features will provide a more comprehensive understanding of customer spending behaviours, aiding in more effective and customized marketing campaigns.

5. Evaluation

a. Evaluation Metrics

Model	Evaluation Metrics	Rationale
Regression Model	RMSE Feature Importance	RMSE is a measure of the average magnitude of the errors between predicted and actual values. For predicting future cash outflow, precision is crucial. Customers rely on these predictions to manage their finances under high interest rates and inflation. RMSE is a good indicator that can help ensure the model provides accurate and reliable predictions, minimizing the risk of significant deviations that could mislead customers. Since RMSE is in the same units as the target variable (e.g., currency), it's also easier to interpret. Feature Importance is used to reduce the noise caused by unimportant features to improve the model's accuracy
XGBoost Algorithm	Recall, precision, F1 score and AUC.	We utilized vital metrics such as recall, precision, F1-score, and AUC to evaluate our model. We prioritize recall to capture as many fraudulent transactions as possible, reducing financial losses. Precision is also monitored to keep false positives low and maintain customer satisfaction, though it less emphasized than recall. The F1-score integrates precision and recall, offering a balanced assessment of the model's accuracy and efficiency. Additionally, AUC helps determine the model's ability to differentiate between fraudulent and legitimate transactions, assisting in setting the optimal classification threshold. Our focus on recall and F1-score ensures the model effectively detects fraud and is practical for real-world use without overwhelming users with false alerts.
Isolation Forest	t-SNE Visualization of High-Dimensional Data	t-SNE serves as a visual validation tool that helps verify the effectiveness of Isolation Forest in identifying anomalies. It is also a powerful communication tool, transforming complex data relationships into clear and understandable visualisations. This ensures that the model's findings are accessible and valuable to the monitoring team, thereby improving the impact and adoption of the model.
k-means Clustering	Silhouette score, Elbow Method, Feature pairs	Silhouette score and the Elbow Method ensure meaningful clusters for customer segmentation. By analyzing feature pairs, we discern customer behavior, guiding marketing strategies. K-means clustering groups customers by spending habits, enabling personalized marketing emails and enhancing bank-customer connections.

b. Results and Analysis

Regression Modelling

After using 80% of the total dataset for training and 20% for testing, the final RMSE of the training set is 2134 and RMSE for the testing set is 2969 using below features:

Feature	Importance
state	0.10886703
age	0.32521686
category	0.17243165
end_of_month	0.22799394
season	0.06100052
city_pop	0.10449001

Table 4. Feature Importance

From the training dataset, the model is more accurate from monthly spending of 0 to 10,000 dollars per category, and the level of accuracy continues to drop from 10,000 to 40,000+ dollars.

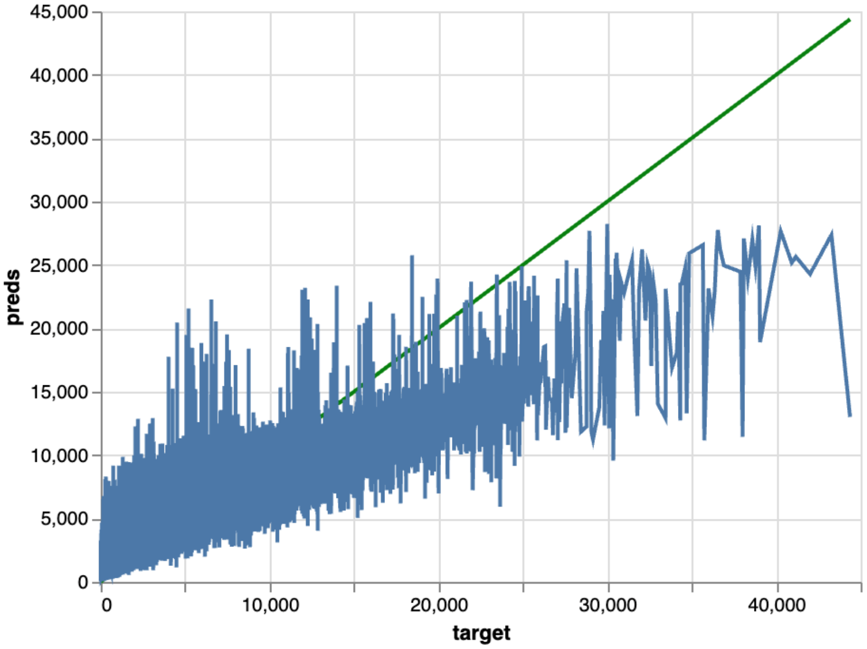


Figure4 . Training Set Target vs Predictions

From the testing set, the same trend has been observed but the accuracy starts to drop from 10,000 dollars.

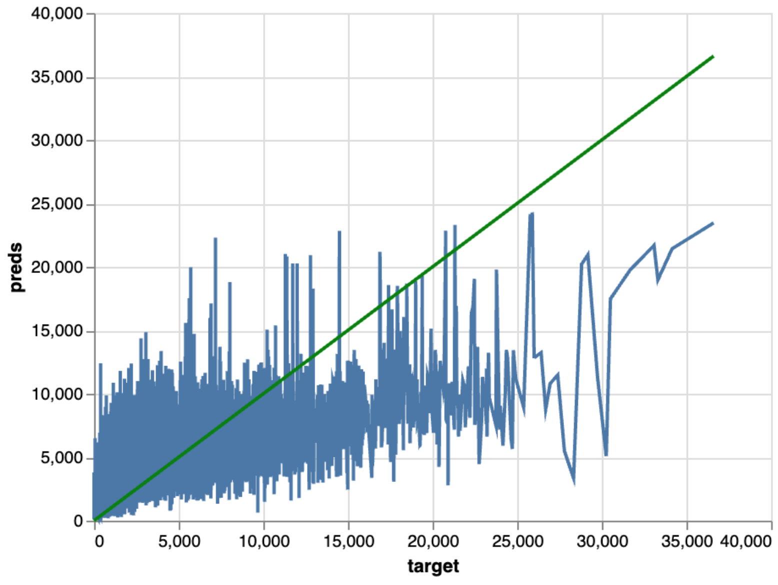


Figure 5. Testing Set Target vs Predictions

Breaking down the RMSE to different categories in the testing set, the 3 categories that have the lowest error are: category 0 – gas_transport, category 10 – entertainment and 7 - food_dining

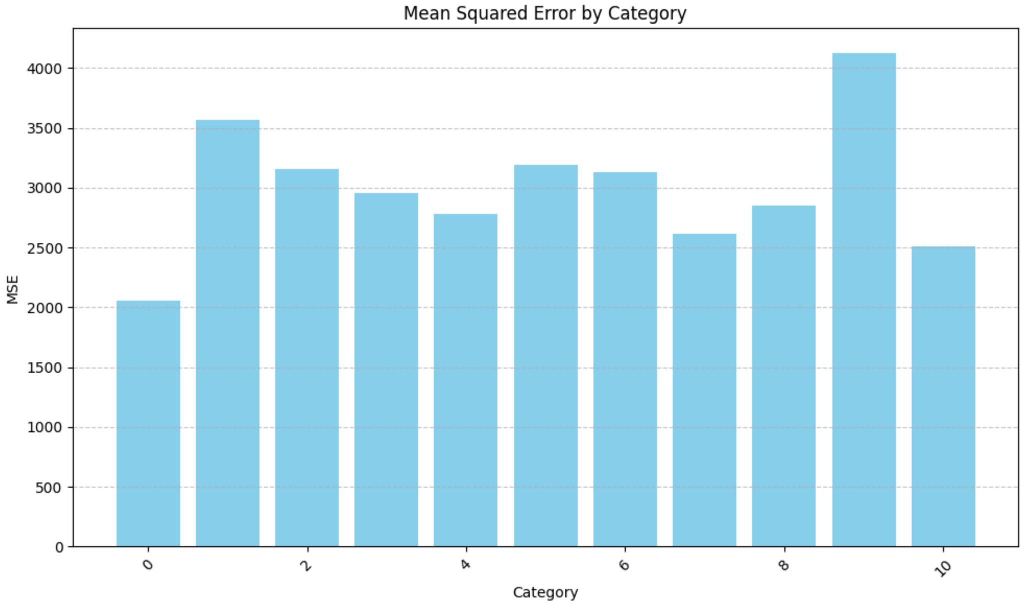


Figure 6. *RMSE by Category*

XGBoost Model

Following the implementation of the XGBoost model for detecting fraudulent transactions, we can assess that the model we chose for this case can flawlessly distinguish between fraudulent and legitimate transactions across all thresholds in the test dataset, which is complemented by Figure 7, where it is possible to appreciate an almost perfect classifier performance with an Area Under the Curve (AUC) of 1.00.

While an AUC of 1.00 is ideal, such a perfect performance might suggest overfitting, mainly if the model performs differently on new, unseen data. In future improvements and developments of the model, it is crucial to validate the model on a separate validation set or use cross-validation to ensure that it generalizes well beyond the training data, as it is essential to review the training process and data handling to confirm there are no issues such as data leakage or biases that could artificially inflate performance metrics.

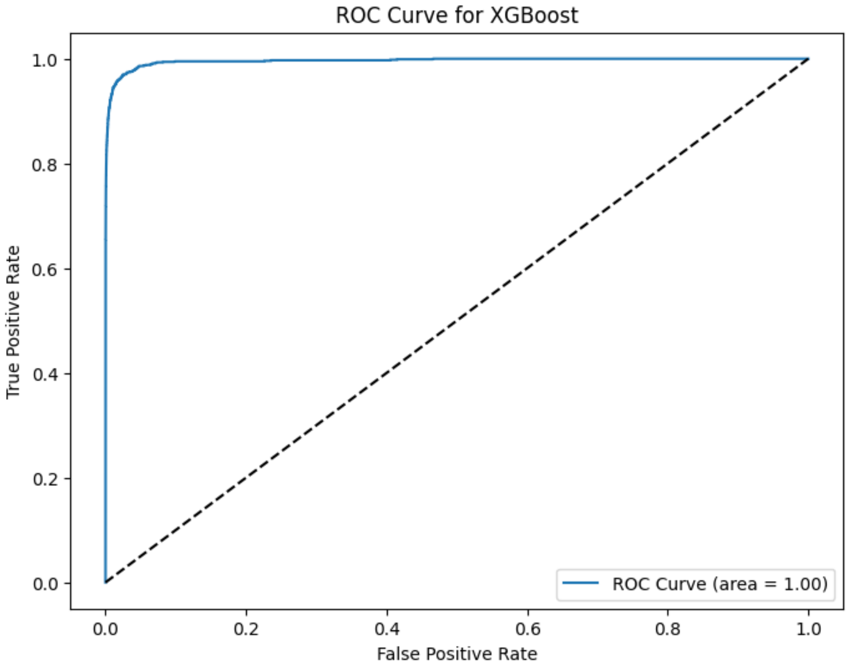


Figure 7. *ROC Curve for XGBoost.*

On the other hand, when carrying out the confusion matrix of the model to examine the classification results in terms of fraudulent transactions, we can note from Figure 8 that there are 848,290 true negatives, indicating a high accuracy in identifying legitimate transactions. However, it also shows 2,884 false positives, which are legitimate transactions incorrectly classified as fraudulent, and 124 false negatives, where fraudulent transactions were not detected. The model correctly identified 883 fraudulent transactions (true positives).

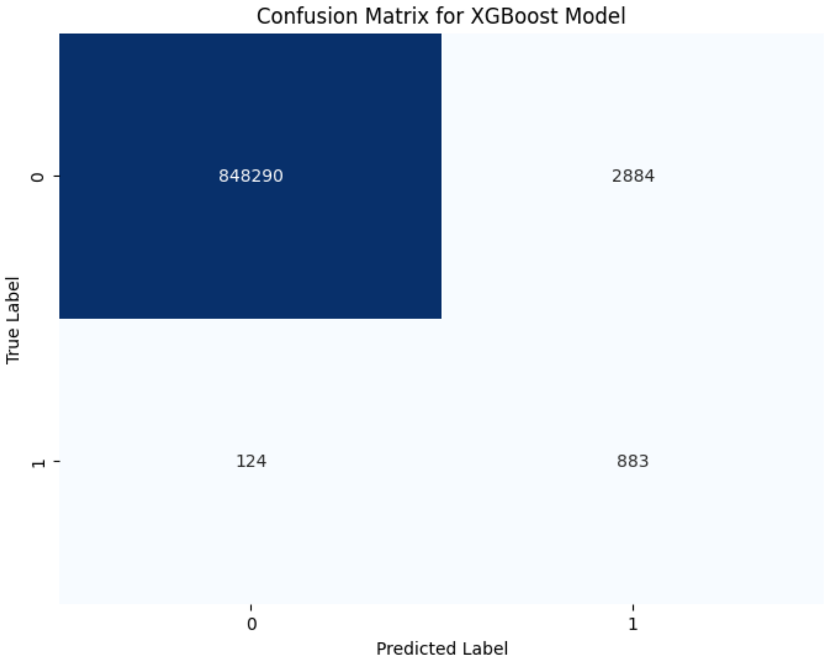


Figure 8. Confusion Matrix for XGBoost model.

There is a high recall of approximately 93%, which denotes the model's ability to capture most fraudulent transactions effectively. However, precision is notably lower at around 61%, reflecting that only a small proportion of transactions flagged as fraudulent are genuinely fraudulent. This imbalance results in an F1-score of about 0.684, indicating a need for improvement in balancing precision with recall to enhance overall model performance.

Cost-sensitive learning must be implemented to make a future transition from a development environment to a robust production environment, where higher penalties are assigned to misclassifications of the fraudulent class might help in reducing false negatives and improving precision and once in production, the model should be continuously monitored for performance drift and retrained with new data; this is crucial as fraudsters constantly change their strategies, and the model must adapt.

Isolation Forest Anomaly Detection

Advanced visualisation techniques were used to analyse the isolation forest model.

The t-SNE visualisation, **Figure 9**, clearly shows distinct data groups, with average points in blue and anomalies in green. This visual separation supports the effectiveness of the Isolation Forest in isolating anomalies from normal data. Anomalous groups appear consistently at the margins of the data clusters, as these points are less frequent and differ significantly from the common patterns. There are still points indistinguishable from each other, suggesting that this model has room for improvement.

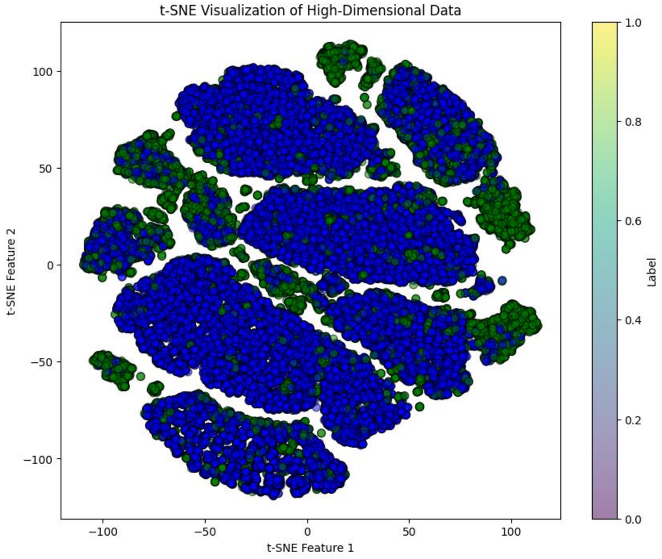


Figure 9. *T-SNE Visualization.*

The histogram in **fig 10** shows the distribution of anomaly scores with a marked threshold. Scores to the left of the red threshold indicate anomalies. Most scores accumulate around zero and to the right, suggesting that most of the observations are normal. Those crossing to the left of the threshold represent the anomalous points. This distribution is a good indicator of the sensitivity of the model and its ability to distinguish between normal and atypical behaviour. Still, the model may be flagging too many points as anomalous.

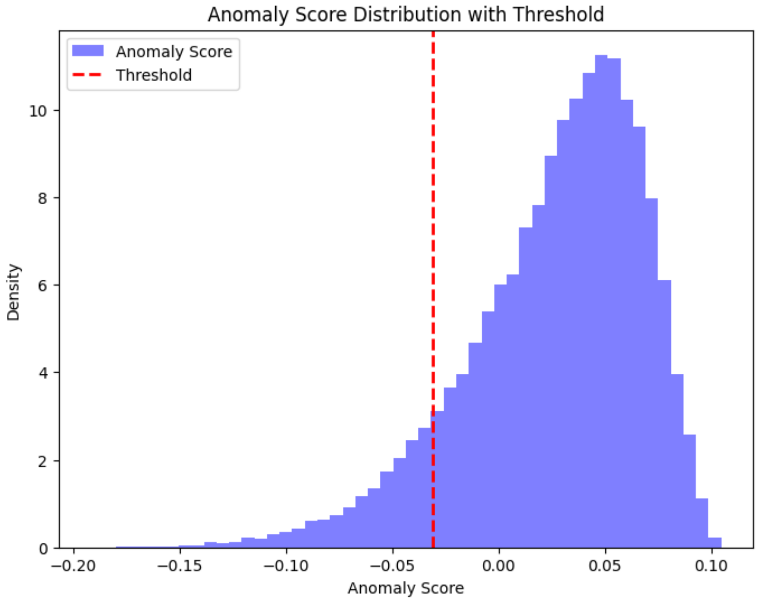


Figure 10. *Anomaly score distribution with score*

This bar chart shows the proportion of points labelled as normal versus anomalous. As can be seen, most of the points are labelled as normal. However, it is still observed that anomalies make up a large proportion of the database, which reinforces the analysis that improvements to the algorithm should be sought.

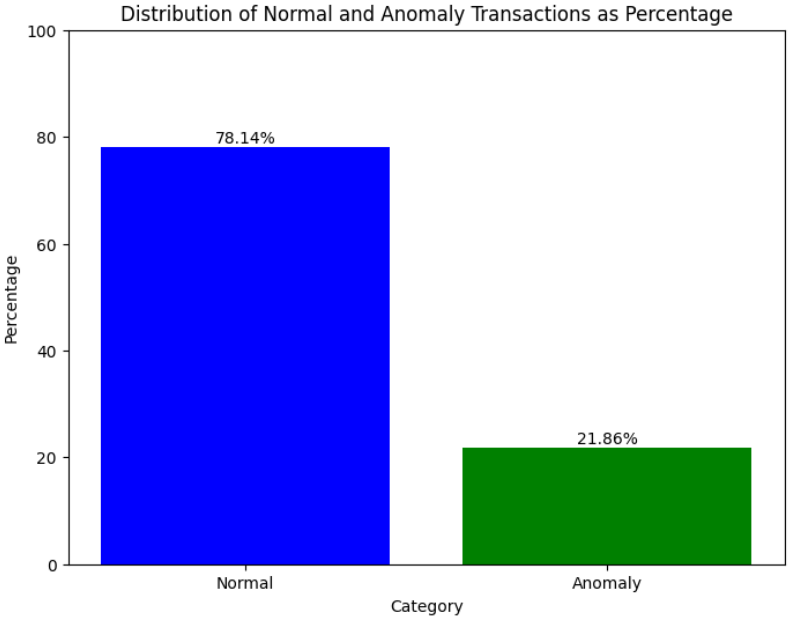


Figure 11. Normal and anomaly point distribution.

In conclusion, the visual tests show a good performance of the model, which manages to separate anomalous vs. normal data clearly. However, the model still needs to be improved as it is classifying a large proportion of the database as anomalous (approximately 20%).

K-means Clustering

For Clustering the graph generated has the number of clusters on the left-right line and WCSS values on the up-down line. The Elbow Method looks for a point where the WCSS line bends like an elbow. This bend usually marks the best number of clusters for our data.

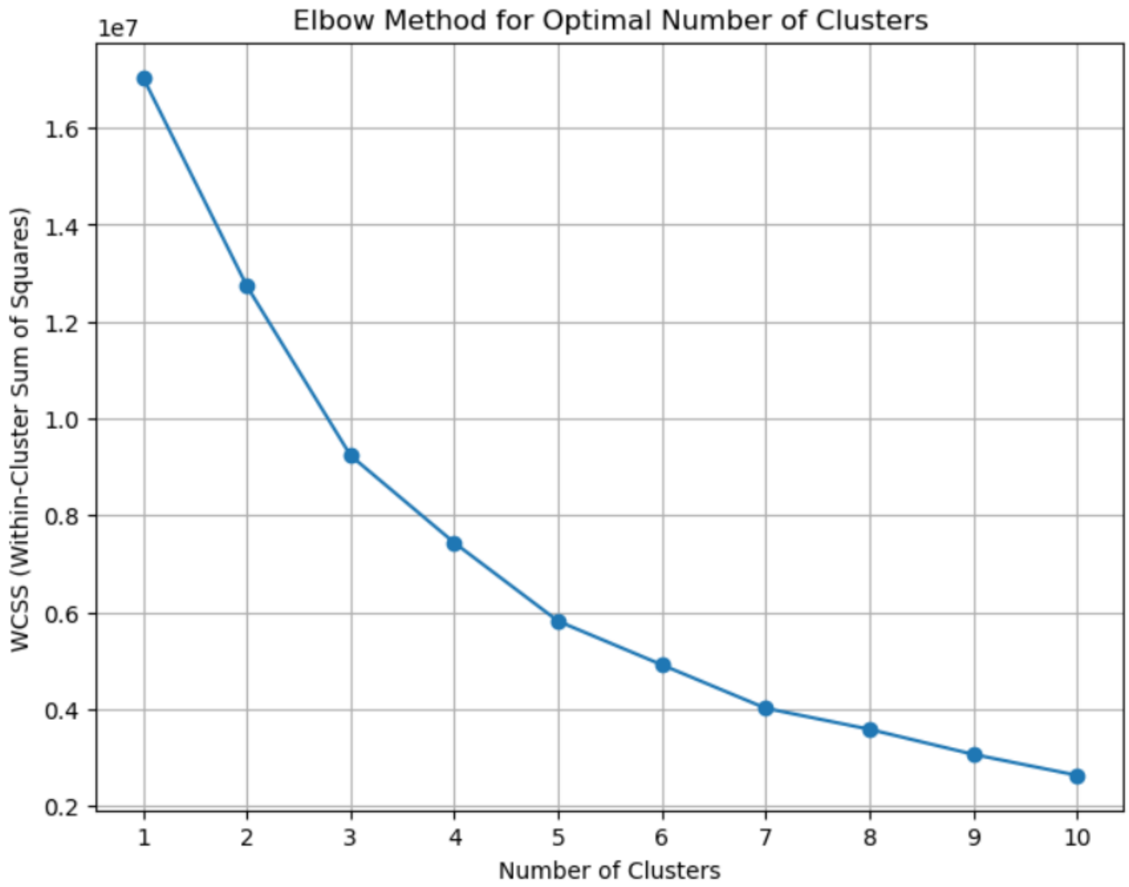


Figure 12. *Plotting Elbow Curve*

When we make the Elbow Method graph, we look for where the line stops dropping much when we add more clusters. This spot is usually the best number of clusters for our data. It's a balance between making sure our clusters are tight (compactness) but not too complex.

And again, the plot below shows how well groups of data are separated. Each colour represents a different group. The red dashed line shows the average score of how well the groups are separated. The score of 0.92 means the groups are very well separated.

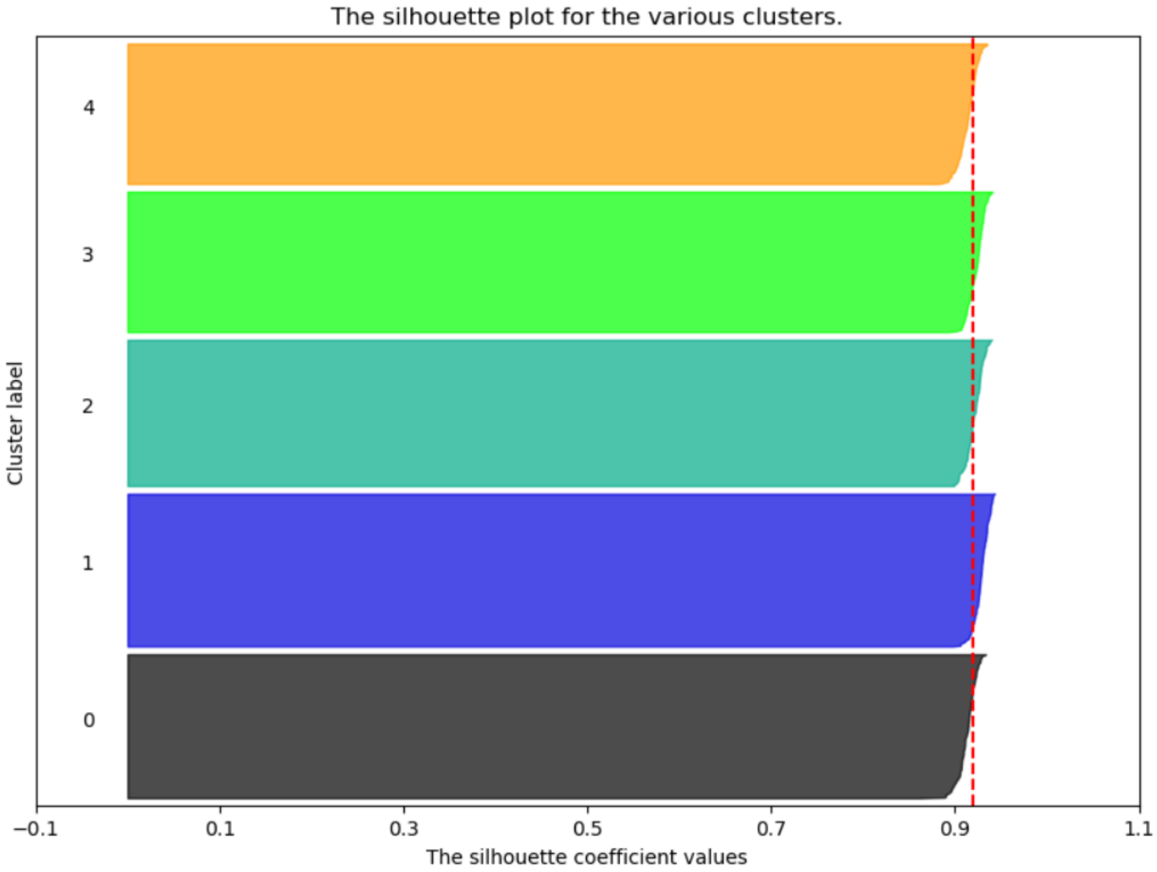


Figure 13. *Plotting Silhouette scores for clusters*

c. Business Impact and Benefits

Regression Modelling

By predicting cash outflows for different categories, the model helps customers gain insights into their spending habits, enabling them to adjust their budgets accordingly. This is particularly valuable for categories with the lowest prediction errors, such as gas and transportation, entertainment, and food and dining.

The new model can help customers save time and effort by previewing their spendings on each category. This convenience significantly enhances customer satisfaction and engagement with the bank's online platform.

However, since RMSE is not non-noticeable, and there are only basic customer and transaction data available for modelling. Important data related to spending habits like income and savings are missing. If DST could be provided with other account data which could be utilised into the regression modelling, the accuracy will be no doubt increased.

DST recommends The Bank to roll out the beta version to different customer groups for model refinements and select certain clients who have been banking for a long time and have regular incomes and savings into The Bank's savings account, as well as everyday transaction account or credit card account which gives the model enough transactional data to learn the spending behaviours from different customers groups.

XGBoost Model

The XGBoost model significantly boosts fraud detection within the business, achieving a high recall rate of 93%, effectively reducing financial losses by identifying most fraudulent transactions. However, the 61% precision rate indicates room for improvement as it also leads to many false positives; this affects operational efficiency and customer satisfaction due to the resources needed for manual reviews.

The model's success in preventing fraud translates into considerable annual savings based on average transaction values. Enhancements focusing on improving precision could further reduce operational costs and enhance customer experiences. Continuous refinement and adaptation to emerging fraud patterns, supported by ongoing stakeholder feedback, will ensure the model's effectiveness and relevance in detecting and preventing fraud, maximizing return on investment.

Anomaly detection Isolation Forest

Implementing the Isolation Forest model aims to transform the detection of anomalies in our company, increasing customer satisfaction and reducing operational costs. With the results obtained, it is possible to equip the monitoring team with a tool that will give them early warning of possible anomalous cases for further investigation, only for transactions over 200.

The model results indicate that it is possible to use it in production. Still, with caution, as the model detects many data as anomalous, it is a future work to reduce the target transactions to be studied. In contrast, the model is improved to segregate the anomalous data better.

This initial project is to help the monitoring team automate the search for anomalous transactions.

Clustering

The clustering model helps the bank's marketing team send targeted emails based on customer spending habits. This increases engagement and conversions, saves money, and improves the efficiency of marketing. Receiving tailored offerings from businesses boosts customer happiness and loyalty. The bank can observe gains in marketing performance and more income by examining measures like open rates and conversion rates. Time is saved, and as the bank gathers more data, automated segmentation may grow. The methodology improves the bank's financial results, customer experience, and marketing effectiveness overall.

d. Data Privacy and Ethical Concerns

We make sure to follow strict rules about keeping customer information private, like GDPR in Europe and APPs in Australia. This is super important because it keeps our customers' trust and makes sure we're doing everything legally, no matter where they are (Petropoulos, 2021).

We also think about fairness and being right when using data. Sometimes, data can be biased, which means it might not be fair, especially for groups like Indigenous people in Australia. This could affect things like loans or other banking services. To fix this, we use smart methods like hiding data details and checking for biases regularly (European Parliament, 2020).

We're committed to being fair and clear with our data. We make sure our models work for everyone and listen to different experiences to make our models better and fairer for everyone.

6. Collaboration

a. Individual Contributions

Student Last Name	Student First Name	Student ID	Contributions
Guo	Sophia	14157150	• Organising team meetings • Taking meeting minutes and action items • Discuss the dataset and findings • Taking the lead in delivery timeline and final project report writing
Montoya	Santiago	24898381	• Developing the code to join all the files into one dataset. • Structuring parts of the written report • Discuss ideas for modelling and preprocessing for each business case. • Working closely with Valeria to address the best way to deal with the transaction dataset.
Kandikattu	Vishal	25413586	• Collaborative and attentive in completing the project • Completing the Report in time According to Team Needs • Sharing Thoughts About the Model During Meetings
Roman	Valeria	24896716	• Create a team working group to share progress and develop the final report. • Setting up all meeting via teams. • Timely completion of my written part of the final report. • Help in improving the final structure of the report.

b. Group Dynamic

Reflecting on our group dynamics during the project at the Bank, effective communication and precise distribution of responsibilities were critical to our success.

Our team used WhatsApp for quick updates and Microsoft Teams for detailed discussions, ensuring continuous connectivity and efficient issue resolution. Regular in-person and online meetings allowed us to stay actively involved.

Each team member focused on a specific area aligned with their skills, enhancing our project's depth and efficiency. We divided the responsibilities into the following: Sophia handled regression analysis, Santiago focused on anomaly detection, Valeria managed fraud classification, and Vishal worked on clustering. This clear division of labour streamlined our efforts.

To ensure effective teamwork, we embraced several strategies:

1- Regular Status Updates: Frequent discussions about progress and challenges kept the team informed and supportive.
2- Collaborative Problem Solving: We tackled technical challenges, such as managing large datasets and enhancing our collective problem-solving capacity.
3- Transparent Documentation: We document and shared all decisions and strategies, maintaining transparency and alignment among the team members.

These practices boosted our efficiency and fostered a collaborative environment crucial for overcoming obstacles and achieving our project goals.

c. Ways of Working Together

Our team used an agile way to manage our project, which means we stayed flexible and could change things quickly. This helped us make our project better step by step and keep up with what we needed to do by checking often.

Meetings were really important for us. We met in person every two weeks at De Vu Café Burwood or at UTS campus for big talks about our progress. We also had online meetings every week using Microsoft Teams to talk about urgent stuff and fix problems fast. This mix of meetings let all of us be part of the project no matter where we were.

We kept track of how we were doing by sharing documents and talking about them during meetings. When we had to make decisions, we all talked about it and agreed before deciding anything. We used tools like WhatsApp for quick chats and Microsoft Teams for formal meetings and sharing files.

d. Issues Faced

Model	Issue	Solution
Clustering	Calculating silhouette scores took too long.	Used a smaller sample of data to speed up calculations, used resample from sklearn.utils.
	Hard to create silhouette plots to show clustering quality.	Fixed errors in the plotting code, ensured all variables were defined and used silhouette_samples correctly.
XGBoost Model	Some models overfitted due to feature selection.	Improved feature selection and feature engineering.
	Could not identify patterns of the positive class (did not predict the frauds).	Applied oversampling techniques, enhanced feature engineering and improved hyperparameters.
Anomaly Detection	Original dataset was too large, causing delays in model training.	Filtered the dataset and extracted sufficient information for the models.
	Difficulty understanding model performance (unsupervised algorithm).	Got help from the team and used visualization techniques to address this problem.
Regression	Extra Tree regression model was overfitting.	Tried different parameters manually to assess performance first and make judgment calls.
	It took too long to run cross-validation for the whole dataset.	Aggregated data to quarter and month end to see differences and choose the right aggregation, which also reduced training time.

To solve these problems as a team, we held several short meetings to discuss each member's progress and solve the doubts and problems we had together. In the future, we can allocate fixed spaces for working together, focusing on give solution to the issues encountered.

7. Conclusion

Regression Modelling

The Extra Tree Regressor model was effective in predicting customers' future cash outflows, particularly for spending categories with monthly expenditures ranging from $0 to $10,000. The model's accuracy decreased for higher spending amounts, highlighting areas for further refinement. Integrating monthly spending transaction data with customer demographics (residency state， age, residency city population) improved the model’s ability to predict spending patterns, demonstrating the value of a comprehensive customer profile in financial forecasting.

This model will be able to enhance customer experience by providing an automated tool for budgeting and financial planning in online banking platform. DST will work with Legal, Banking Data and Platform Team to roll out the beta version to different customer groups for model refinements, targeting customers who has long term relationship with the bank that can offer rich transaction data.

XGBoost Model

The XGBoost model for fraud detection achieved a high recall rate of approximately 93%, effectively identifying most fraudulent transactions. However, a lower precision rate of about 61% resulted in numerous false positives, indicating a need for better balance to enhance operational efficiency and customer. Future efforts should focus on fine-tuning the decision threshold and further developing feature engineering. Additionally, continuous model updates and monitoring will be essential to adapt to new fraud tactics and ensure sustained effectiveness. Regular stakeholder feedback and adjustments based on real-world performance will help refine the model, making it more aligned with operational requirements and effective in practical scenarios.

Anomaly detection

The Isolation Forest model has shown promising results in initial evaluations, validating its potential for practical implementation. However, it should be noted that its primary function is the preliminary identification of possible anomalies. For future work, it is crucial to focus on refining the model's ability to isolate anomalies. This could be achieved through improved feature engineering, such as introducing new variables that provide more specific details about the location of transactions rather than just using coordinates. Exploring interactions between existing variables would also be beneficial in improving the distinction between normal and anomalous data.

In addition, employing feature-importance techniques will help transparently explain to the monitoring team why certain transactions are considered anomalous by the model. This approach requires increased computational capacity to accelerate model experimentation and optimisation, thus ensuring that improvements are practical and efficient. Implementing these strategies will improve the model's performance and strengthen confidence in its results, making it more effective in proactively detecting anomalies.

Clustering

The Data Science Team through clustering successfully helped the marketing team send personalized emails by segmenting customers based on their spending behaviours using k-means clustering. We identified the optimal number of customer groups using the Elbow Method and Silhouette scores, creating clear and meaningful segments. This analysis revealed key spending patterns and behaviours, providing valuable insights for targeted marketing.

The DST met its goal by enabling the marketing team to craft more personalized and engaging email campaigns, improving customer satisfaction and engagement. Stakeholders' requirements were fulfilled by delivering actionable insights that aligned with the marketing strategy.

8. References

Abreu, G. (2021). Predicting the amount spent by customers using ML. Medium. Retrieved May 24, 2024, from https://medium.com/geekculture/predicting-the-amount-spent-by-customers-using-ml-df4dddf5c1df.

Deshpande, K. (2019.). Predicting my expense from historical transactions. Medium. Retrieved May 24, 2024, from https://medium.com/@keyoordeshpande/prediting-my-expense-from-historical-transactions-e4d423f019b0

European Parliament. (2020). The ethics of artificial intelligence: Issues and initiatives. European Parliamentary Research Service.

Marutho, D., Handaka, S. H., Wijaya, E., & Muljono, N. (2018). The determination of cluster number at K-Mean using elbow method and purity evaluation on Headline News. 2018 International Seminar on Application for Technology of Information and Communication. *https://doi.org/10.1109/isemantic.2018.8549751*

Petropoulos, G. (2021). Challenges of the General Protection Regulation for the use of machine learning in the banking sector. *Journal of Financial Regulation and Compliance.*

Roul, R. K., & Sahay, S. K. (2016). Semi-supervised clustering using seeded-KMeans in the feature space of ELM. *https://doi.org/10.1109/indicon.2016.7838892*

Thongnim, P., Charoenwanit, E., & Phukseng, T. (2023). Cluster Quality in Agriculture: Assessing GDP and harvest patterns in Asia and Europe with K-Means and Silhouette scores. *https://doi.org/10.1109/iementech60402.2023.10423469*