# Machine Learning for Transaction Analysis Spending Prediction, Fraud Detection, Anomaly Identification, and Behavioral Clustering

Phuong A. Pham
Tran M. Hoang
Karan Bhutani
Arjun Khanijau
Department of Transdisciplinary Innovation, University of Technology Sydney

Machine Learning Algorithms and Applications

May, 2024

## Table of Contents

# LIST OF TABLES

# LIST OF FIGURES

# SECTION 1.

# Executive Summary

The project, conducted by Group 11 for the Machine Learning Algorithms and Applications course at the University of Technology Sydney, aimed to apply advanced machine learning techniques to analyze and extract insights from a comprehensive transactional dataset collected by Bank's over three years. The primary objectives were to enhance financial planning, detect fraudulent transactions, optimize marketing strategies, and provide proactive customer support.

## 1.1 Problem statement and context

The Bank has accumulated extensive transactional data from 2018 to 2022, including over 4.26 million transactions. The bank sought to harness this data to gain deeper insights into customer behavior, detect fraudulent activities, and enhance their service offerings. The highly imbalanced dataset, with only 0.1% fraudulent transactions, posed a significant challenge. The project aimed to address this challenge by developing robust machine learning models to extract actionable insights from the data.

## 1.2 Project objectives

2. Develop a Regression Model: Predict customers' total spending for the next month to aid in better financial budgeting.
3. Create a Classification Model: Identify fraudulent transactions to enhance the bank's fraud detection capabilities.
4. Implement Clustering Techniques: Segment customers based on spending behaviors to tailor marketing campaigns effectively.
5. Utilize Anomaly Detection Methods: Detect abnormal spending patterns to provide proactive customer support.

The significance of this project lies in its potential to transform Bank's operational efficiency and customer service by leveraging data-driven insights. Machine learning models enable the bank to stay agile in a competitive market, improve customer satisfaction, optimize financial products, mitigate risks, and maintain a competitive edge.

### 1.3 Achieved Results

2. Financial Planning Optimization: Hoang
   - Developed a K-Nearest Neighbors (KNN) regression model to predict monthly customer spending.
   - Achieved consistent RMSE and MAE scores, indicating reliable predictions despite low R-squared values.
3. Fraud Detection Enhancement: Pham
   - Created a KNN classification model to identify fraudulent transactions.
   - Achieved high accuracy, precision, recall, and F1 scores across training, validation, and test sets, effectively identifying fraudulent activities.
4. Proactive Customer Support: Karan
   - Utilized Isolation Forest and Autoencoder models for anomaly detection.
   - Identified transactions with high anomaly scores, providing insights into abnormal spending patterns and potential fraudulent activities.
5. Personalized marketing campaigns: Arjun
   - Implemented K-Modes clustering to segment customers based on spending behaviors.
   - Identified distinct customer segments, enabling targeted marketing strategies and personalized banking products.

### Conclusion

The project successfully demonstrated the value of integrating machine learning into Bank's operations. By predicting customer spending, detecting fraud, and segmenting customers, the bank can improve its financial planning, enhance fraud prevention, and deliver more personalized and proactive customer service. The outcomes of this project underscore the importance of data-driven decision-making in the banking industry, paving the way for future improvements and innovations.

# SECTION 2.

# Business Understanding

## 2.1 Business use cases

After three years of collecting transactional data, Nexus Bank has accumulated a vast amount of information related to customer interactions and transaction history. Leveraging machine learning algorithms presents several opportunities to extract actionable insights from this wealth of data.

As competition in the banking industry intensifies and customer expectations evolve, the ability to accurately predict customer behavior becomes increasingly critical. ML algorithms offer the capability to adapt and learn from new data continuously, enabling Nexus Bank to stay agile and responsive in a dynamic financial environment. Thus, after three years of data collection, integrating ML into our operations can empower Nexus Bank to enhance customer satisfaction, optimize financial products, mitigate risks, and maintain our competitive edge by delivering on the following use cases:

**Use Case 1: Financial Planning Optimization:** Nexus Bank can utilize the ML model to predict customers' total spending amount for the next month. By providing this forecast, the bank can help customers better budget their finances, assisting them in making informed financial decisions.

**Use Case 2: Fraud Detection Enhancement:** With the ML model's predictions, Nexus Bank can assist its Compliance Team in identifying fraudulent behavior by predicting whether a transaction is fraudulent or legitimate. This proactive approach can help mitigate financial risks and protect customers' assets.

**Use Case 3: Personalized Marketing Campaigns:** Leveraging the ML model's insights, Nexus Bank can tailor its marketing emails to groups of customers presenting similar spending behaviors. This targeted approach can enhance the effectiveness of marketing campaigns, leading to higher engagement and conversion rates.

**Use Case 4: Proactive Customer Support:** Nexus Bank can empower its Customer Support team to reach out to customers exhibiting abnormal spending patterns compared to their usual behavior. By detecting anomalies in real-time, the bank can offer proactive assistance and address potential issues before they escalate, enhancing customer satisfaction and loyalty.

## 2.2 Key objectives

**Objective 1:** Develop a regression model to predict customers' total spending amount for the next month, aiding in better financial budgeting.

**Objective 2:** Create a classification model to identify fraudulent behavior by predicting transaction authenticity.

**Objective 3:** Implement clustering techniques to tailor marketing emails to customer groups with similar spending behaviors.

**Objective 4:** Utilize anomaly detection methods to proactively reach out to customers with abnormal spending patterns.

**Objective 5:** Provide actionable recommendations to Nexus Bank based on insights from all four models to optimize financial budgeting, fraud detection, targeted marketing, and proactive customer support strategies.

Based on the above objectives, let's delve into the stakeholders involved in this project and their respective requirements-

**Stakeholder 1 - Customers:** Nexus Bank's primary stakeholders are its customers, who rely on our bank's services for their financial needs. These include individuals, businesses, and organizations seeking banking solutions such as savings accounts, loans, and investment opportunities. Our customers expect efficient and secure financial transactions, personalized services, and proactive support from Nexus Bank.

**Stakeholder 2 - Compliance Team:** They play a crucial role in ensuring adherence to regulatory requirements and preventing financial crimes such as fraud and money laundering. They will be relying on accurate predictions from models to identify suspicious transactions and mitigate compliance risks, safeguarding the bank's reputation and maintaining regulatory compliance.

**Stakeholder 3 - Marketing Team:** They are responsible for developing and executing marketing strategies to attract and retain customers. They plan to leverage machine learning insights to segment customers based on their spending behaviors and preferences, enabling targeted marketing campaigns that drive engagement and acquisition. By sending customized marketing communications, they aim to enhance customer satisfaction and loyalty.

**Stakeholder 4 - Customer Support Team:** They are tasked with providing assistance and resolving inquiries from customers. They wish to utilize machine learning algorithms to detect abnormal spending patterns and identify potential issues that may require intervention. By reaching out to customers proactively, they strive to enhance customer experience, address concerns promptly, and foster long-term relationships.

# SECTION 3.

## Data Understanding

The dataset contains information of 4,260,904 transactions from December 2018 to December 2022, including both legitimate and fraudulent transactions using credit cards of 1,000 customers. It includes 132 transaction files containing customer properties and a file containing personal information of 1,000 customers. These files are combined on two common columns (cc_num and acct_num).

**TABLE 3. 1:** Dataset information

| **Dataset** | **Transaction file** | **Customer files** |
|---|---|---|
| **Number of files** | 132 | 1 |
| **Name of files** | *transactions_0.csv* to *transactions_131.csv* | *customers.csv* |
| **Dimension** | (4260904, 10) | (1000, 15) |
| **Common columns** | cc_num<br>acct_num | |
| **Attributes** | cc_num<br>acct_num<br>trans_num<br>unix_time<br>category<br>amt<br>is_fraud<br>merchant<br>merch_lat<br>merch_long | ssn<br>cc_num<br>first<br>last<br>gender<br>street<br>city<br>state<br>zip<br>lat<br>long<br>city_pop<br>job<br>dob<br>acct_num |

**TABLE 3. 2:** Statistics for the dataset variables

| Name in the original dataset | Rename | Count | Unique | Top | Frequency |
|---|---|---|---|---|---|
| **first, last** | Card holder | 4,260,904 | 974 | Christopher Lee | 14549 |
| **cc_num** | Credit card number | 4,260,904 | 983 | 347208496498560 | 10912 |
| **acct_num** | Account number | 4,260,904 | 983 | 11546128003 | 10912 |
| **dob** | DOB | 4,260,904 | 962 | 1999-12-23 | 16747 |
| **age** | Age | 4,260,904 | 80 | 41 | 146553 |
| | Age group | 4,260,904 | 4 | 25-44 | 1853411 |
| **job** | Job | 4,260,904 | 505 | Computer games developer | 33,859 |
| **street** | Street | 4,260,904 | 983 | 2531 Diane Landing Apt. 510 | 10912 |
| **zip** | Zipcode | 4,260,904 | 939 | 1581 | 16,362 |
| **city** | City | 4,260,904 | 726 | Brooklyn | 74,282 |
| **state** | State | 4,260,904 | 51 | CA | 505,426 |
| **city_pop** | Population of city | 4,260,904 | 764 | 2504700 | 73,200 |
| **trans_num** | Transaction number | 4,260,904 | 4,260,904 | - | - |
| | Trans_year_month | 4,260,904 | 49 | 2022-12 | 260,373 |
| **category** | Category | 4,260,904 | 14 | shopping_pos | 420268 |
| **amount** | Amount | 4,260,904 | 81,218 | 1.02 | 1879 |
| **merchant** | Merchant | 4,260,904 | 21,977 | Smith PLC | 5737 |
| **is_fraud** | Is_fraud | 4,260,904 | 2 | 0 | 4255870 |

**TABLE 3. 3:** Statistics for the numeric variables

| | Mean | Std | Min | 25% | 50% | 75% | Max |
|---|---|---|---|---|---|---|---|
| **Age** | 46.5 | 17.63 | 17 | 32 | 44 | 58 | 97 |
| **Zipcode** | 51903.84 | 29967.2 | 1571 | 27505 | 49202 | 78704 | 99705 |
| **Population of city** | 303801.14 | 569623.09 | 105 | 20103 | 67593 | 242037 | 2906700 |
| **Amount** | 68.99 | 161.85 | 1.0 | 9.1 | 44.49 | 81.58 | 41300.53 |

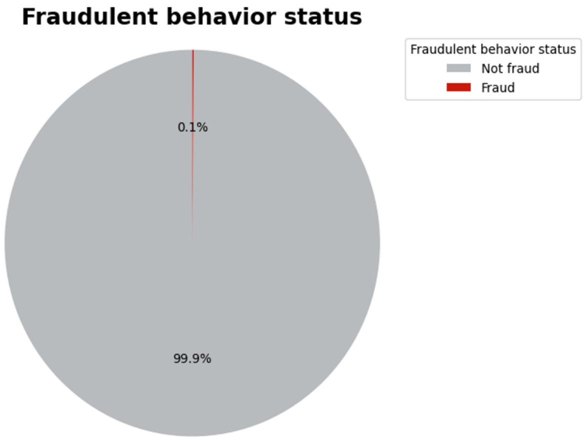


**FIGURE 3. 1:** Transaction class distribution

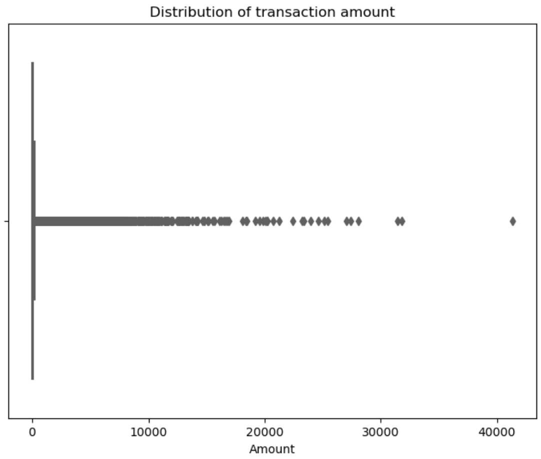


**FIGURE 3. 2:** Distribution of transaction amount

Main insights:

- There is a total of 2,261,904 transactions. In which, there are 4,255,870 valid transactions (99.9%), and 5,034 fraud cases (0.1%).
- The dataset is highly imbalanced.

**TABLE 3. 4:** Overview of transaction categories

| No. | Category | Number of transactions | Proportion of fraudulent |
|---|---|---|---|
| 1 | shopping_pos | 420,268 | 0.15% |
| 2 | home | 403,237 | 0.03% |
| 3 | grocery_pos | 401,000 | 0.29% |
| 4 | kids_pets | 372,055 | 0.03% |
| 5 | gas_transport | 365,233 | 0.11% |
| 6 | food_dining | 320,989 | 0.03% |
| 7 | entertainment | 312,478 | 0.05% |
| 8 | shopping_net | 307,013 | 0.38% |
| 9 | personal_care | 300,730 | 0.05% |
| 10 | misc_pos | 281,977 | 0.06% |
| 11 | health_fitness | 265,986 | 0.03% |
| 12 | misc_net | 191,290 | 0.34% |
| 13 | grocery_net | 181,614 | 0.05% |
| 14 | travel | 137,034 | 0.05% |

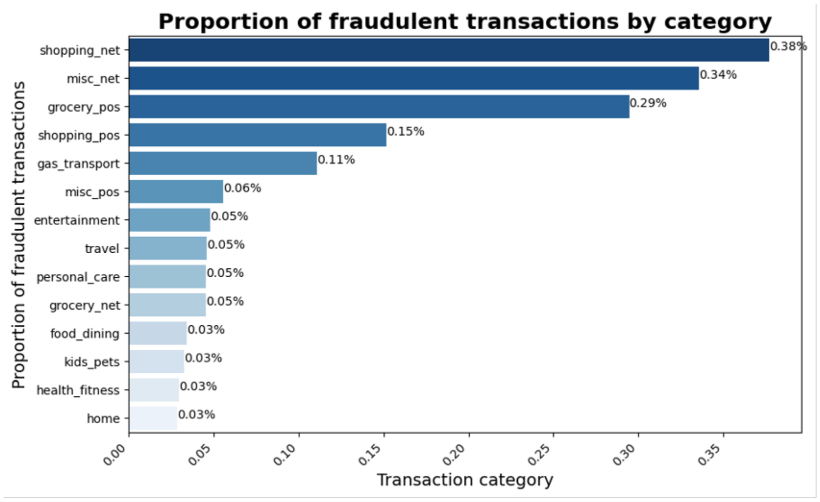


**FIGURE 3. 3:** Proportion of fraudulent transactions across various categories

Main insights:

- ***shopping_pos*** is the category with the highest amount of transactions, yet this category has a relatively low proportion of fraudulent transaction at only 0.15%.
- Category ***home*** has a significant number of transactions (403,237), yet the proportion of fraudulent transactions is notably low at 0.03%.
- Despite having fewer transactions (307,013), the proportion of fraudulent transactions of ***shopping_net*** category was the highest among all categories at 0.38%. Miscellaneous online transactions (***mics_net***) also have high rates of fraud despite of few transactions (191,290). This indicates that online shopping transactions and miscellaneous online transactions are more vulnerable to fraud.

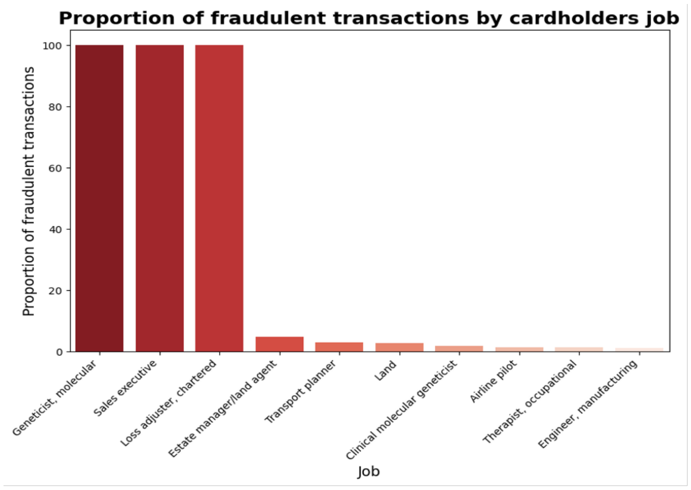


**FIGURE 3. 4:** Proportion of fraudulent transactions by cardholder job

**TABLE 3. 5:** Top 5 jobs with the highest proportion of fraudulent transaction

| | Job | Number of *fraud* transactions | Number of transactions | Proportion of fraudulent |
|---|---|---|---|---|
| 1 | Geneticist, molecular | 12 | 12 | 100% |
| 2 | Sales executive | 7 | 7 | 100% |
| 3 | Loss adjuster, chartered | 7 | 7 | 100% |
| 4 | Estate manager/land agent | 18 | 376 | 4.79% |
| 5 | Transport planner | 11 | 369 | 2.98% |

**TABLE 3. 6:** Top 5 jobs with highest occurrence of fraudulent transaction

| | Job | Number of *fraud* transactions | Number of transactions | Proportion of fraudulent |
|---|---|---|---|---|
| 1 | Computer games developer | 53 | 33,859 | 0.16% |
| 2 | Teaching laboratory technician | 53 | 14,251 | 0.37% |
| 3 | Agricultural consultant | 50 | 13,121 | 0.38% |
| 4 | Accountant, chartered public finance | 45 | 26,965 | 0.17% |
| 5 | Clinical molecular geneticist | 41 | 2,216 | 1.85% |

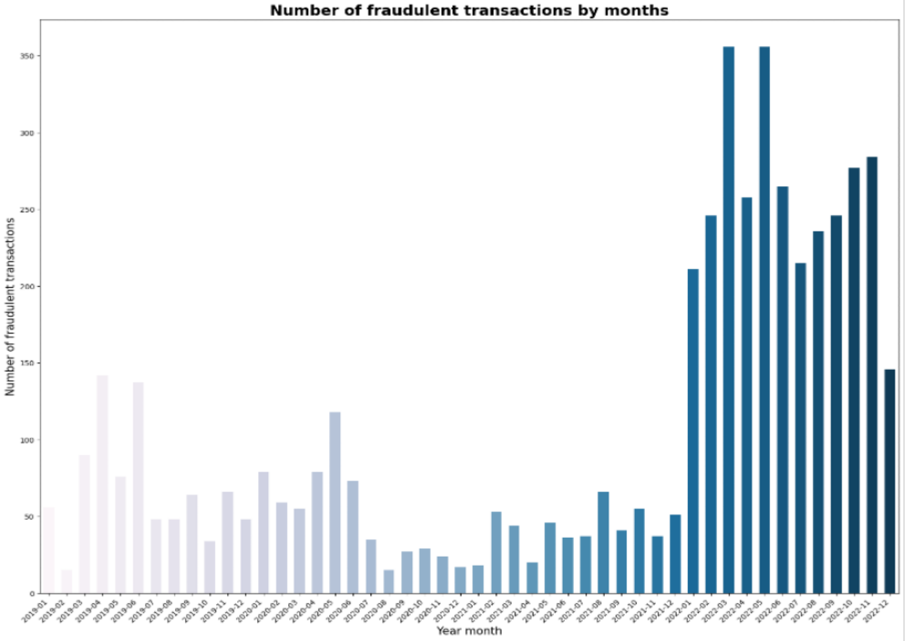


**FIGURE 3. 5:** Number of fraudulent transactions by month

Insights: The number of fraudulent transactions increased significantly at the beginning of 2022, and peaked at March and May 2022 with nearly 350 cases per month.

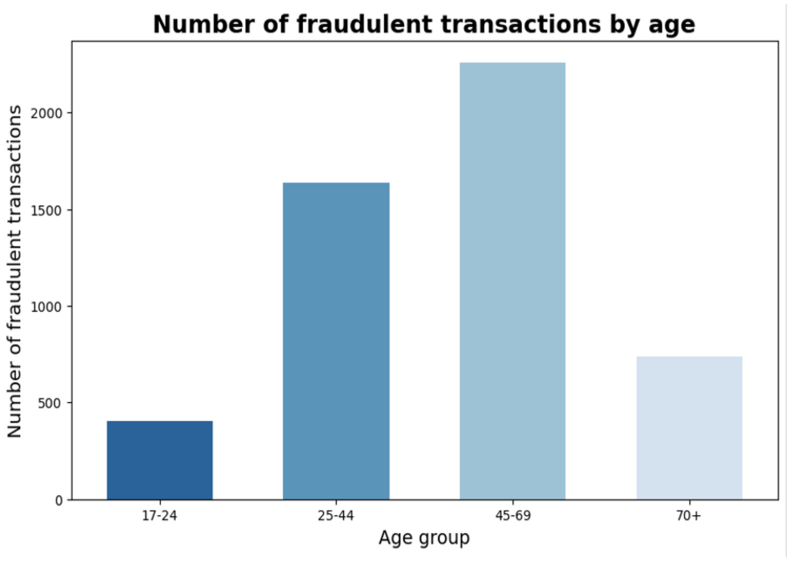


FIGURE 3. 6: Number of fraudulent transactions by age

Insights:

- There is a significant concentration of fraudulent transactions among customers aged 45-69.
- Customers aged 25-44 are also vulnerable to fraudulent activities.

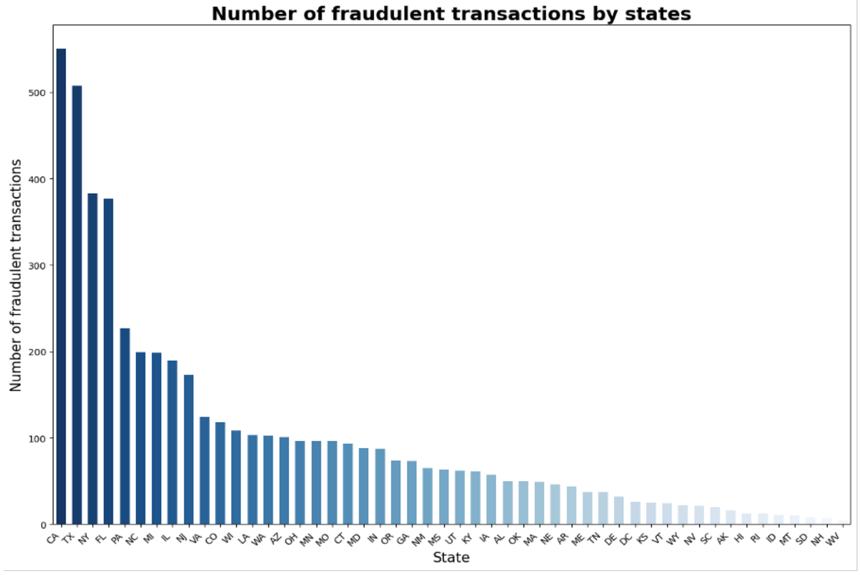


FIGURE 3. 7: Number of fraudulent transactions by states

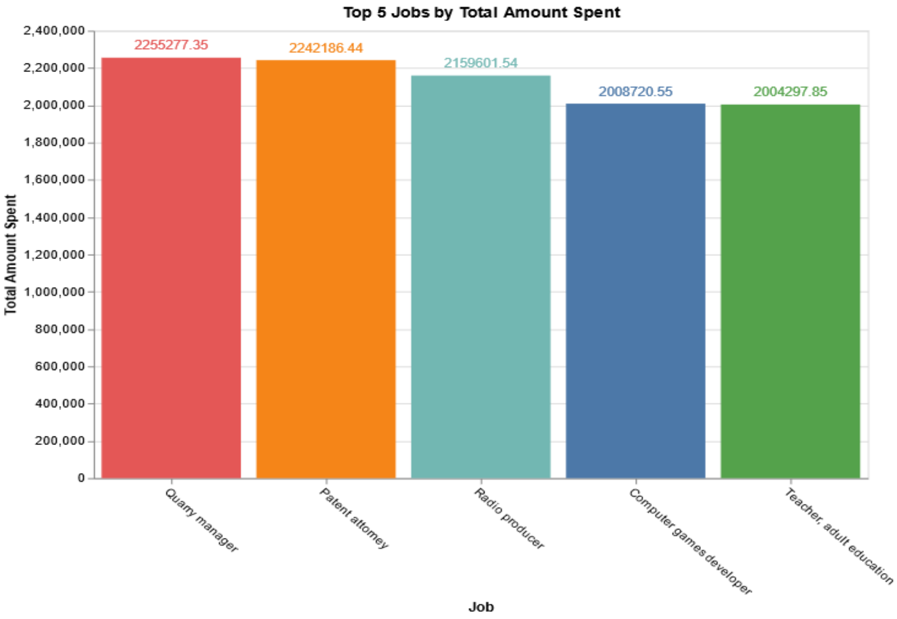


**FIGURE 3. 8:** Top 5 jobs by total amount spent

Insights:

- Customers who working as quarry manager, patent attoney, radio producer, computer games developer and teacher, adult education are likely spend more than other jobs.
- The bank should focus on Quarry Manager and Patent attorney what categories that they spend the most so start planning the marketing or advertising offer to them.

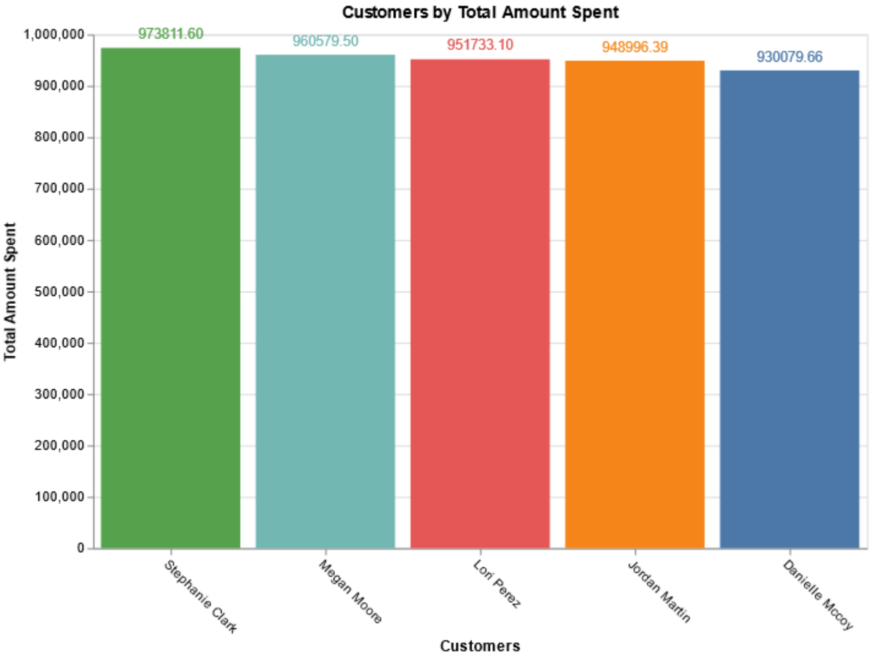


**FIGURE 3. 9:** Top 5 customers by total amount spent

Insights: The amount of top five customers who spend the most are closely to each other; the bank should continue explore and focus on these clients to keep them with the business and offer them more services or products.

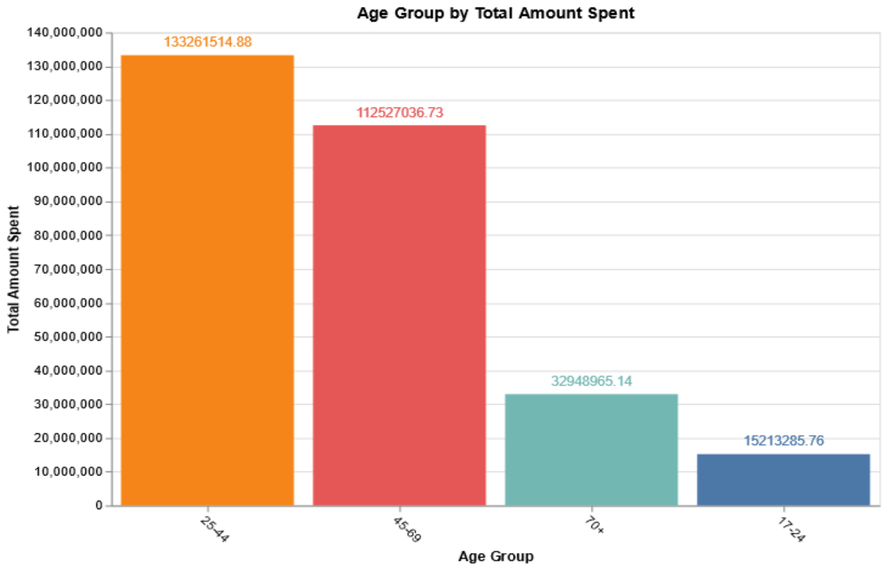


**FIGURE 3. 10:** Age group by total amount spent

Insights: The age group from 25 to 44 spends more than the other groups significantly because this age group is mainly in workforce; therefore, they have more opportunities and demands to spend their money.

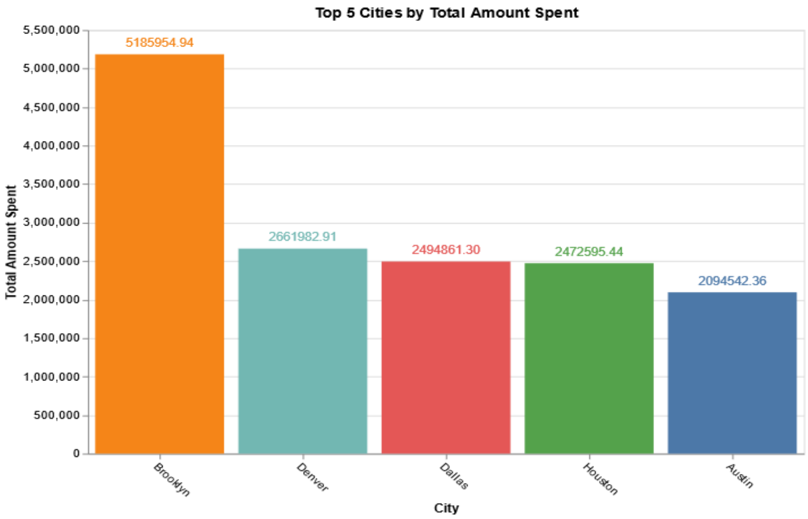


**FIGURE 3. 11:** Top 5 cities by total amount spent

Insights: The Brooklyn city shows that customers in this city spend their money for many activities more than other cities significantly. The investigation of the customers in this city including: age group, job, category should be highly considered.

# SECTION 4.

## Data Preparation

In the data pre-processing stage:

- We merged 132 transaction files and customers.csv
- Unix time is converted to datetime format for human-readable, making it easier to further manipulate information about time of transactions.
- We then renamed the columns which helps the audience understand the project more easily.
- A new column named ***card holder*** was generated by merging column ***first*** and ***last***.
- A new column named ***age*** was generated by extracting the year from ***dob*** and subtract the birth year from the current year (2024).
- A new column named ***age group*** was generated based on customers' age. There are four age groups: under 24, 25-44, 45-69, and over 70 years old.
- Reorder columns
- Unnecessary columns, including ***ssn, lat, long, gender, merch_lat, merch_long, first, last, transaction time*** are dropped to make training model process runs faster. Attribute ***gender*** should be dropped for ethical reasons, while attribute **ssn** should be dropped for unique identifier.

In the data processing stage:

- Check duplicate values
- Check NaN
- We used Isolation Forest method to handle outliers

  We set n_estimators = 200, which means the number of trees to be used in the ensemble are 200.

  There are 425,900 outliers.
- Encoding
- We split the dataset into train, validation and test set (64% for training, 16% for validating and 20% for testing)

  We have 2,454,402 transactions for training, 613,601 transactions for validating, and 767,001 transactions for testing.
- Handle imbalanced data.

# SECTION 5.

# Modelling

## 5.1 Regression model for predicting the amount that customers will spend in the next month

The regression model has been applied to this business use case, after finishing the Data Processing the dataset has 16 features, 3,835,381 transactions with 2,454,643 for training, 613,661 for validation, and 767,077 for testing.

Moreover, this project applied feature selection to select the most important features for the model; also, it helped to reduce the overfitting issue. The feature selection techniques that be used in this case were mutual_info_regression, and SelectKBest; moreover, 10 best variables have been selected in this project, including: **'card holder'**, **'credit card number'**, **'account number'**, **'age'**, **'age group'**, **'job'**, **'street'**, **'zipcode'**, **'city'**, **'population of city'**, **'category'**, **'merchant'**.

For choosing machine learning model, group 11 has tried five regression models namely:

- Linear Regression
- Ridge Regression
- Lasso
- ElasticNet
- K-nearest neighbors

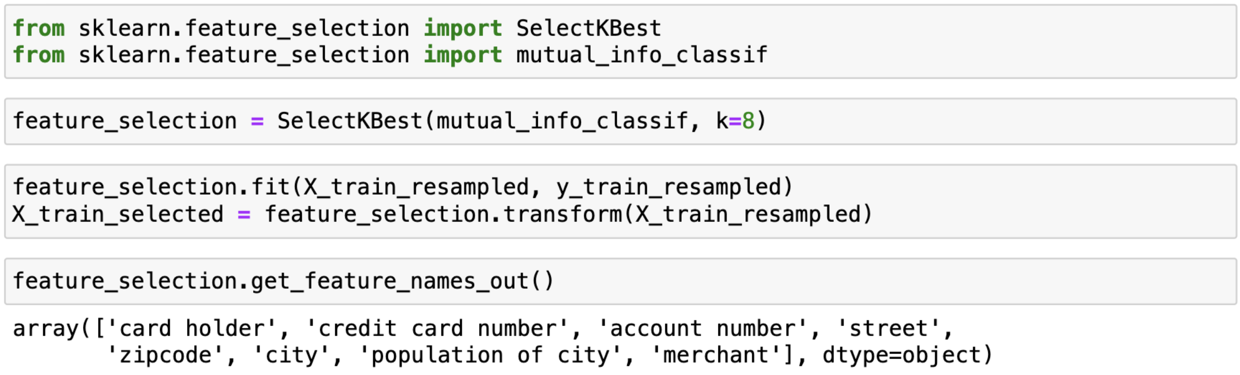
After training five models and KNN return the highest results as the best model; therefore, the team continue applying hyper-parameter tuning was Grid Search. In addition, this case was imported KNeighborsRegressor from sklearn.neighbors to train the KNN mode; also, import GridSearchCV from sklearn.model_selection for applying Grid Search technique.

During the tuning model, the 'n_neighbors' has been set in range [1, 5, 10, 20, 30, 50, 80, 100] and let the model tried one by one; at the end, the number 50 has generated as the best n_neighbors for this model; therefore, the final notebook just keep the parameter [50] only due to training the KNN model took a substantial amount of time of waiting.

### 5.2 Classification model for classifying fraudulent and non-fraudulent transactions

After the feature engineering stage, we have 16 attributes and 3,835,004 transactions, including 2,454,402 transactions for training, 613,601 transactions for validating, and 767,001 transactions for testing.

Feature selection plays a crucial role in improving the performance and efficiency of the model. Additionally, it may help reduce overfitting and faster the training process since this dataset is very large with nearly 4 million transactions. Because this dataset contains both numeric and categorical features, I chose SelectKBest and mutual_info_classif for feature selection. Eight features which have the most significant impact on the target variable are ***'card holder', 'credit card number', 'account number', 'street', 'zipcode', 'city', 'population of city',*** and ***'merchant'***.



```
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import mutual_info_classif
```

```
feature_selection = SelectKBest(mutual_info_classif, k=8)
```

```
feature_selection.fit(X_train_resampled, y_train_resampled)
X_train_selected = feature_selection.transform(X_train_resampled)
```

```
feature_selection.get_feature_names_out()
```

```
array(['card holder', 'credit card number', 'account number', 'street',
       'zipcode', 'city', 'population of city', 'merchant'], dtype=object)
```

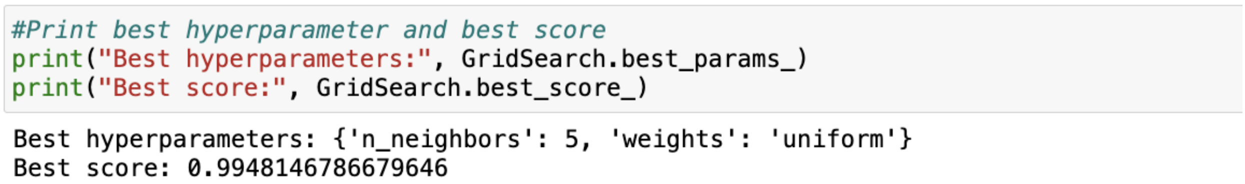
We utilize three machine learning models, including K-Nearest Neighbors, Decision Tree, and Random Forest. The results turned out that KNN is the best model for this dataset, hence, we will dive deep into this algorithm.

We used GridSearch Cross-Validation as a technique for fine-tuning hyperparameter.

First, we import KneighborsClassifier from sklearn.neighbors and GridSearch CV class from sklearn.model_selection to be used for hyperparameter tuning.

We then define parameter grids with n_neighbors = [5, 10], 'weights' = ['uniform'], and fit so that it will search through the set of parameter values to find which combination can lead to the best model performance and minimize overfitting.

This is the best hyperparameters and best score after applying GridSearch CV.



```
#Print best hyperparameter and best score
print("Best hyperparameters:", GridSearch.best_params_)
print("Best score:", GridSearch.best_score_)
```

```
Best hyperparameters: {'n_neighbors': 5, 'weights': 'uniform'}
Best score: 0.9948146786679646
```

## 5.3 Anomaly detection model for identifying customers with abnormal spending patterns

### Rationale for Using Isolation Forest and Autoencoders to identify anomalous transactions

### Overview

Identifying anomalous transactions in financial datasets is crucial for fraud detection. Two advanced methods, Isolation Forest and Autoencoders, are particularly effective due to their distinct approaches and complementary strengths.

### Isolation Forest

Isolation Forest is an ensemble-based anomaly detection method. Unlike traditional methods focusing on clustering or density estimation, it explicitly isolates anomalies. The algorithm isolates observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature. The key insight is that anomalies are few and different, making them easier to isolate.

### Advantages:

Efficiency: Isolation Forest is computationally efficient with a linear time complexity concerning the number of data points.

Scalability: It scales well with large datasets due to its linear time complexity.

Simplicity: Requires fewer parameters to tune compared to other anomaly detection algorithms.

### Feature Consideration:

Isolation Forest considers multiple features, such as transaction amount (amt), transaction time (time_stamp), merchant category (category), and other transactional attributes. This diverse set of features helps the model isolate anomalies more effectively, capturing unusual patterns that deviate significantly from normal behaviour.

Implementation steps:

1. Feature Selection: Pre-process and select relevant features from the transaction dataset.
2. Model Training: Train the Isolation Forest model on the selected features.
3. Anomaly Scoring: Compute anomaly scores for each transaction. Transactions with higher scores are flagged as anomalous.
4. Threshold Setting: Determine a threshold score to classify transactions as normal or anomalous.

### Autoencoders

Autoencoders are a type of neural network used for unsupervised learning, specifically for learning data representations. They consist of an encoder that compresses the input into a latent-space representation and a decoder that reconstructs the input from this representation. The reconstruction error (difference between the input and output) is used to identify anomalies. Anomalies tend to have higher reconstruction errors because they differ significantly from the patterns learned by the model during training.

### Advantages:

Feature Learning: Autoencoders can learn complex data representations and capture non-linear relationships between features.

Adaptability: They can adapt to various types of data distributions and are not limited to specific data shapes.

Robustness: Effective at handling high-dimensional data and can be fine-tuned to improve detection performance.

### Feature Consideration:

Autoencoders consider a wide range of transactional features, similar to Isolation Forest. The encoder part compresses these features into a lower-dimensional space, capturing the essential characteristics of normal transactions. The decoder then attempts to reconstruct the original transaction. High reconstruction errors indicate anomalies.

Implementation Steps:

1. Data Normalisation: Normalise the features to ensure efficient training.
2. Model Architecture: Define the encoder and decoder architecture with appropriate layers and activation functions.

3. Model Training: Train the autoencoder on the dataset to minimise the reconstruction error.
4. Anomaly Detection: Calculate reconstruction errors for each transaction and set a threshold to classify anomalies.

### Complementary Use

Combining Isolation Forest and Autoencoders leverages their complementary strengths. Isolation Forest provides an efficient, scalable solution with fewer parameters, while Autoencoders offer robust feature learning capabilities. Using both methods can enhance the overall detection performance, capturing a broader range of anomalies that might be missed by a single method.

Steps for combined use:

1. Apply Isolation Forest: Identify initial set of anomalies based on anomaly scores.
2. Train Autoencoder: Use the same dataset to train the autoencoder.
3. Cross-Validation: Validate the anomalies detected by both methods and refine thresholds for better precision and recall.
4. Comparison and Analysis: Compare anomalies detected by both methods to understand overlaps and differences.

### Conclusion

Isolation Forest and Autoencoders are powerful tools for anomaly detection in transactional data. Their theoretical foundations, coupled with practical advantages and complementary features, make them suitable for identifying fraudulent activities. Implementing these models requires careful consideration of the features and appropriate tuning to achieve optimal performance. Combining their outputs can provide a robust and comprehensive solution for anomaly detection in financial datasets.

## 5.4 Clustering based on similar spending behaviors using K-Modes and Hamming distance

K-Modes is a clustering algorithm similar to K-Means but designed for datasets with predominantly categorical features. It determines cluster centroids as the modes (most frequent values) of categorical features. K-Modes uses dissimilarity measures such as the Hamming distance rather than Euclidean distance to compute the dissimilarity between data points and centroids. These metrics are suitable for categorical data since they can handle non-numeric attributes.

## Justification for Using K-Modes Clustering for Customer Segmentation

### 1. Interpretability

- The centroids of clusters represent the most frequent category occurrences in each cluster, reflecting the predominant characteristics of customers within each cluster.
- This makes it easy to interpret and tailor marketing strategies to each segment based on their centroid profiles.

### 2. Deterministic

- K-Modes provides consistent and predictable clustering results due to its clear objective of minimizing intra-cluster dissimilarity based on categorical attributes, making it easier to understand and communicate the results to stakeholders.
- Unlike other clustering methods like DBSCAN or hierarchical clustering which may produce varying results depending on the parameters or may be sensitive to noise. K-Modes ensures reliable and actionable insights for customized marketing strategies.

### 3. Scalability

- K-Modes works well with large datasets, which is ideal for customer data that can be extensive.
- It efficiently handles a large number of data points, making it suitable for real-world marketing scenarios.

### 4. Customization Potential

- Allows for easy updating of clusters as new customer data comes in.
- Marketers can continuously refine and customize their strategies based on the evolving clusters.

## Limitations of K-Modes Clustering

**1. Sensitive to Initial Centroids**

- **Issue:** K-Modes clustering's performance can be sensitive to the initial placement of centroids.
- **Impact:** Depending on the initial centroids, the algorithm may converge to different local optima, leading to variations in cluster assignments.

**2. Assumption of Circular/Spherical Clusters**

- **Issue:** K-Modes assumes that clusters are spherical and of equal size.
- **Impact:** It may struggle with non-linear or irregularly shaped clusters, leading to suboptimal results if clusters have varying densities or shapes.

**3. Need to Specify Number of Clusters (K)**

- **Issue:** The number of clusters must be predefined, which can be challenging without prior domain knowledge or an automated method to determine the optimal K.
- **Impact:** Choosing an incorrect K may result in clusters that do not accurately represent underlying patterns in the data.

## Features Relevant for Clustering

### Demographic Features

**1. Age Group**: Age group can provide insights into how spending behaviors vary across different age demographics. We will be using age group instead of age as our segmentation goal is to identify broad age segments (e.g., young adults, middle-aged, seniors) and understand how spending behaviors differ between these groups.

**2. Job:** Occupation can influence spending habits and preferences.

### Transactional Features

**3. Category:** The category of transactions (e.g., groceries, entertainment, travel) can reveal spending patterns and preferences.

**4. Amount:** The transaction amount directly reflects spending behavior and can be a crucial feature for segmentation. To integrate with K-Modes, the values for 'amount' have been discretized into 3 categories: low, medium, and high, determined by their quartile positions.

**5. Merchant:** Different merchants may attract different customer segments, so including merchant information can help identify spending preferences.

## Selecting the optimal number of clusters (k)

We're employing the Elbow Method alongside WCSS (Within-Cluster Sum of Squares). WCSS quantifies the dispersion of data points within each cluster, crucial for categorical data like ours. The aim is to minimize WCSS, indicating higher similarity within clusters.

The Elbow Method aids in pinpointing the optimal cluster count by identifying the elbow point. This point represents the trade-off between the goodness of fit (minimizing WCSS) and the simplicity of the model (number of clusters).

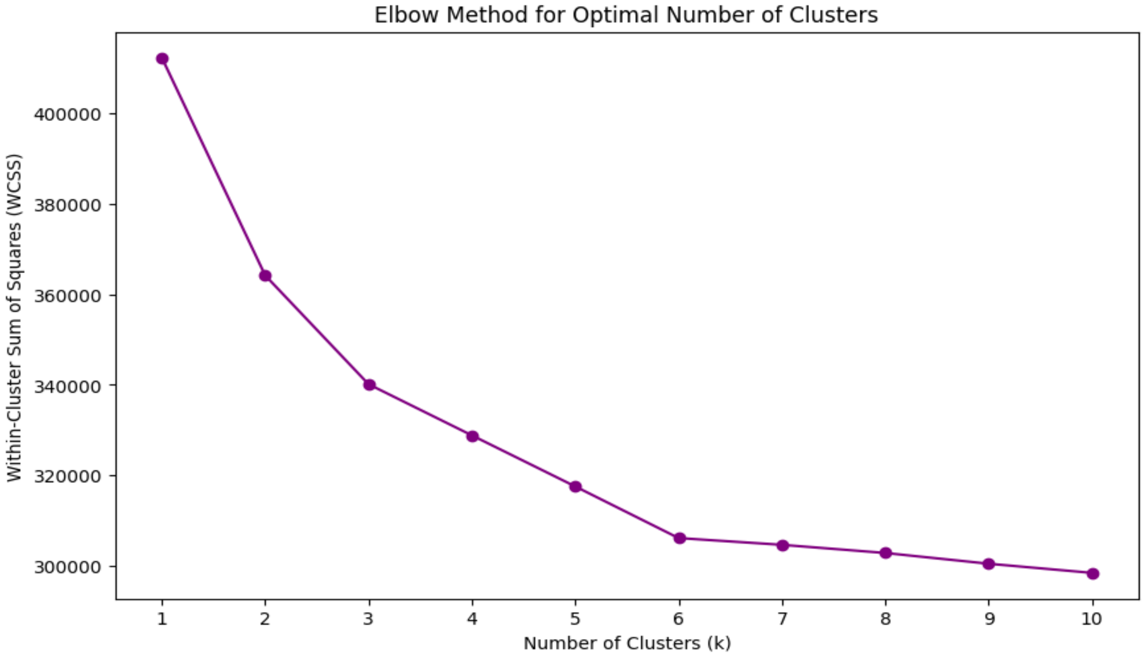


**FIGURE 5. 1:** Finding the optimal number of clusters (k)

k=6 is selected as the optimal number of clusters, as adding more clusters beyond this point does not significantly decrease WCSS.

# SECTION 6.

# Evaluation

## 6.1 Use case 1: Predict customers total spending amount for the next month

a. Evaluation Metrics

The business use case 1 is a regression task; also, to evaluate the efficiency of the regression model, there are some evaluation metrics that can be used, including:

**Root Mean Squared Error (RMSE)**

Root Mean Squared Error used to measure the difference between the predicted values and actual values; on the other word, the lower RMSE indicates the higher accuracy of the model (StatisticHowTo, 2024). In this business case of predicting the amount that customers will spend for the next month, the lower result of RMSE addresses the more accuracy of the model prediction.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(yi - \widehat{yi})^2}$$

**Mean Absolute Error (MAE)**

Mean Absolute Error measures the average absolute difference between the predicted values and actual values; on the other word, the lower RMSE indicates the higher accuracy of the model (Ahmed, 2023). In this business case of predicting the amount that customers will spend for the next month, the lower result of MAE addresses the more accuracy of the model prediction.

$$\text{MAE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|yi - \widehat{yi}|}$$

**R-squared ($R^2$)**

The R-squared is a number that demonstrates how well the dependent variable variation can be explained by the independent variable. Moreover, the range is go from 0 to 1 and if that close to 1 means a good fit of the model to the data (Fernando, 2024). In this business case of predicting the amount that customers will spend for the next month, the R-squared value close to 1 means the model has a good fit in predicting the amount spend based on the selected variables.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(yi \; - \widehat{yi})^2}{\sum_{i=1}^{n}(yi - y\bar{i})^2}$$

b. Results and Analysis

**TABLE 6. 1:** Regression model evaluation

| | Train Set | Validation Set | Test Set |
|---|---|---|---|
| RMSE | 0.24257 | 0.24745 | 0.24724 |
| MAE | 0.20006 | 0.20415 | 0.20398 |
| R Squared | 0.08001 | 0.04452 | 0.04378 |

According to the table above, the model performs well through the results of RMSE and MAE which consistent values; hence, it can be understood that the model is not overfitting or underfitting. However, the R Squared results of this model are lower than expected; therefore, seem like this model is not reliable enough for the prediction.

In this business use case, there are several regression models have been applied to figured out the best model for the prediction purpose, including:

- Linear Regression
- Ridge Regression
- Lasso
- ElasticNet
- K-nearest neighbors

All the models return the values of RMSE around 0.25 and only KNN model has the highest accuracy score; unfortunately, the R-squared value of all of them are similar which significantly low. After reviewing the results and discussing with team; group 11 has addressed some insights and further improvements:

- This project has applied two methods for feature scaling including MinMaxScaler and StandardScaler; this could be a factor that affect to the results and other methods such as RobustScaler, MaxAbsScaler or Normalizer should be considered.

- This project has applied both Hyper-parameter tuning Grid Search and Random Search; those two methods are return almost same results of RMSE and MEA.

c. Business Impact and Benefits

The machine learning model used in this business case is significantly helpful for the bank to understand their potential customers; also, with the prediction the amount that customers will spend next month, it assists the bank to prepare and consider to provide any benefits to their clients and keep them stay with the business. For example, those customers who spend the most money in buying technology then the bank can provide them the promotion such as discount coupon from the suppliers; by doing this, the organisation can keep the client continue using their services and open connections with many suppliers.

On the other hand, for those customers who spend not much money during the month, and the bank may consider to start some marketing activities for two reasons; firstly, attract those customers to continue using their product such as credit card with many promotions from the suppliers, and encourage them to spend money for buying products. Moreover, if they are spending their money but using other banks, then the marketing is a key to convince them to choose this service which provide many benefits.

Again, the machine learning model is significant useful for the bank in this situation; however, a further improvement the model with higher accuracy and reliability will support the business in evaluate the potential customers and attract more clients in the future.

## 6.2 Use case 2: Identify fraudulent behavior by predicting if a transaction is a fraud or not

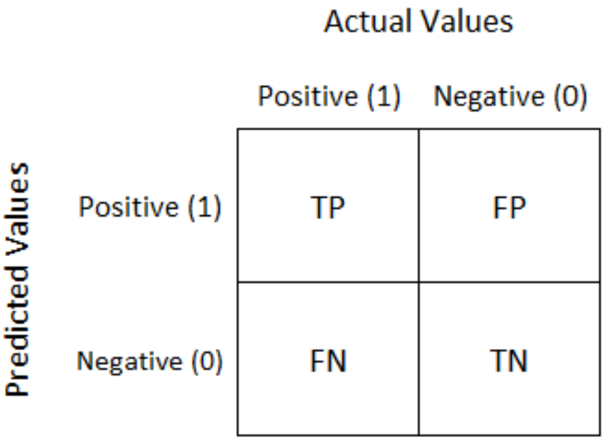
a. Evaluation Metrics

When dealing with classification models, there are several evaluation metrics which can be used to assess the efficiency of models:

**Confusion matrix:** This is a summary of predicted fraud transactions compared to the actual values.

- TP stands for True Positive predictions
- TN stands for True Negative predictions
- FP stands for False Positive predictions

- FN stands for False Negative predictions



| | | Actual Values | |
|---|---|---|---|
| | | Positive (1) | Negative (0) |
| Predicted Values | Positive (1) | TP | FP |
| | Negative (0) | FN | TN |

**FIGURE 6. 1:** Confusion matrix for binary classification problem

**Accuracy:** Accuracy measures how accurate a model's predictions are overall, determined by the proportion of correctly classified instances to the total number of instances.

In the context of credit card fraud detection, this evaluation metric indicates the proportion of correctly classified transactions (both fraudulent and legitimate) out of total number of transactions.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

**Precision:** Precision measures the proportion of true positive predictions among all positive predictions made by the model. It is calculated by the proportion of true positives to the sum of True Positives and False Positives.

In the context of credit card fraud detection, precision indicates the accuracy of the model in correctly identifying fraudulent transactions, minimizing non-fraudulent transactions which are classified as fraudulent by the automatic system.

$$\text{Precision} = \frac{TP}{TP + FP}$$

**Recall:** Recall measures the proportion of True Positive predictions among all actual positive instances. It is calculated as the ratio of True Positives to the sum of True Positives and False Negatives.

In the context of credit card fraud detection, this evaluation metric indicates the model's ability to capture all fraudulent transactions, minimizing false negatives which are fraudulent transactions classified as non-fraudulent.

$$\text{Recall} = \frac{TP}{TP + FN}$$

**F1 Score:** The F1 score is the harmonic mean of precision and recall, providing a balance between the two metrics. The F1 score reaches its best value at 1 and worst at 0.

$$\text{F1 score} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

**ROC curve**: ROC curve is a common tool with binary classification. The dotted line represents the ROC curve of a purely random classifier. The blue line represents the result of trained model. A good classifier stays as far away from the dotted line as possible toward the top-left corner (Polanitzer, 2021).

b. Results and Analysis

**TABLE 6. 2:** Classification model evaluation

| | **Train set** | **Validation set** | **Test set** |
|---|---|---|---|
| **F1 score** | 0.99712 | 0.99516 | 0.99317 |
| **Precision score** | 0.99465 | 0.99461 | 0.98642 |
| **Recall score** | 0.99961 | 0.99572 | 1.0 |
| **Accuracy** | 0.99711 | 0.99516 | 0.99312 |

The model performs well in correctly identifying both fraudulent and non-fraudulent transactions with high F1, precision, recall, and accuracy score across train, validation, and test sets.

While the precision score slightly decreased when we train test set, the value remains high, indicating that the majority of transactions classified as fraudulent are indeed fraudulent.

A high recall score indicates that KNN model effectively captures amost fraudulent transactions in the dataset.

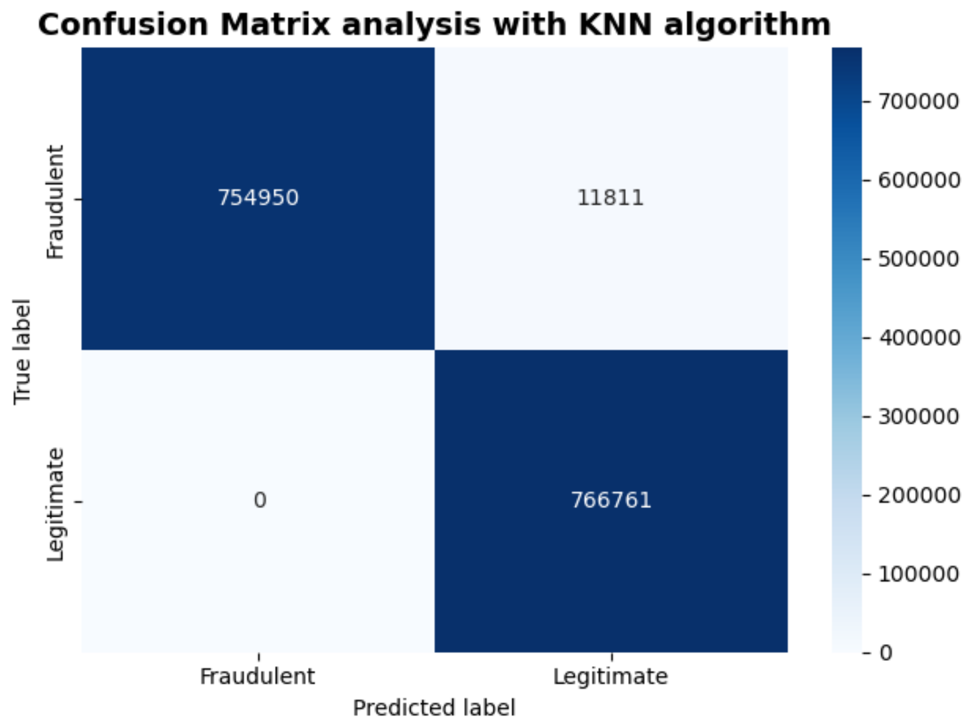


**FIGURE 6. 2:** Confusion matrix analysis with KNN model

**TABLE 6. 3:** Confusion matrix with KNN model

| | Meaning | KNN |
|---|---|---|
| True Positives | Transactions which were correctly predicted as fraudulent when these transactions are actually fraudulent. | 766,761 |
| True Negatives | Transactions which were correctly predicted as not fraudulent when these transactions are actually not fraudulent. | 754,950 |
| False Positives | Transactions which were incorrectly predicted as fraudulent when these transactions are actually not fraudulent. | 11,811 |
| False Negatives | Transactions which were incorrectly predicted as not fraudulent when these transactions are actually fraudulent. | 0 |
| Total correct predictions | | 1,521,711 |
| Total incorrect predictions | | 11,811 |

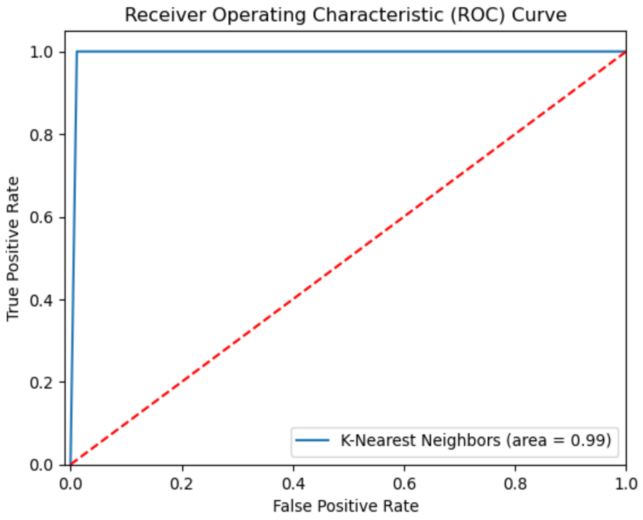


**FIGURE 6. 3:** AUC ROC Curve of KNN model

c. Business Impact and Benefits

Credit card fraud detection is a process of data exploration which will provide the best results in revealing and avoiding fraud. Personal information of users' transactions such as transaction number, account number, product category, amount, merchant, etc. are used to be run through a trained classification model which figures out patterns so that it can classify whether a transaction is legitimate or fraudulent.

False Positives represents a costly error where legitimate transactions are incorrectly flagged as fraudulent. Incorrectly predicting a transaction as fraudulent when it is actually legitimate may incur several costs and negative consequences for both the bank, the customers, and the technical infrastructure:

Bank: Incorrect classification can lead to increase operational costs since the bank has to allocate resources to review flagged transactions, verify their legitimacy, handle customer complaints. Another consequence could be missed revenue opportunities when customers get in trouble in completing a high-value purchase. The bank may also lose customer trust by flagging non-fraudulent transactions as fraudulent transactions.

Customers: One of the serious consequences would be customers inconvenience when conduct transactions. They expect efficiency and speed when purchasing items online, hence, a false positive alert may delay the check-out process. If the bank blocks their accounts since they suspend that transaction is fraud when it is indeed not a fraudulent transaction, customers have to contact the bank to resolve their issues, leading to time lost and inconvenience. They may perceive the bank over-cautious, and hesitate to use the bank's services.

Technical infrastructure: When a transaction is flagged as potentially fraudulent, SMS messages, emails, push notifications or codes are sent to customers which may cost the bank a small amount of money.

The project results may help the bank not only to reduce cost by minimizing False Positives, but also to enhance customer experience, help them to be less likely to encounter disruptions or frustration, leading to higher customer satisfaction and loyalty enforcement. To minimize false positives, it is necessary to analyse the characteristics of transactions misclassified as fraudulent, identify common patterns such as transaction amounts, transaction location, merchant, etc. Additionally, the Compliance team needs to examine the demographic and behavioral profiles of customers associated with false positive transactions. Factors which should be considered include frequency, spending habits, etc. It is data scientists' responsibility to investigate the root causes behind fraudulent transactions which are flagged as non-fraudulent by the model.

## 6.3 Use Case 3: Helping Customer Support team to reach out to customers with abnormal behaviors from their usual spending patterns (anomaly detection)

a. Evaluation Metrics

In unsupervised learning, particularly for anomaly detection, we lack labelled data to directly evaluate the performance using traditional metrics like accuracy, precision, recall, etc. However, there are several ways to assess the effectiveness of unsupervised models, such as Isolation Forest and Autoencoders, for identifying anomalous transactions:

### Evaluation Methods

1. **Reconstruction Error Analysis (for Autoencoders):**

- Reconstruction Error Distribution: Plot the distribution of reconstruction errors. Anomalies typically exhibit higher reconstruction errors compared to normal transactions.

- Threshold Setting: Determine an appropriate threshold for the reconstruction error to classify anomalies. This can be done using visual inspection or statistical methods like setting the threshold at a certain percentile of the error distribution.

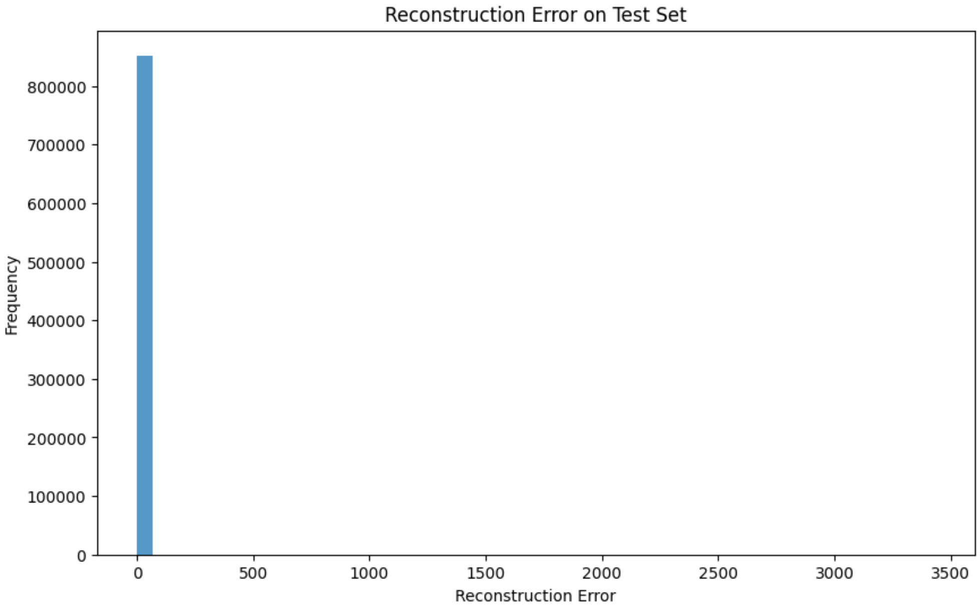


**FIGURE 6. 4:** Reconstruction Error (Autoencoders)

**Model Prediction:** The Autoencoder reconstructs the test data to mimic the original transactions.

- **Reconstruction error calculation:** The Mean Squared Error (MSE) between the original and reconstructed data is computed to quantify the reconstruction error.
- **Threshold Setting:** An anomaly detection threshold is established at the 95th percentile of the reconstruction errors. Transactions with errors above this threshold are flagged as anomalies.
- **Anomaly Identification:** Transactions exceeding the threshold are identified as anomalies.
- **Data Annotation:** These anomalies are added to the original test data for further inspection and validation.

2. **Anomaly Score Analysis (for Isolation Forest)**

- **Score distribution:** Analyse the distribution of anomaly scores generated by the Isolation Forest. Higher scores indicate a higher likelihood of being anomalous.

- **Threshold setting:** Set a threshold for anomaly scores to classify transactions as anomalous. This can be based on the score distribution or domain knowledge.

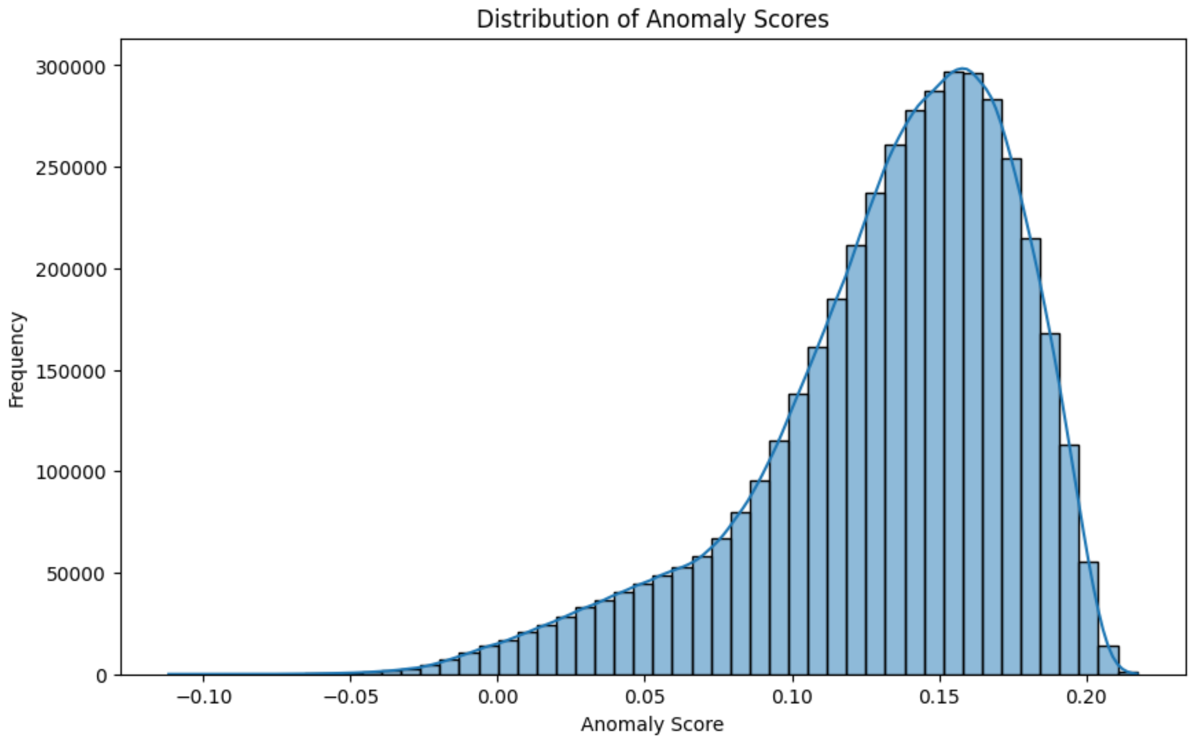


**FIGURE 6. 5:** Distribution of anomaly scores for every transaction

3. **Case studies:**

- **Investigate Individual Anomalies:** Conduct case studies on a few detected anomalies to understand the nature of these transactions. This involves examining the transaction details and context to verify if they are genuinely unusual.

b. Results and Analysis

Below is the Inlier v/s Outlier scatter plot. The points in red represent the anomalous transactions and blue represent the Normal transactions identified by the models above.

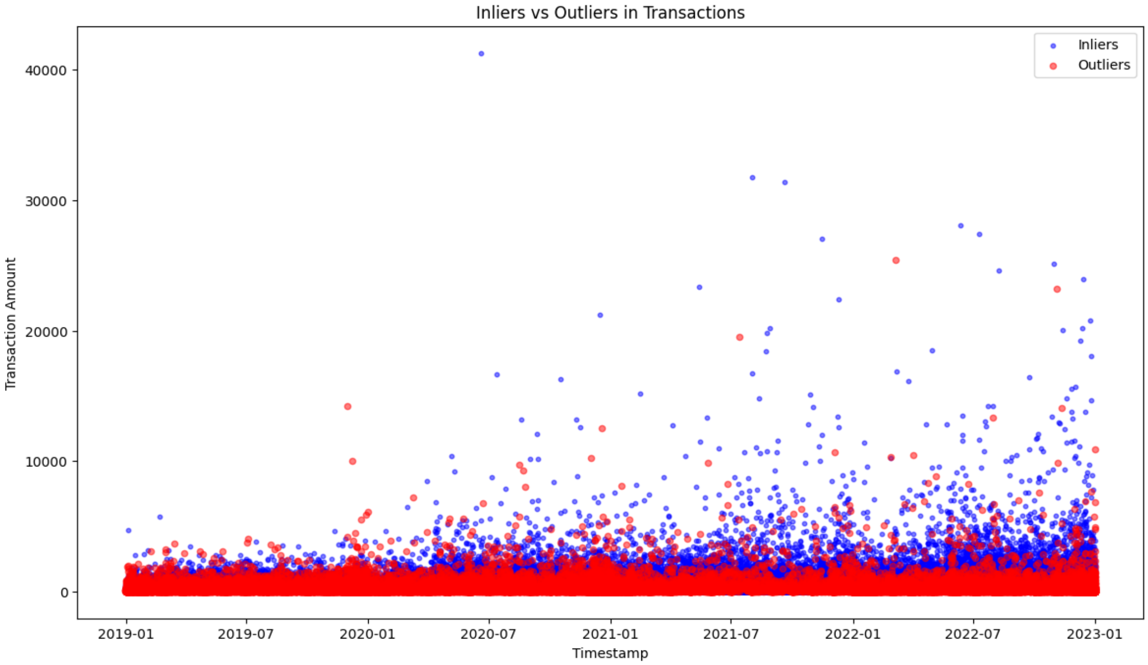


**FIGURE 6. 6:** Inlier v/s Outlier for Isolation Forest

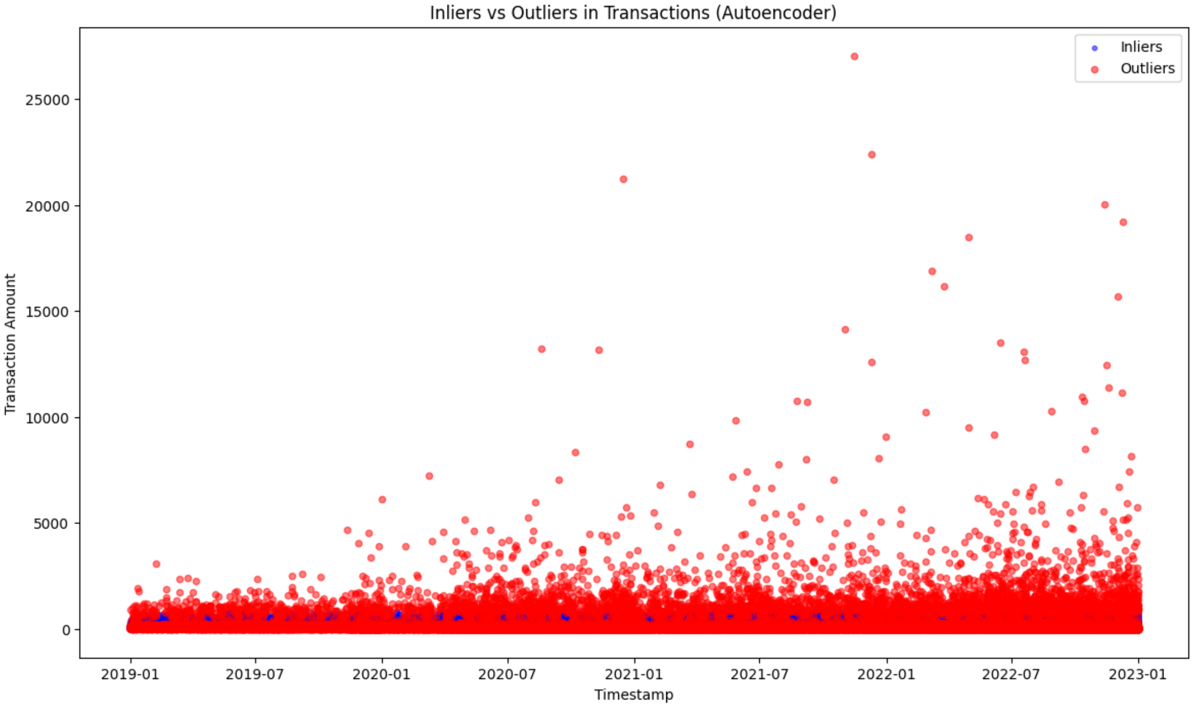


**FIGURE 6. 7:** Inlier v/s Outlier for Autoencoders

### Distribution of Anomalous Transactions by Amount

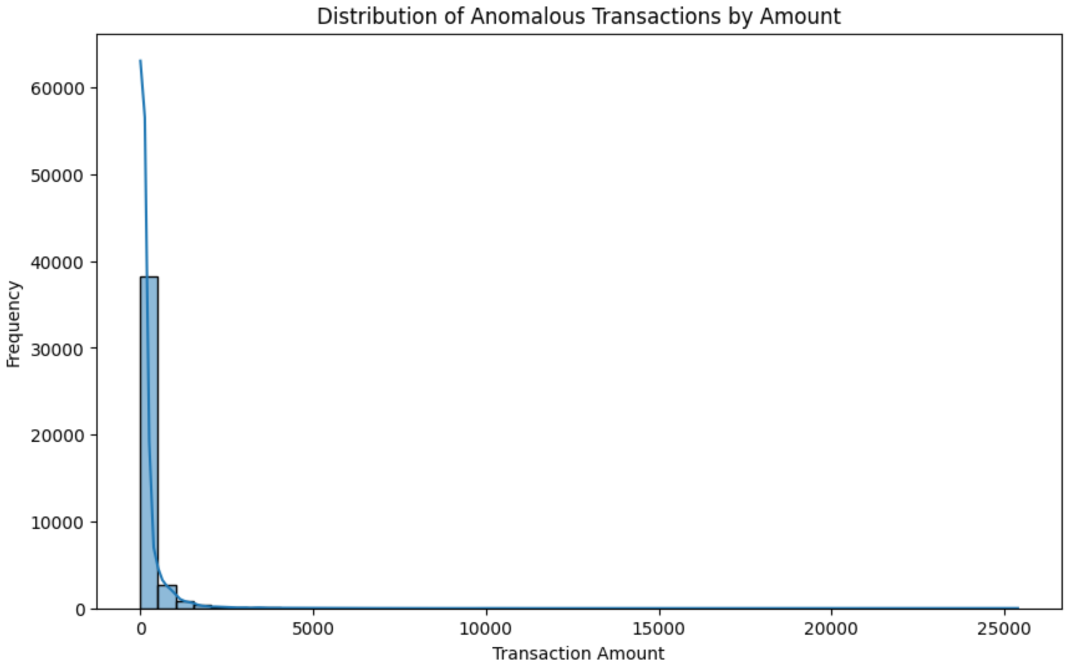


**FIGURE 6. 8:** Results for Isolation Forest

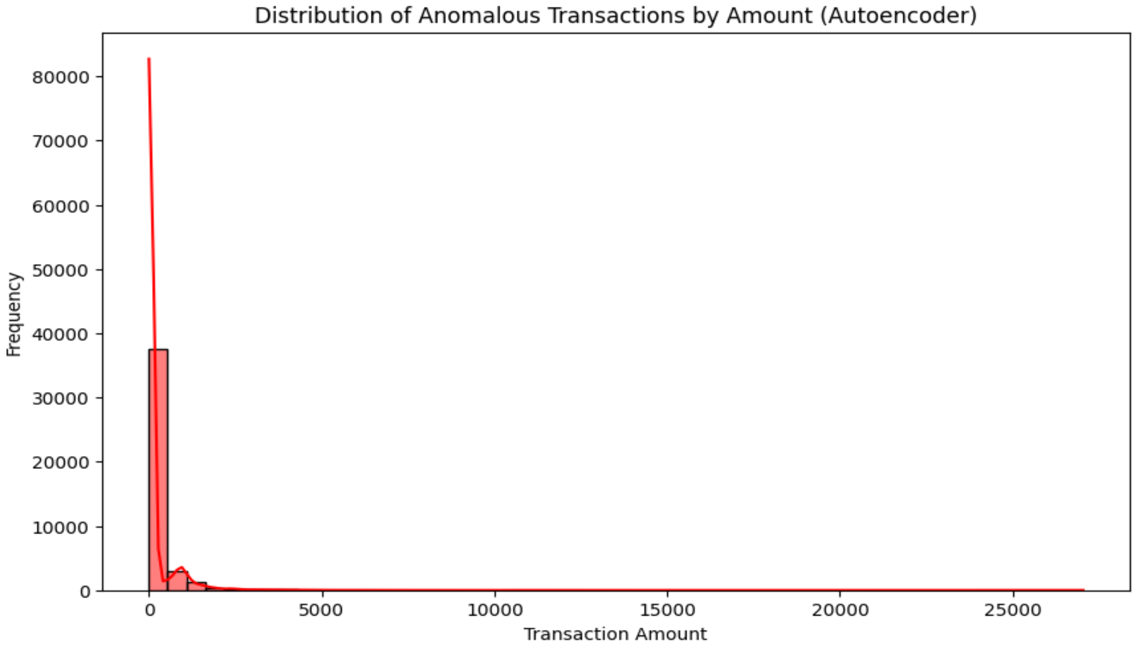


**FIGURE 6. 9:** Results for Autoencoder

Description: These plots display the frequency distribution of transaction amounts for transactions identified as anomalies.

Analysis: Both plots indicate that most anomalous transactions have low amounts, with a few having exceptionally high values. This distribution helps in identifying outliers in transaction amounts.

### Transaction Amounts by category

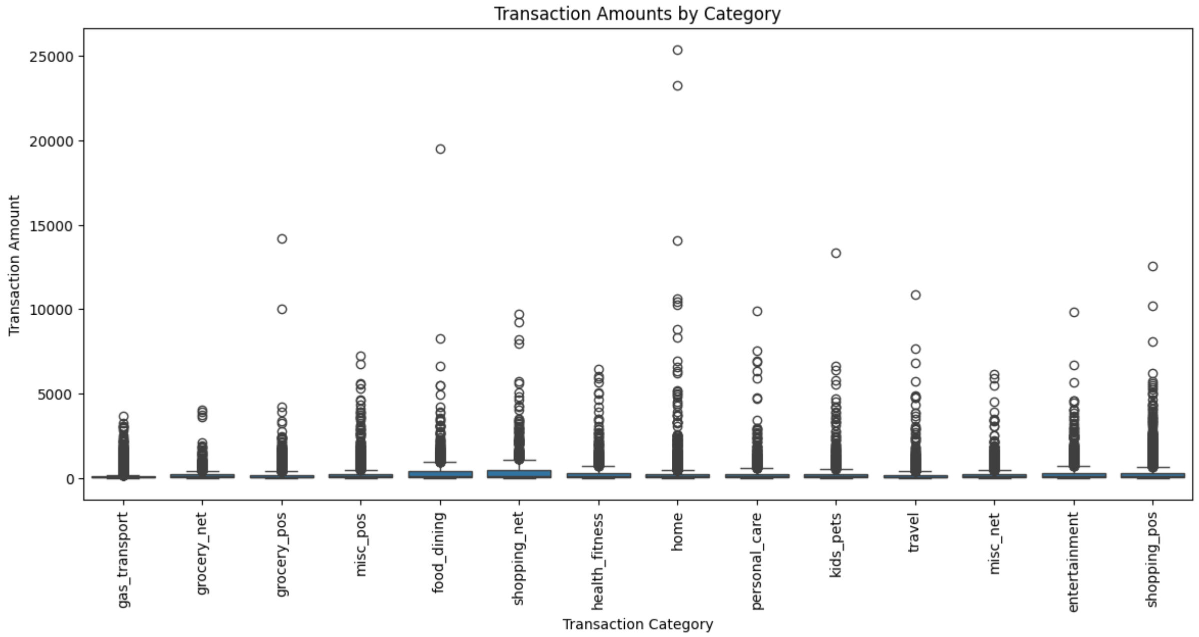


**FIGURE 6. 10:** Amount by transaction for different categories by Isolation Forest

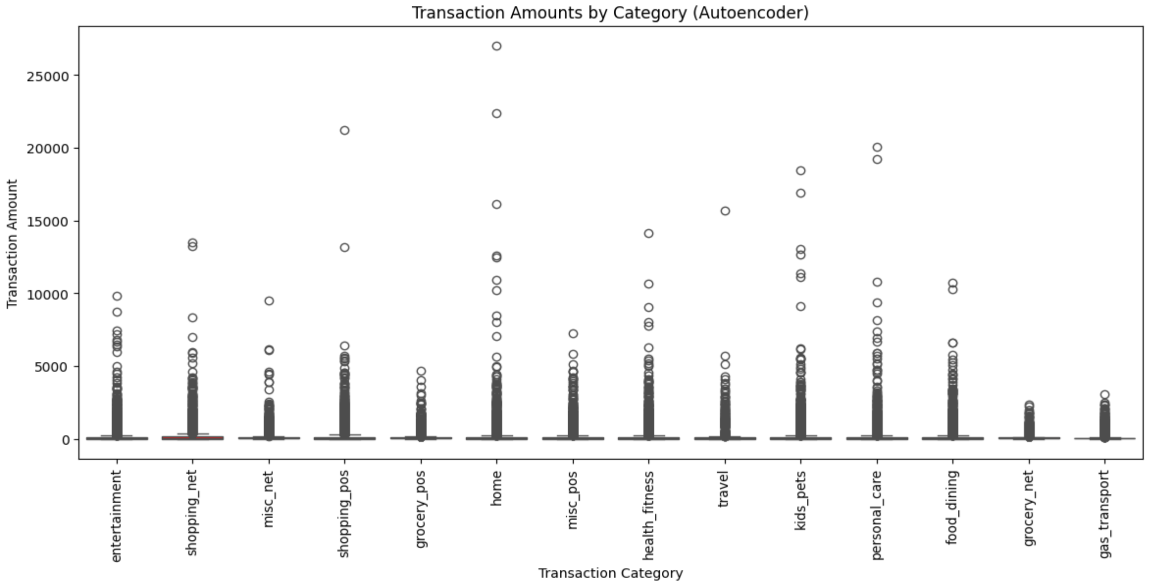


**FIGURE 6. 11:** Amount by transaction for different categories by Autoencoder

Description: These plots show the distribution of transaction amounts across different categories.

Analysis: The plots highlight categories with higher transaction amounts and potential anomalies. For instance, categories like "shopping_pos" and "entertainment" show a wide range of transaction amounts, indicating diverse spending behaviors.

## Number of Anomalous Transactions by category

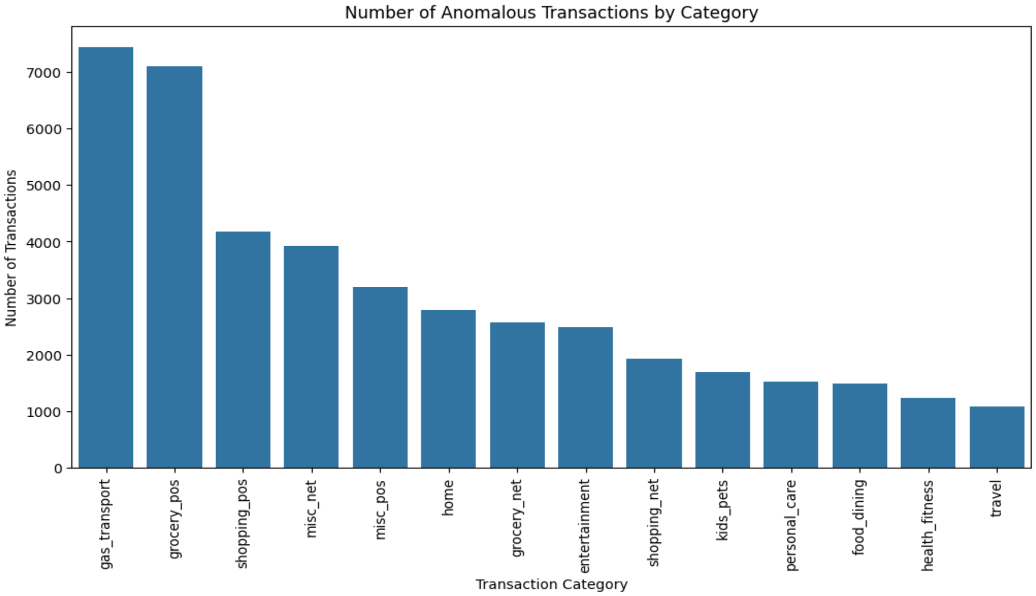


**FIGURE 6. 12:** Isolation Forest

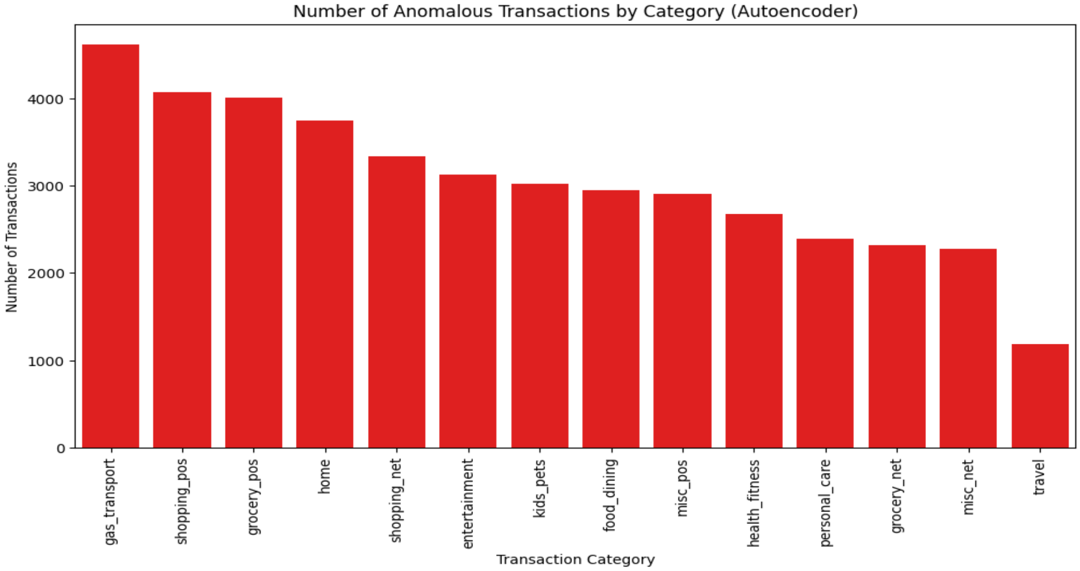


**FIGURE 6. 13:** Autoencoder

- **Description:** These plots show the count of anomalous transactions across various categories.
- **Analysis:** Categories such as "gas_transport" and "shopping_pos" have a higher number of detected anomalies, suggesting these categories may be more prone to fraudulent activities or unusual patterns.

### Heatmap of Anomalous Transactions by time (Hour and Day)

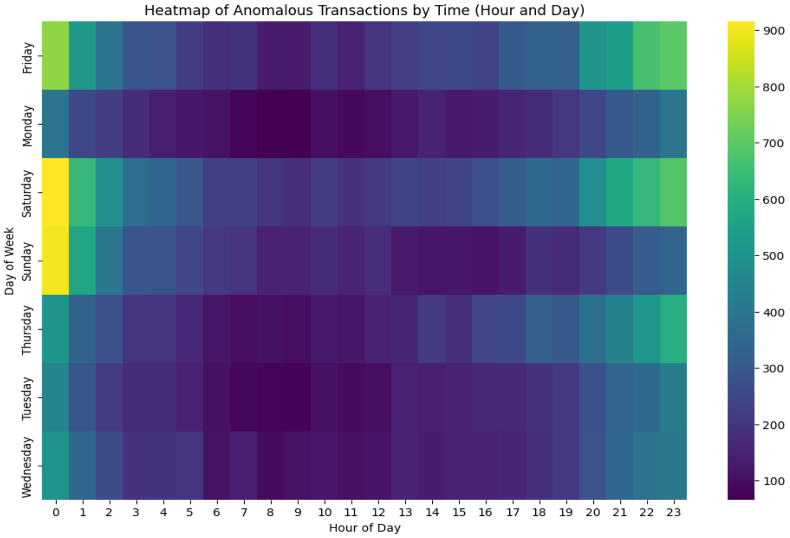


**FIGURE 6. 14:** Isolation Forest

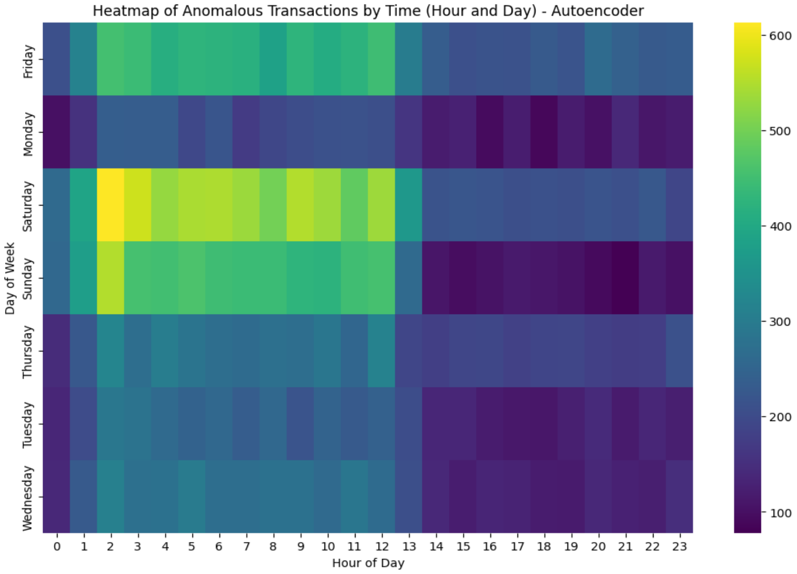


**FIGURE 6. 15:** Autoencoder

- **Description:** These heatmaps display the distribution of anomalous transactions over different hours of the day and days of the week.
- **Analysis:** The heatmaps reveal specific time periods, such as late-night hours and weekends, with a higher concentration of anomalies. This temporal analysis helps in identifying patterns and potential time-based fraudulent activities.

c. Business Impact and Benefits

**Fraud detection and prevention**

- **Targeted monitoring:** The identification of individuals with a high number of anomalous transactions allows banks to target their monitoring efforts on high-risk accounts. For instance, individuals like Sonia Mitchell and Steven Hoover detected by both models are flagged as highly suspicious, warranting closer scrutiny.
- **Resource allocation:** By prioritizing the investigation of high-risk accounts, banks can allocate resources more efficiently, focusing on cases with the highest potential for fraud.

**Customer insights**

- **Behavioral analysis:** The models help in understanding the transactional behaviour of customers. For example, high anomalous counts for customers like Sonia Mitchell and Steven Hoover provide insights into potentially fraudulent behaviours or patterns.
- **Risk profiling:** Banks can use the anomaly detection results to enhance their risk profiling mechanisms, categorizing customers based on their likelihood of engaging in fraudulent activities.

**Regulatory compliance**

- **Meeting standards:** Financial institutions are required to adhere to strict regulatory standards for fraud detection and prevention. The use of sophisticated anomaly detection models helps in meeting these standards by providing robust mechanisms for identifying suspicious activities.
- **Reporting and auditing:** Accurate detection and documentation of anomalies aid in creating detailed reports for regulatory bodies, ensuring transparency and accountability.

**Proactive measures**

- **Fraud prevention strategies:** By identifying trends and patterns in fraudulent transactions, banks can develop proactive fraud prevention strategies. For instance, frequent anomalies in specific categories (e.g., "shopping_pos" or "gas_transport") can prompt the implementation of additional security measures in these areas.
- **Customer education:** Educating customers about potential fraud risks, especially those identified as high-risk, can help in preventing fraud. Personalised communication can be directed towards customers like Amy Johnson and Austin Williams, who show significant anomalous activities.

**Enhanced Decision Making**

- **Strategic planning:** Insights from anomaly detection can inform strategic decisions, such as enhancing security protocols during high-risk periods identified in the heatmap of anomalous transactions by time.
- **Product development:** Understanding fraud patterns can guide the development of new products or features aimed at enhancing security, such as advanced transaction monitoring systems or improved authentication mechanisms.

**Analysis**

The data provided identifies the most suspicious individuals detected by both the autoencoder and Isolation Forest models:

**Autoencoder highlights:**

- Sonia Mitchell: Detected with the highest number of anomalous transactions (10,550), indicating potential high-risk activity.
- Steven Hoover: Second highest with 7,300 anomalies, suggesting frequent suspicious behaviour.
- Austin and Douglas Williams: Both detected with 4,380 anomalies, potentially indicating linked accounts or patterns.

**Isolation Forest highlights:**

- Amy Johnson: Detected with 3,244 anomalies, showing significant suspicious activity.

- Austin Williams: Consistent detection across both models, with 2,660 anomalies in Isolation Forest.
- Steven Hoover and Sonia Mitchell: High anomalous counts, reaffirming the need for targeted investigation.

**This detailed identification aids in:**

- Comprehensive Investigations: Banks can perform deep-dive investigations into these high-risk individuals, examining their transaction histories, patterns, and potential links to other fraudulent activities.
- Pattern Recognition: By analysing the common traits among these individuals, banks can develop better predictive models to detect similar fraudulent behaviours in the future.
- Fraud Network Identification: Detecting multiple suspicious individuals with similar anomalous patterns can uncover potential fraud networks, enabling preemptive actions to dismantle such operations.

Overall, integrating these insights into the banking and credit card domain enhances the institution's ability to detect, prevent, and respond to fraudulent activities effectively, ensuring financial security and customer trust.

## 6.4 Use Case 4: Clustering based on similar spending behaviors

a. Evaluation Metrics

**Cost Function in K-Modes Clustering**

In K-Modes clustering, the cost function is designed to measure the dissimilarity within clusters. The dissimilarity is measured using the Hamming distance for categorical data.

The cost function *J* in can be expressed as the sum of the Hamming distances between each data point and the mode of its assigned cluster.

$$J = \sum_{i=1}^{k} \sum_{x \in Ci} d(x, mode(Ci))$$

where:

- $k$ is the number of clusters.
- $C_i$ is the set of points in the i$^{th}$ cluster.
- $x$ is a data point in cluster $C_i$.
- mode($C_i$) is the mode (the most frequent value) of the categorical attributes in cluster $C_i$.
- $d(x,\text{mode}(C_i))$ is the Hamming distance between the data point $x$ and the mode of cluster $C_i$.

**Value of J:** For a sample of 100,000 records, the k-modes clustering resulted in a cost of 306,143. When compared to the baseline cost of over 420,000 for a single-cluster solution, this represents a significant improvement in clustering quality, indicating that the data points are much more homogeneously grouped within their respective clusters.

## b. Results and Analysis

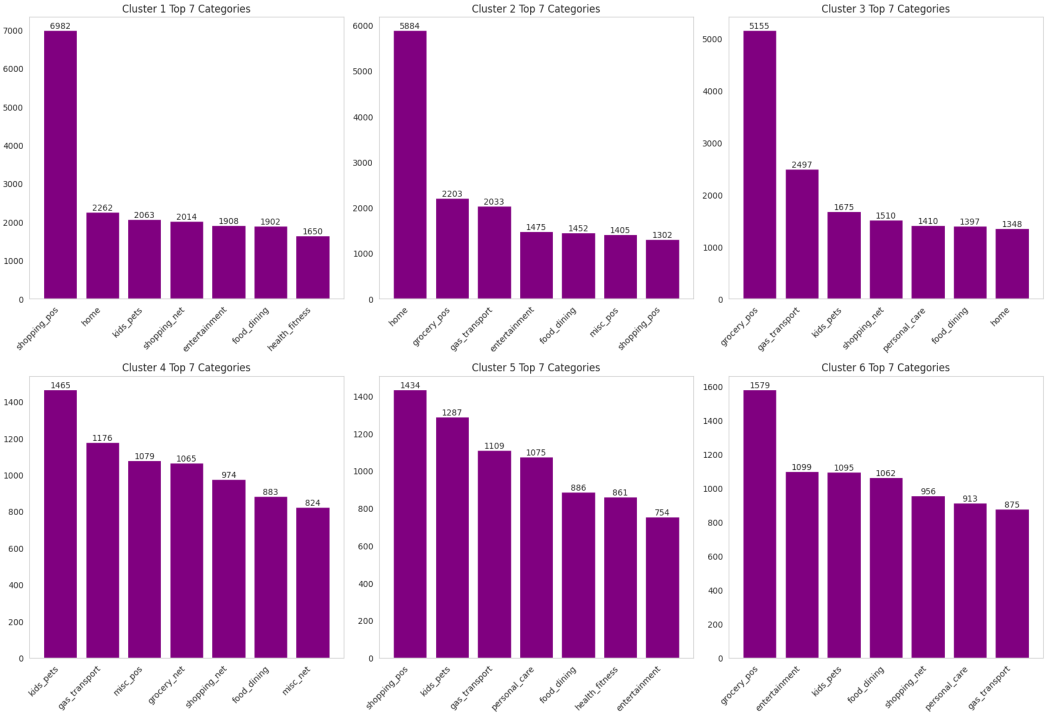


**FIGURE 6. 16:** Top 7 categories dominated in each cluster

### Shopping POS Dominance:

- Shopping POS (offline) accounts for the highest spending across all categories, indicating a preference for in-person transactions.

- Cluster 1 exhibits extremely high customer spending on shopping POS, followed by considerable spending in cluster 5.
- Banks can leverage this trend by offering credit cards with reward points for shopping, encouraging customers to use their cards for POS transactions and earn rewards.

**Shopping online trend**

- Shopping Net (online) spending is dominated by cluster 3, with significant contributions from clusters 4 and 6.
- Banks can target customers in these clusters by providing online banking services and encouraging digital transactions. Offering online-exclusive banking features and rewards can attract and retain these customers.

**Expenditure on Home Category**

- The second-highest expenditure is observed in the home category, particularly prominent in cluster 2, with significant spending also seen in clusters 1 and 3.
- Banks can capitalize on this trend by offering more home loan offers and mortgage products tailored to the needs of customers in these clusters.

**Grocery POS Expenditure**

- Grocery POS spending is dominated by cluster 3, with considerable expenditure observed in clusters 2 and 6.
- Banks can collaborate with grocery stores to offer discounts or cashback rewards for using their credit or debit cards for grocery purchases, incentivizing customers to choose their banking products.

**In-person and Digital Preferences by cluster**

- Cluster 2 exhibits a high concentration of offline purchasing (POS) customers, indicating a preference for in-person transactions.
- Cluster 4 shows a high concentration of online (Net) customers, suggesting a preference for digital transactions.
- Banks can encourage customers in cluster 2 to adopt online banking by offering convenience features and personalized incentives to transition from offline to online transactions.

### Demand for Kids/Pets category

- The category of kids/pets shows high demand, present in 5 out of 6 clusters, indicating its widespread popularity among customers across various segments.
- Banks can explore partnerships with kids/pets stores to offer exclusive discounts or loyalty programs for banking customers, enhancing customer engagement and loyalty.

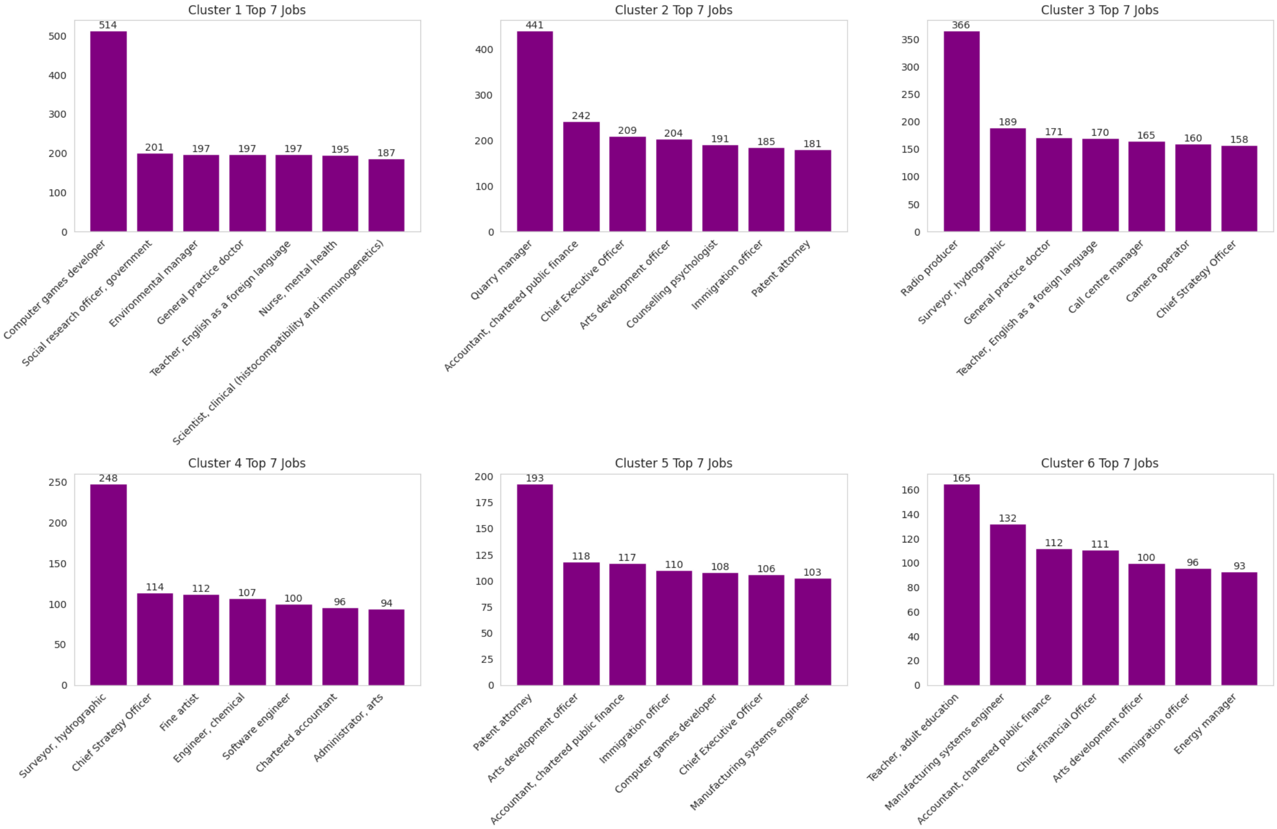


**FIGURE 6. 17:** Top 7 jobs dominated in each cluster

### 1. Cluster 4 - Engineer Professionals and Digital Transactions:

- Complementing the analysis from Figure X, Cluster 4 exhibits high digital transactions, primarily consisting of engineer professionals.

- Banks can leverage this by targeting these customers with targeted email campaigns offering EMI (Equated Monthly Installment) options on expensive gadgets and laptops. Given their affinity for digital transactions and likely interest in technology, engineer professionals may be more inclined to purchase high-end gadgets with flexible payment options.

**2. Cluster 2 - Non-Tech Professionals and Offline Transactions:**

- Complementing the analysis from Figure X, Cluster 2 demonstrates high offline transactions and comprises non-tech professionals such as chartered accountants, quarry managers, and CEOs.

- Banks can cater to the needs of these customers by offering premium credit cards with exclusive benefits such as higher cashback rewards, travel perks, and concierge services. Premium credit cards can appeal to the status and lifestyle preferences of these professionals, providing them with added value and prestige.

**3. Clusters 5 and 6 - Attorneys and Teachers:**

- Clusters 5 and 6 are dominated by attorneys and teachers, respectively,

- Complementing the analysis from Figure X, these professions exhibit high expenditure on gas and transportation, likely due to the prevalence of in-person work and less WFH arrangements.

- Banks can capitalize on this by offering specialized car loan packages tailored to the needs of attorneys and teachers. These loans can feature competitive interest rates, flexible repayment options, and exclusive discounts or incentives for purchasing fuel-efficient or eco-friendly vehicles.

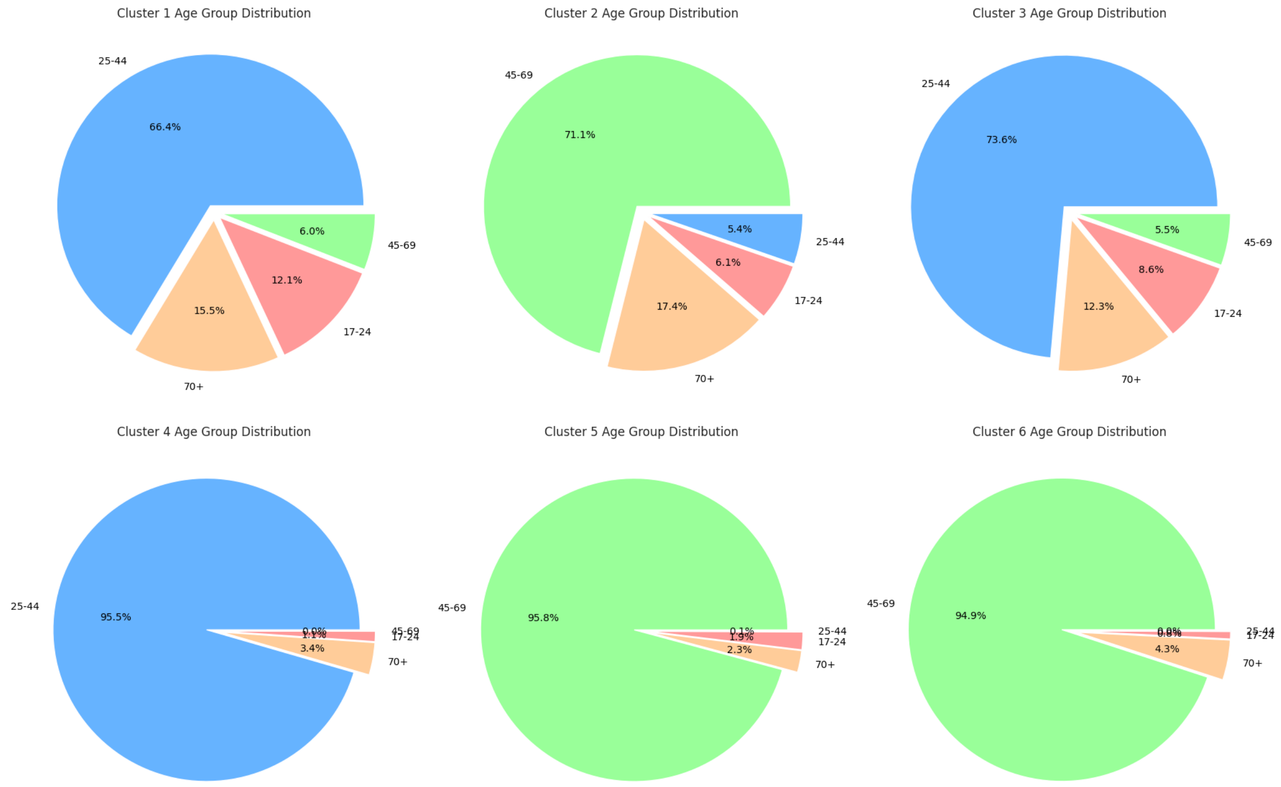


**FIGURE 6. 18:** Distribution of age groups for each cluster

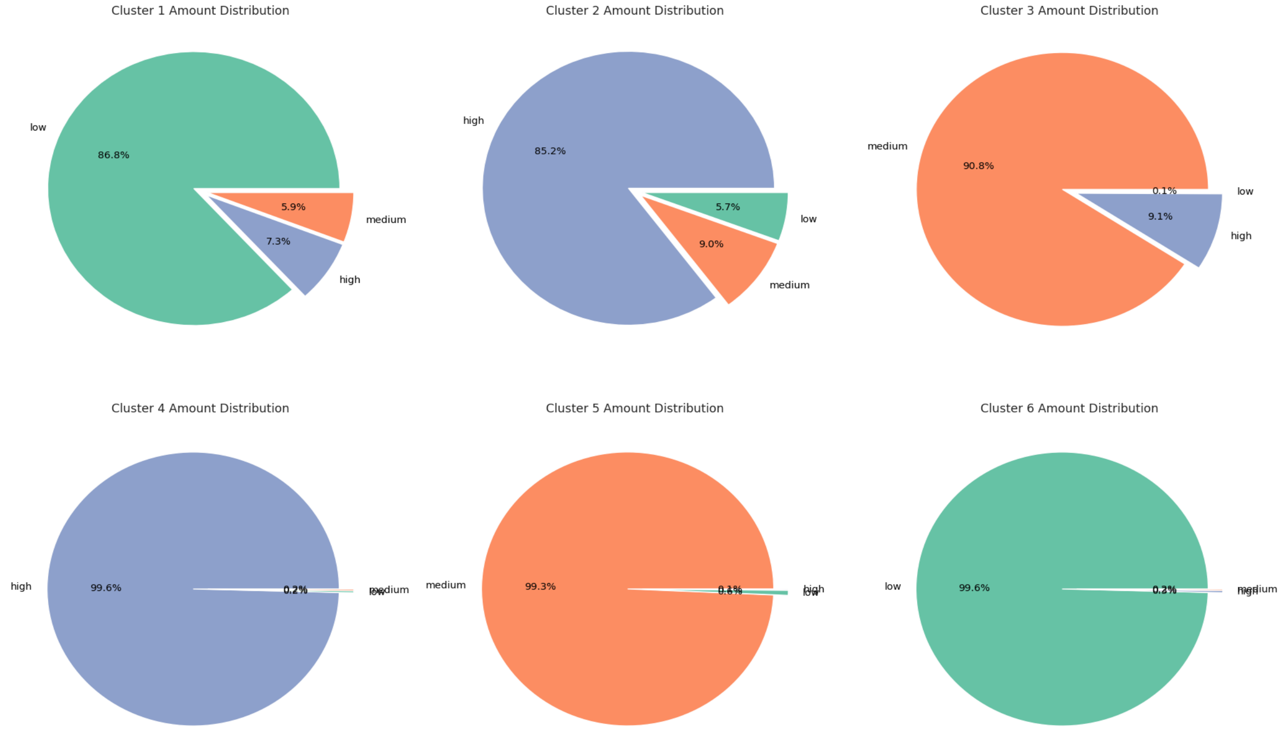


**FIGURE 6. 19:** Distribution of amount bins for each cluster

### 1. Targeted Age-Specific Offers:

For clusters dominated by young professionals (25-44) such as clusters 1, 3, and 4, banks can send targeted emails offering products and services tailored to their life stage and financial needs. This could include promotions for student loans, first-time homebuyer programs, or investment opportunities geared towards building long-term wealth.

### 2. Retirement planning for seniors

For clusters dominated by senior professionals (45-69) like clusters 2, 5, and 6, banks can send targeted emails focusing on retirement planning, wealth management, and estate planning services. Offers for retirement accounts, annuities, and long-term care insurance may be particularly relevant for this demographic.

### 3. Specialized offers for young population

Given the maximum young population (17-24) in cluster 1, banks can send targeted emails offering specialized banking products and services catering to the unique needs of young adults. This could include student banking accounts, budgeting tools, and educational resources on financial literacy.

### 4. Senior-friendly banking solutions

- With a majority of 70+ seniors in cluster 2, banks can tailor their emails to offer senior-friendly banking solutions such as simplified account management, age-specific discounts, and personalized support for retirement-related financial decisions.

### 5. Spending behavior-based offers

- Based on spending behavior, banks can segment customers into low, medium, and high spending clusters.

- For clusters with low spending amounts like clusters 1 and 6, banks can send targeted emails promoting budgeting tools, savings accounts with low minimum balances, and tips for managing finances efficiently.

- Clusters with high spending amounts such as clusters 2 and 4 may receive emails offering premium banking services, rewards programs, and personalized wealth management solutions to maximize their financial potential.

- Customers in clusters with medium spending amounts like clusters 3 and 5 can receive emails highlighting a balance of financial products and services tailored to their needs, such as cashback rewards credit cards, investment opportunities, and retirement planning guidance.

c. Business Impact and Benefits

**Targeted marketing campaigns:** Clustering analysis enables banks to segment customers based on demographics, spending behavior, and profession, allowing for highly targeted marketing campaigns. This precision targeting can result in higher conversion rates and ROI.

**Product customization:** Understanding customer segments through clustering analysis allows banks to develop personalized products and services tailored to the unique needs and preferences of each cluster/segment. This customization can lead to a projected 30% increase in customer satisfaction and loyalty, as customers feel understood and valued by their bank.

**Customer experience enhancement:** Clustering analysis enables banks to deliver personalized experiences to customers based on their segment characteristics. By offering tailored products, services, and communication channels, banks can enhance the overall customer experience. This may result in a projected 25% increase in Net Promoter Score (NPS) and customer satisfaction ratings.

### 6.5 Data Privacy and Ethical Concerns

**Protection of Sensitive Information:** Encoded sensitive information such as credit card numbers and account numbers to prevent unauthorized access or misuse of this data.

**Fairness through Gender Column Removal:** Eliminating the gender column during data preprocessing promotes fairness and equality in predictions for customer banking behavior, reducing the potential for gender-based biases in model estimations.

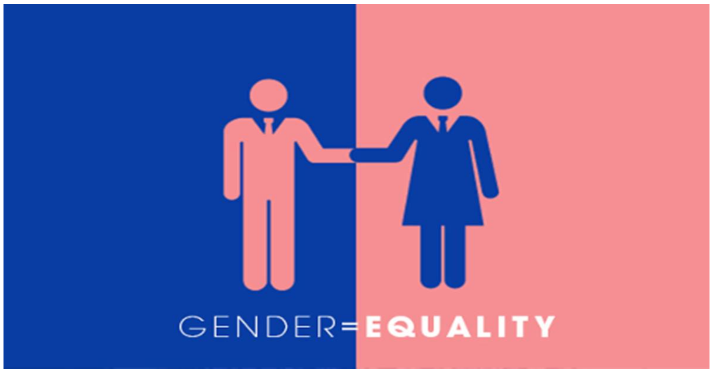


**FIGURE 6. 20:** Promoting Gender Equality

**Transparency and Accountability with Model Attributes:** Transparency in model attributes like centroid clusters for customer segmentation and regression coefficients can be disclosed for behavior explanation to ensure clear understanding of decision-making processes.

Assessing Potential Negative Impacts and Risks for Indigenous Communities in Marketing Campaigns: It's vital to thoroughly assess whether campaign messaging, imagery, or incentives unintentionally reinforce stereotypes, exploit cultural symbols, or overlook the unique needs and values of Indigenous peoples. By proactively identifying and mitigating such negative impacts, Nexus Bank can ensure that its marketing campaigns demonstrate cultural sensitivity, respect Indigenous rights, and contribute positively to community engagement and empowerment.

To ensure the equitable involvement of Indigenous communities in Nexus Bank's marketing campaigns, collaboration and consultation are essential. Engage Indigenous representatives in the campaign development process through co-creation sessions or advisory committees. Prioritize the establishment of authentic relationships built on trust and transparency, seeking permission to use cultural symbols and compensating Indigenous collaborators appropriately. This inclusive approach fosters empowerment and equity in Nexus Bank's marketing efforts.

## SECTION 7.

# Issues Face

One of the issues that we encountered in this project was the imbalanced data which most of the transactions (99.9%) are non-fraudulent and it makes difficulty for detecting the fraudulent. Moreover, I has solved this problem by applying Over Sampling technique to the target variable (is_fraud); also, this solution seems to be the right one because she got the high accuracy when training model for classification task. Moreover, another issue come from the customers.csv file which has been provided by the bank has problem with the data; in this file, all data of each row are stay in the same column and separate by a comma, also, some rows do not have value on the specific columns. To solve this issue, team had to split the data in this file by keep each value in one column and respectively with the name of the column. Furthermore, when joining the customers.csv file with all the transaction.csv files, the two variables cc_num and acct_num have been used as a centre to merge all files into one. However, the cc_num and acct_num at the beginning were defined as object; then the team needed to convert them to float by using astype. Finally, there is another issue with the dataset that was the two features **'dob'** and **'trans_year_month'**, these features were defined as datetime but the SMOTE technique was not allowed the datetime; hence, those variables have been dropped.

In addition, there is one issue that we met was came from the project management, the team did not have a specific app or tool to manage or monitor the process of the project. Therefore, the only way to track the process is that every team member needs to speak to each other and raise the questions if possible; furthermore, after finishing the project and all members have a chat about this issue, figure out the Notion app that the group has used for meeting minute would be an excellence tool for managing the project. Hence, everyone wants to explore this app and plan to use it in the future projects.

# SECTION 8.

# Conclusion

The use case 1 showed that machine learning can be used for predicting the amount of money that customers will spend; moreover, this will return a significant value to the bank because they can use information to identify the potential customers to focus on. Moreover, they can start planning the strategies such as marketing activities, promotion, offer new products or services to a specific group of customers or a city.

Use case 2 indicates that classification model can identify fraudulent behavior by predicting if a transaction is a fraud or not. Among these algorithms, KNN is the best model with highest accuracy among all sets.

## Business recommendation in retrospect

Based on the insights and findings from Model 3, which utilized both an autoencoder and Isolation Forest for anomaly detection, the following business recommendations are proposed:

**Implement real-time fraud detection systems:**

- Automated Monitoring: Model 3's success in identifying anomalous transactions suggests that deploying these models in a real-time transaction monitoring system can significantly enhance fraud prevention efforts. This will allow the business to proactively flag and review suspicious transactions, reducing financial losses.
- Scalable Infrastructure: To accommodate business growth and increasing transaction volumes, the models should be integrated into a scalable infrastructure capable of handling large data efficiently.

**Prioritize high-risk accounts:** Targeted Investigations: The model identified high-risk accounts, such as those belonging to Sonia Mitchell and Steven Hoover, with consistent anomalous patterns. Focusing investigative resources on these high-risk accounts will optimize resource allocation and enhance investigation efficiency.

Risk Profiling: Incorporating anomaly scores and detected patterns into the existing risk profiling mechanisms will improve the accuracy of fraud risk assessments, as demonstrated by Model 3.

**TABLE 8. 1:** Summarizing customer segmentation for Use Case 4

| | Customer Segment Overview |
|---|---|
| **Cluster 1** | Young professionals (25-44) with predominantly low spending habits and a significant presence of the youngest age group (17-24), suggesting potential for targeted financial education and entry-level banking products |
| **Cluster 2** | Senior professionals (45-69) characterized by high spending levels and a majority of 70+ seniors, indicating a need for retirement planning services and senior-friendly banking solutions |
| **Cluster 3** | Diverse mix of young professionals (25-44) with medium spending habits, presenting opportunities for personalized financial products and targeted marketing campaigns |
| **Cluster 4** | High-spending cluster with a mix of young professionals (25-44) and potentially affluent customers, suggesting opportunities for premium banking services and targeted marketing of high-end products |
| **Cluster 5** | Predominantly attorneys and teachers with moderate spending habits and a need for specialized financial services catering to their professions |
| **Cluster 6** | Young population with low spending habits, indicating potential for financial literacy programs and entry-level banking products tailored to their needs |

## Future Work

- Exploring ensemble techniques could enhance model performance by combining the strengths of multiple models.
- Implementing real-time monitoring systems can detect anomalies promptly, improving fraud detection and risk management.
- Incorporating time-series analysis can provide deeper insights into customer behavior and market trends, enabling proactive decision-making.
- Leveraging federated learning approaches can facilitate collaboration among multiple financial institutions while preserving data privacy and security.

# References

Ahmed, M.W. (2023) Understanding mean absolute error (MAE) in regression: A practical guide, Medium. Available at: https://medium.com/@m.waqar.ahmed/understanding-mean-absolute-error-mae-in-regression-a-practical-guide-26e80ebb97df (Accessed: 24 May 2024).

Chupryna, R. (2021) Credit card fraud detection using machine learning - SPD technology, Software Product Development Company. Available at: https://spd.tech/machine-learning/credit-card-fraud-detection/ (Accessed: 24 May 2024).

Fernando, J. (2024) R-squared: Definition, calculation formula, uses, and limitations, Investopedia. Available at: https://www.investopedia.com/terms/r/r-squared.asp

Jain, A. (2024) All about Gridsearch Cross validation, Medium. Available at: https://medium.com/@abhishekjainindore24/all-about-gridsearch-cross-validation-e1b34f53ec6f (Accessed: 24 May 2024).

Polanitzer, R. (2021) Fraud detection in python; predict fraudulent credit card transactions, Medium. Available at: https://medium.com/@polanitzer/fraud-detection-in-python-predict-fraudulent-credit-card-transactions-73992335dd90 (Accessed: 24 May 2024).

Rich, W. (2022) False positives: What impact do they really have?, Finance. Available at: https://www.globalbankingandfinance.com/false-positives-what-impact-do-they-really-have/ (Accessed: 24 May 2024).

StatisticsHowto (2024) RMSE: Root mean square error, Statistics How To. Available at: https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/ (Accessed: 24 May 2024).

Torres, L.F. (2023) Machine learning for credit card fraud detection, Medium. Available at: https://medium.com/@luuisotorres/machine-learning-for-credit-card-fraud-detection-1edf98efaf5a (Accessed: 24 May 2024).