Machine Learning Algorithms and Applications (36106)

Assessment 3 | Consolidated Report

Nathan Collins | Hemang Sharma | Rafia Tasneem

Assessment 3:	Group Project
Type:	Group & Collaborative Assessment
Deliverables:	Jupyter Notebook (x3) Final Report - Limit 5000 words
Weight:	100 pts
Due:	Friday, 26 May, by 23:59

Assessment Criteria:

Soundness of justification for selected technique.
Quality of code and visualisations.
Accuracy of results and evidence supporting claims.
Breadth of evidence of collaborative work (e.g. meeting minutes, details of contributions etc).
Criticality and specificity in evaluating assumptions and potential ethical issues.
Appropriateness of communication style to audience.

Section 1: Business Understanding

[1.1] Objective, Situation, Data Mining Goals

The aim of this project was to apply analytical and statistical methods to derive insights into the “MLAA Bank” customer and transactional data gathered over a three-year duration (2019 to 2022). The success of the models will empower MLAA Bank stakeholders and employees to implement strategies that bring value to the bank or the end customers, establishing longevity in business operations.

Three core objectives to aid in bringing value to MLAA Bank and its customers were ultimately pursued and are explained in detail below.

Core Objectives:

1. Predicting customer lifetime value (CLV) by Hemang Sharma.
2. Identifying customer segments by Rafia Tasneem.
3. Identifying high-value customers (HVC) by Nathan Collins.

Predicting CLV

By predicting CLV through transactional histories, a forecast of the customer’s lifetime value to Bank MLAA may be understood. From a business perspective, the objective is to recognise customers who frequently demonstrate high value and formulate marketing strategies to enhance customer retention to increase profitability ultimately. Predicting CLV offers Bank MLAA the capacity to allocate resources and customise personalised offers to optimise customer satisfaction and loyalty.

Identifying Customer Segments and Fraudulent Transactions.

Identifying customer segments is a standard technique applied in marketing and client monitoring. The strategy offers a business the ability to learn about the preferences and habits of its customers by analysing relevant proponents.

Applying this strategy to the bank's scenario, the objective will be to accurately comprehend and identify consumer groups through transactional behaviour and demographic information, empowering Bank MLAA to develop and tailor marketing initiatives, refining their services to respond to the requirements and demands of each segment.

The primary objective of fraud detection in banking using machine learning models is to identify and stop fraudulent activities within the financial system. Machine learning models may assist in automating the fraud detection process by examining enormous volumes of transactional data and looking for patterns and anomalies that may imply fraudulent conduct.

Identifying HVC's

An HVC is an individual who bears significant influence on the bank's bottom line and relies heavily on Bank MLAA services more than most customers. Depending on the bank's business model, an HCV may include an individual with significant capital residing in one or multiple bank accounts, as indicated by their volume of spending. The larger the sum, the greater the loaning capacity Bank MLAA may offer future customers, resulting in larger returns through interest. Likewise, frequency and quantity of transactions may also serve as valuable metrics of HVCs, as the dependence on the bank's service may indicate longevity through a dependence on the service provided or provide marketing application. Overall, an HVC establishes more overall business in the long run.

Identifying HCVs may provide further utility for targeting specific advertisement material based on spending preferences or provide utility to neighbouring merchants engaging in these transactions. These may include advertisements based on the category of interest or promotions from the bank itself with tailored account plans for specific deposit thresholds at intervals. Such outreach may increase customer loyalty and further aid in increasing Bank MLAA's profit margins.

[1.2] Project Plan

1. Data Collection: The data used in this project includes transactional data from the bank's database, which comprises customer information and purchase history.

2. Data Cleaning and Preprocessing: Missing values, outliers, and inconsistencies in the data were uncovered. This included imputing values, removing outliers, and resolving any data quality issues.

3. Data Exploration: Exploratory data analysis was performed to gain insights into the data structure, quality, and relationships between variables. This involved descriptive statistics, visualisation, and further identification of quality issues.

4. Feature Engineering and Selection: Depending on the model, specific features were dropped, converted to numeral values or isolated in distinct data frames to carry out modelling with only the necessary features. For CLV, additional features such as transaction frequency, total transaction amount, and average transaction amount to capture relevant information were created.

5. Model Selection and Instantiation: Modelling choices: **Predicting CLV** - Linear Regression Model, Random Forest and Gradient Boosting model. Each model is then tuned either by using GridSearch or polynomial feature selection or

   both. **Identifying Customer Segmentation** -. K-means Clustering, Gaussian Mixture Model (GMM) is used for clustering. Additionally, for Anomaly Detection - Isolation Forest, Local Outlier Factor and Support Vector Machine were used. Fraudulent transactions were also predicted using Logistic Regression, Random Forest classification and XGBoost Classification.

   Identifying HVC's - K-means adjusted for clustering, Hierarchical Dendogram

6. Model Evaluation: The performance of each model using appropriate evaluation metrics such as mean squared error (MSE), root mean squared error (RMSE), R-squared (R2) for regression analysis, F1 score, precision_score, accuracy_score for classification and silhouette analysis for KMeans Classification.

[1.3] Ethical Considerations & Evaluating Assumptions

Ethical and privacy implications can arise from data handling and application. The dataset provides personal details about customers, such as their **address**, **social security number** and **spending habits**. It is important to handle the final output responsibly, in addition to ensuring preventative cyber security measures protect Bank MLAA's customer's privacy in the event of security breaching.

The dataset, while large, only provides an indication of the spending habits of MLAA Bank spending habits in the USA and thus is subject to bias. Because of this, it cannot be extrapolated to represent broader populations who operate outside of MLAA Bank and internationally.

Ethicality must be exercised from the finalised extrapolated insights. This is especially important when delegating with peripheral and proximal merchant entities who may seek to value from these individuals. As such, actions of data use **must be within the bounds of consent** and **serve the customer's interests** in addition to Bank MLAA's interests.

Finally, the dataset appears to be collected throughout a historically vague time period, the **COVID-19 pandemic** - which saw deviations from existing behavioural trends and spending habits. It is likely supplementary data beyond this period will need to be analysed to more accurately gauge spending habits.

All insights derived from the modelling phase must be examined as an unbiased third party where **assumptions remain empirically-based**. This may include seeing a profession that typically pays high, a suburb that typically houses wealthy individuals or a frequent number of smaller transactions and assuming the customer offers value. Should the project reach its deployment phase, insights still remain speculative and historically determined.

Section 2: Data Understanding and Preparation

[2.1] Understanding the Data

The provided dataset includes the demographic and transactional data of the clients from MLAA bank's database. Contained were **131** .csv files detailing information of the transactions of the customers, in addition to the client's demographic information stored in a separate .csv file (see Table 1). All files were concatenated to form a single file, storing all transactional and demographic details of the clients, namely 'merged_data.csv' or 'BD_cleaned.csv'. Given the size of the final file, the dataset was decided to be stored locally.

Exploratory data analysis was performed to gain insights into the data structure, quality, and relationships between variables. This involved descriptive statistics, visualisation, and identification of any data quality issues, such as missing values or duplicates.

Table 1 | Data frames applied for analysis and modelling.

Data Frame	Feature
customers.csv	"*Customer details*" 1,000 entries, 15 features
transactions.csv	"*Concatenated transactions*" 4261035 entries, 10 features

The primary goal was to understand the variables in relation to the business objectives. By inspecting each variable's contents, key themes were identified:

Customer-centric data: such as their corresponding "**age**", "**gender**", "**street**", "**occupation**" and "**location**."

Transaction-centric data: such as the "**transaction_number**", "**account_number**", "**category**", amount "**amt**", "**is_fraud**" and "**unix time**".

Merchant-centric data: such as the number of merchant's "**name**", merchant's "**latitude**" and "**longitude**" coordinates.

For a complete list of the feature and descriptions, see the report appendix

[2.2] Data Preparation

Prior to modelling, data must be prepared by applying a range of industry-standard cleaning and manipulation techniques. These are illustrated on the following page.

Data Cleaning

To reliably ensure the accuracy and applicability of the data, cleaning and preprocessing are performed first. The first comprises the removal of duplicated entries so as not to weigh a specific outcome higher than others. Duplicates are typically hidden behind the ID column, however no duplicate entries were noticed in this dataset.

The next technique involves amending missing values. Visualising these values with the missingno library revealed 0 missing values (Figure 1).

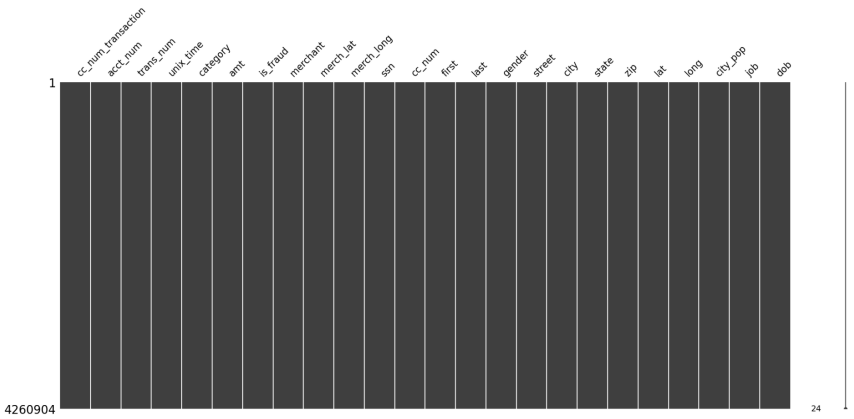


Figure 1: 'missing no.' visualisation of no missing variables present within the data frame.

Outliers removal was addressed in the preprocessing phase of each objective, as some data was more applicable to specific goals and only needed to be addressed when necessary.

Moreover, categorical values such as the customer's gender were encoded into numerical values through label encoding, where necessary, in a corresponding objective. To prevent any characteristic from predominating the clustering procedure, excess features were excluded and standardisation was applied to scale numerical features.

Next, statistical characteristics were screened by reviewing the count, mean, max, min and std. By displaying the ranges in an array, it helped identify possible outliers.

Feature Engineering

Additional features such as “**transaction_frequency**”, “**transaction_total**” and “**transaction_average**” for CLV prediction were created. In the case of applying a clustering algorithm, presenting these features as stand-alone data frames needed to be conducted. Further demographic information, such as his age, is derived from the “**dob**” feature. Converting “unix_time” into relative months also proved valuable when gauging monthly spending capacity.

Other engineering took the form of “**Recency**”, “**Frequency**”, and “**Monetary Value**”, or **RFM**”, detailing the most recent purchases, how frequently they purchase, and the total amount spent along with the time lapse between the customer’s first transaction.

[2.3] Exploratory Data Analysis

Prior to modelling, the data was explored through graphical representations. In Figure 1, a heatmap is presented, providing a visual representation of regions with the highest transaction volumes. The darker shades indicate areas where transactions are more frequent, offering insights into the geographical distribution of transaction activity.

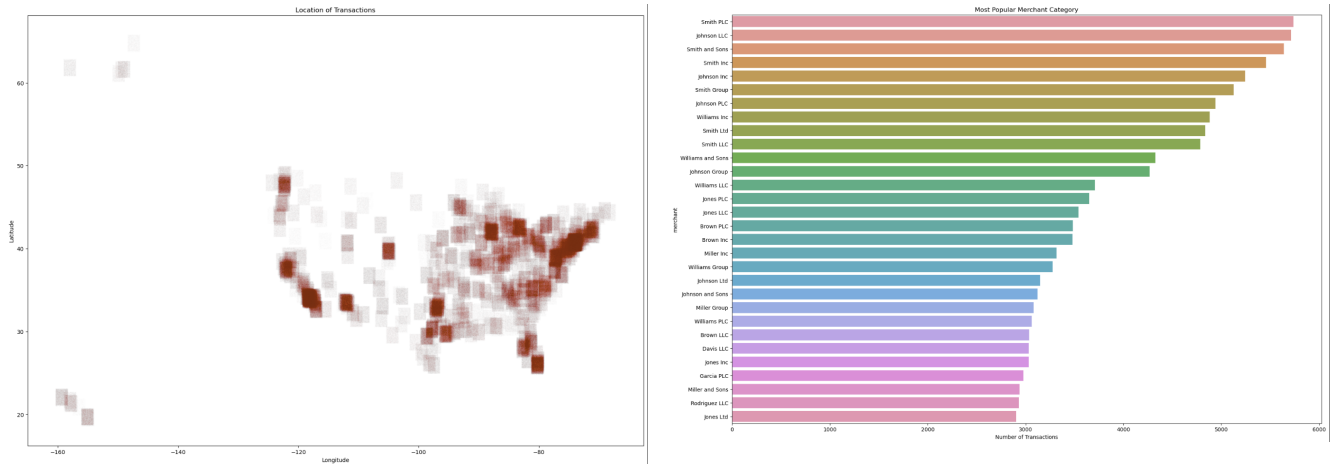
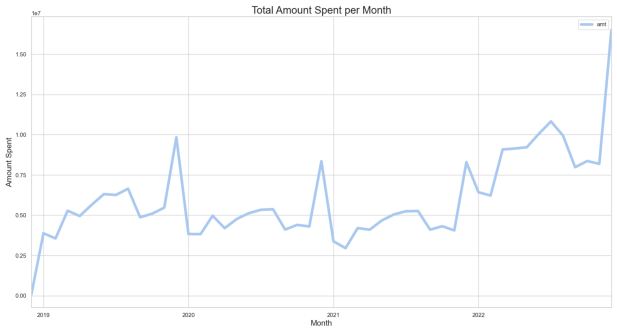
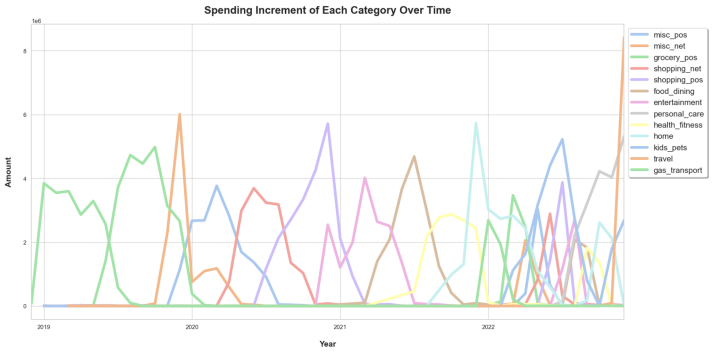


Figure 2 (left): Scatter Plot, visualising the regions with the highest transaction volumes.
Figure 3 (left): Bar Graph, visualising the merchants with the highest transaction volumes.

After discerning the most popular states where a transaction took place, the relative cities were also visualised - offering a more granular view of transactional activity at the city level. Figure 3 presents a bar chart that highlights the most frequently visited merchants within these cities and states. By analysing this chart, we can identify the merchants that attract the most transactional activity, enabling the identification of key players in the market.

By further exploring time and category variables, an understanding of the spending habits of MLAA Bank's customers could be derived. It is here the influence of the pandemic was noticed, with smaller spending capacity taking place throughout the time period (Figure 4, 5), while spikes in travel were noticed towards its conclusion (Figure 6).





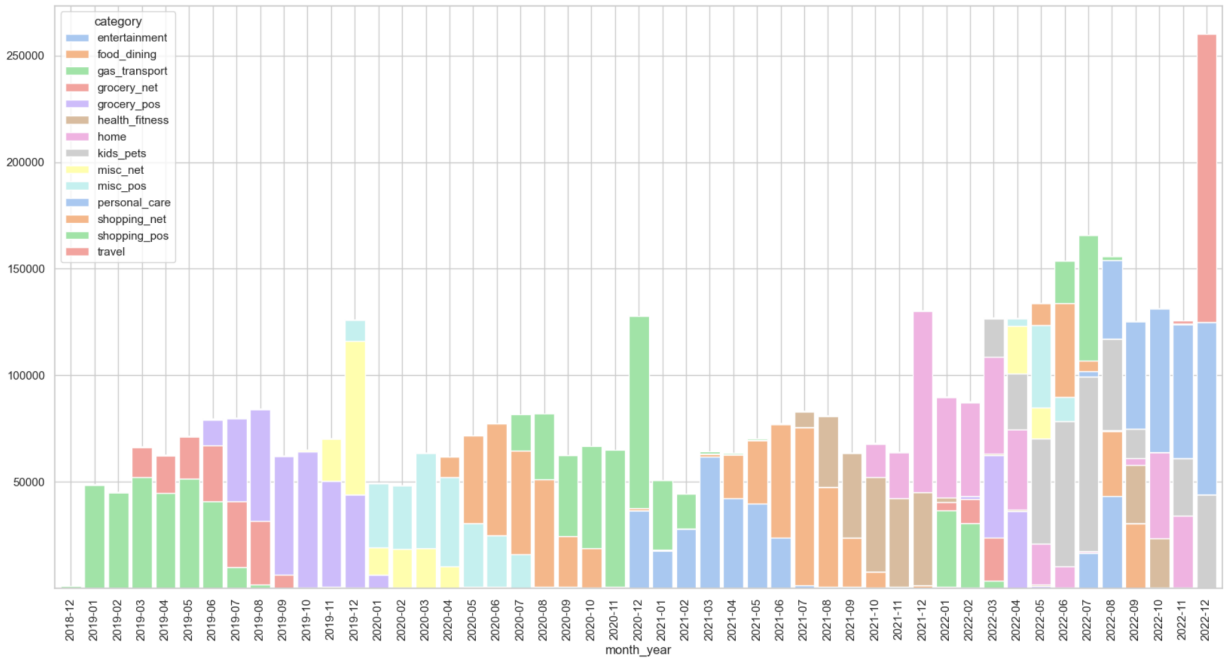


Figure 4 (left): Line Graph illustrating the total spending amount over the time frame.
Figure 5 (right): Line Graph, illustrating the spending priorities and volumes taking place during the time frame.
Figure 6 (bottom): Bar Chart, spending categories illustrated by month.

By discerning the Top 30 spending overall based on SSN (Social Security Number), insights into objective three, discerning HVCs, for example, started to develop (.

Figure 7 (left): Bar chart, visualising the highest total SSN spending.
Figure 8 (right): Bar chart, visualising the highest SSN spending per month.

For other goals, such as discerning fraudulent transactions to meet objective two, other exploratory extrapolation needed to be conducted. Figure 9 illustrates a simple pie chart distinguishing the difference between fraud and non-fraudulent transactions.

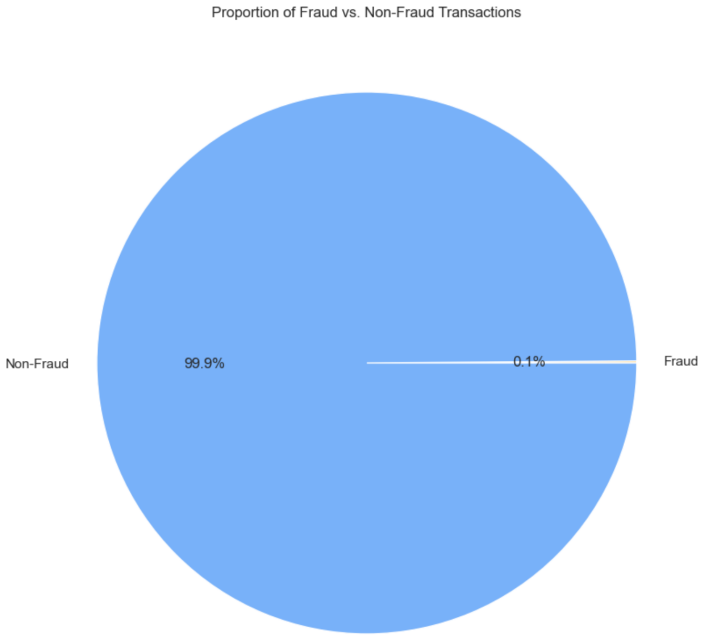


Figure 9 (right): Pie Graph, illustrating fraudulent transaction percentages.

Other goals required understanding the customer's background and discerning if this influenced spending habits. In the examples below, age and gender are explored in relation to total spending, typically dropping off as the customer grows older and typically higher overall in the female cohort.

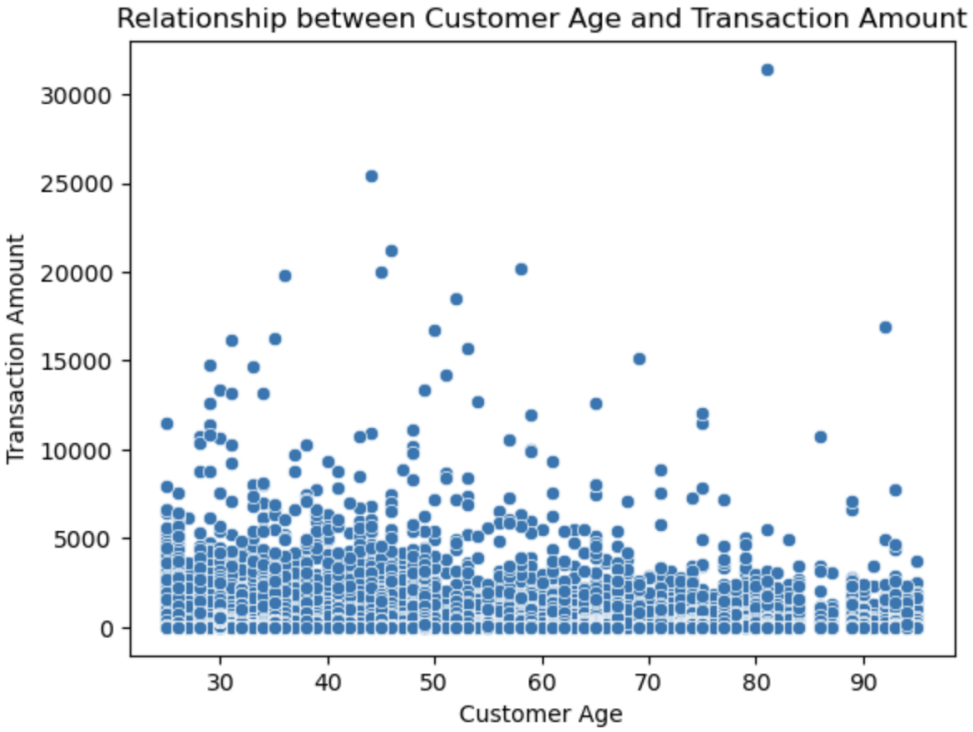


Figure 10: Scatter plot, age of customers and their relative spending.

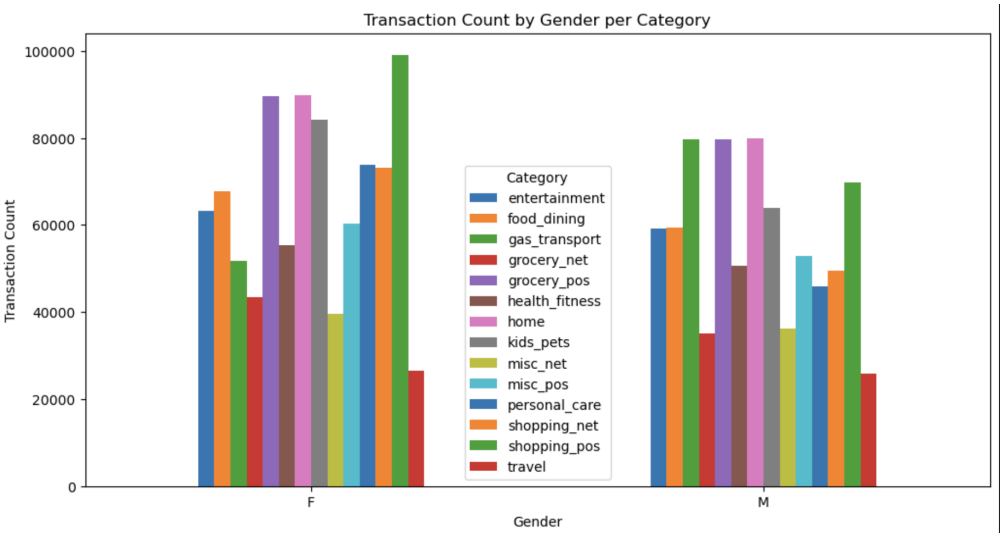


Figure 11: Scatter plot, age of customer and their relative spending.

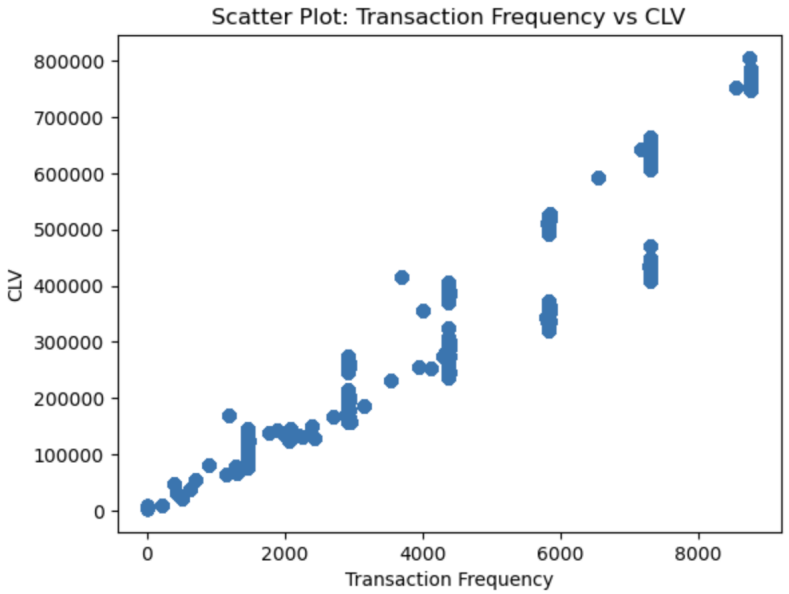


Figure 12 (left): Scatter plot, Transaction frequency against CLV.

The final charts correspond to determining CLV-associated goals. Figure 12 presents a plot showing the relationship between transaction frequency and customer lifetime value (CLV). This plot helps us understand how transaction frequency impacts CLV, providing insights into the importance of customer engagement and loyalty, while Figure 15 showcases average transaction amount versus CLV and assists in understanding insights into the value of individual transactions.

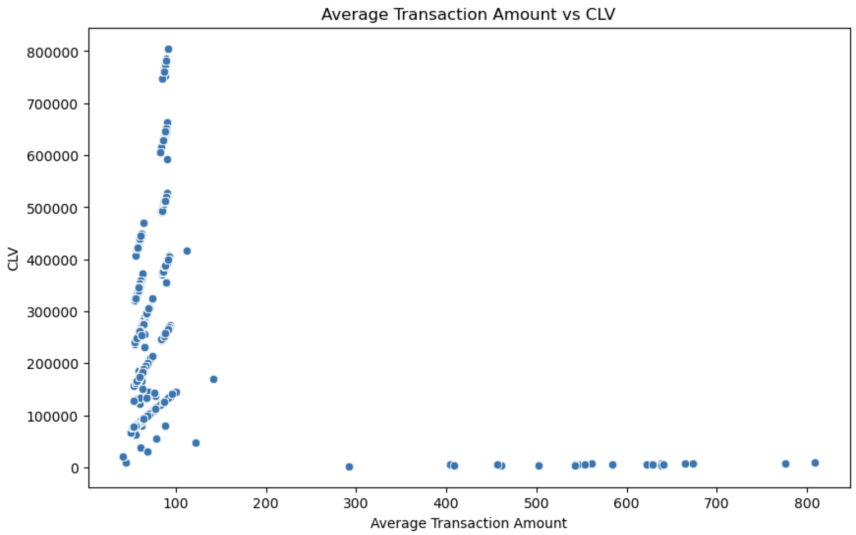


Figure 13 (right): Scatter plot illustrating average transaction amount against CLV.

Section 3: Modelling

[4.1] Applying Techniques

Three separate modelling pathways were applied to achieve three separate objectives. Each technique was carried out through an iterative process, where findings from one technique fed directly into the insights and application of the subsequent method. Following the completion of each step in the analysis phase, a performance metric was also utilised to gauge the output's prediction capacity.

These varied according to the model.

Models for Predicting "Customer Lifetime Value"

For the prediction of CLV, 6 machine-learning models were created, tested and compared.

a. **Linear Regression Model:** Utilised a linear regression model to capture linear relationships between the predictor variables and CLV.

b. **Linear Regression Model with Polynomial Features:** To account for non-linear relationships, polynomial features in conjunction with linear regression were employed.

c. **Random Forest:** Random forest model to capture complex interactions and non-linear patterns in the data was employed.

d. **Random Forest with Hyperparameter Tuning:** To optimise the performance of the random forest model, hyperparameter tuning using GridSearch to find the best combination of parameters was conducted.

e. **Gradient Boosting:** A Gradient boosting model to build an ensemble of weak learners and improve predictive accuracy was utilised.

f. **Gradient Boosting with Hyperparameter Tuning:** To further enhance the performance of the gradient boosting model, hyperparameter tuning using GridSearch to find the optimal parameter values was performed.

Model for Identifying “Customer’s Segments”

To segment customers into different groups, three clustering techniques were used, which are K-Means Clustering, BIRCH Algorithm (Balanced Iterative Reducing and Clustering using Hierarchies) and Gaussian Mixture Model.

K-Means Clustering

The dataset is divided into groups or clusters using the unsupervised machine learning technique K-means clustering. This algorithm targets to combine related data pieces depending on how similar their features are. A total of six demographic and transactional-based features, such as “frequency_of_transaction”, ”average_purchase_value”, ”total_spending”, ”recency_of_purchase”, ”customer’s age”, ”gender”, were selected for segmentation.

The algorithm starts by selecting the number of centroids, which is the centre of the clusters. For determining the ideal number of clusters, a technique named “Elbow Method” is used. The number of clusters is considered as “k”, where for each value of ‘k’, WCSS was calculated. WCSS is a measure of how far off each data point is from a cluster's centroid, squared, which helps in determining the clusters' degree of compactness. The values are plotted on a graph against the K, where the y-axis shows the appropriate WCSS values, while the x-axis shows the number of clusters as shown below:

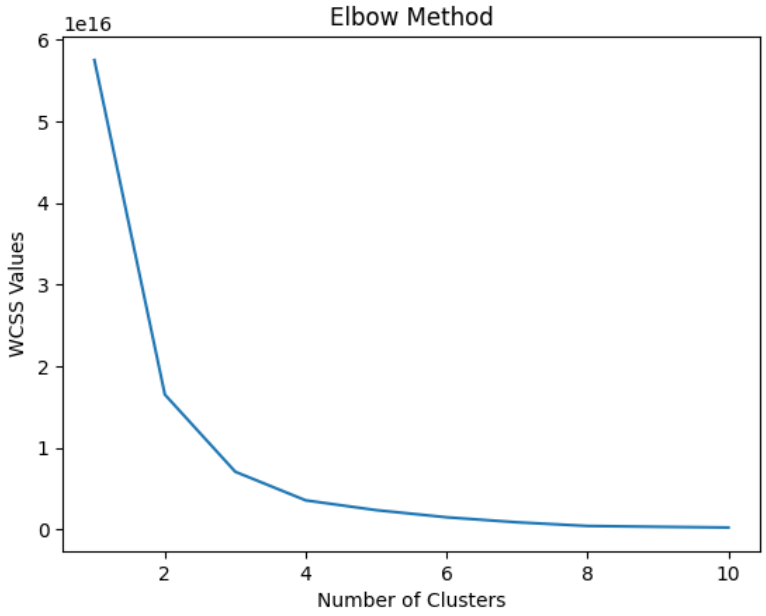


Figure 14: K-Means Elbow Method plot.

The graph presents an elbow shape which considers that the elbow point corresponds to the optimum number of clusters. Examining the graph, it is noticeable that at K=3, the WCSS starts declining towards the bottom. Hence, the optimum number of clusters for the analysis is considered to be 3. The data are then fitted into the KMeans model after the clusters have been identified. Based on a distance measure, often the Euclidean distance, each data point in the dataset is allocated to the closest centroid. The feature values are then used to measure the distance. The method updates the centroid of each cluster after allocating each data point to a cluster. By averaging the feature values of all the data points allocated to that cluster, the new centroids are calculated. When the algorithm converges, each data point belongs to a distinct cluster, and the centroids reflect the centres of those clusters.

Gaussian Mixture Model

A probabilistic model called a Gaussian Mixture Model (GMM) is used to depict complicated data distributions. It consists of a number of Gaussian (normal) distributions, each of which represents a portion of the total distribution. The underlying data distribution can be efficiently modelled and represented by GMMs by calculating the parameters of the component Gaussian distributions and their related weights. After sorting the data and calculating the quartiles, interquartiles, the box plots are generated which gives an overview of the distribution of the generated data.

Isolation Forest

An unsupervised machine learning approach called Isolation Forest is used to find anomalies. It is intended to find examples in a dataset, also known as anomalies or outliers, that dramatically differ from the rest of the data. The number of fraudulent transactions is identified in this regard. Similar to random forest classification, n_estimators of 100 are assigned to determine the number of random decision trees. The data is fitted into the model and trained to predict the outcome of fraudulent transactions. The **local outlier factor model** and **support vector machine** were also used to compare the results.

Additionally, classification techniques such as **Logistic Regression, Random Forest classification** and **XGBoost classification** techniques were also performed to predict the fraud transactions.

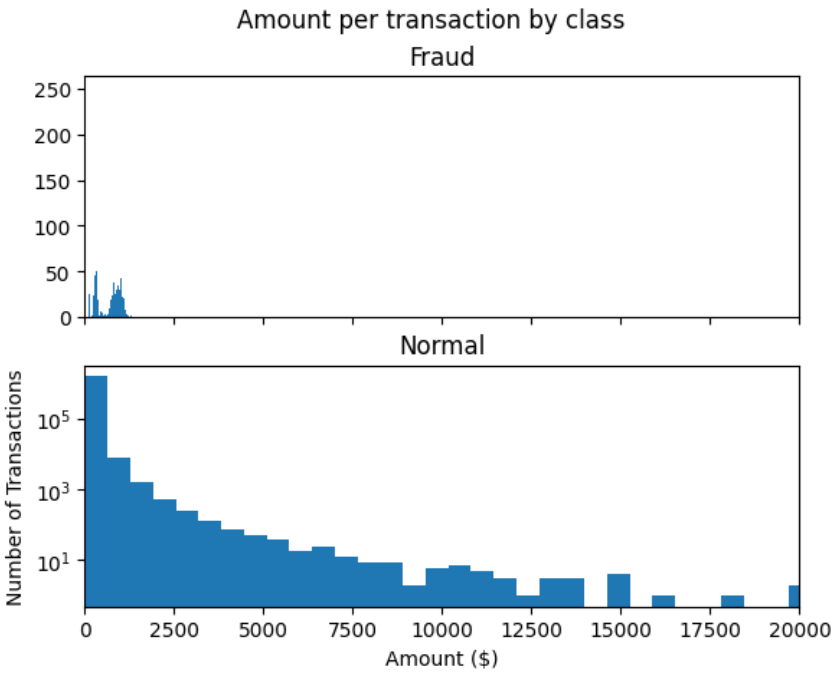


Figure: The chart demonstrates the number of normal and fraud transactions.

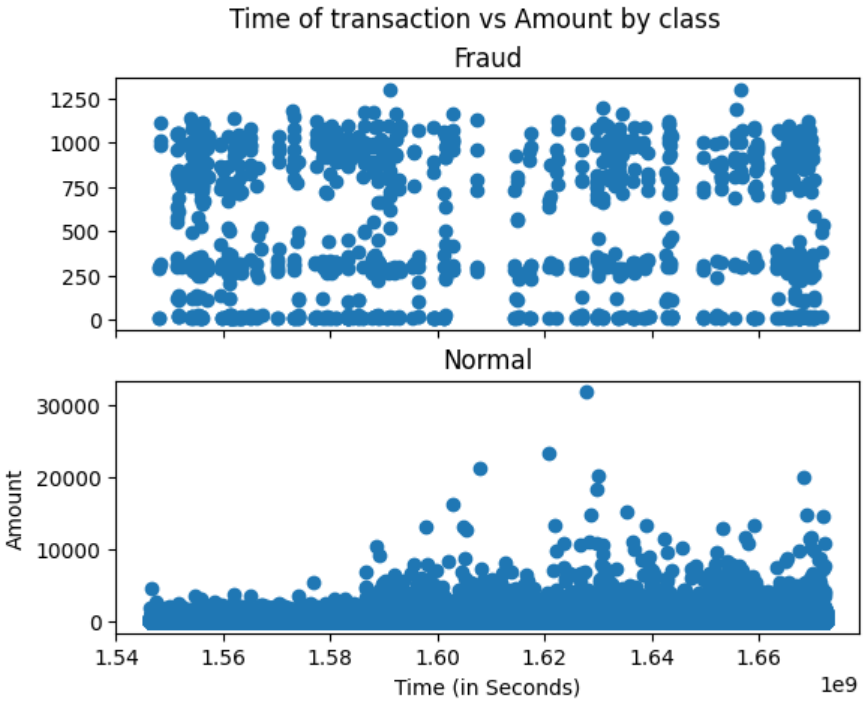


Figure: Scatter plots to identify how often fraudulent transactions occur.

Model for Identifying “High-valued Customers”

Supplementary K-Means Classification

Similar to “Customer Segments”, the approach applied an unsupervised K-Means classifier to determine Bank MLAA’s HVCs. The model was selected and pursued due to its reputation in the industry as popular with a renowned capability in delivering reliable outputs. K-Means offers a highly visual tuning functionality by adjusting the k value (number of clusters), based on assigning centroids. As the model additionally offers scalability to handle large datasets, this was also a contributing factor, given the dataset exceeds 4,000,000 entries.

Following engineering three features and mild data refinement, the average transaction cost of each customer (“transaction_average”) was determined, where results were allocated to a separate data frame with further engineered features, such as a customer’s “transaction_total” and “transaction_count”. As customer value is attributed to the amount and frequency of spending, these features were deemed a suitable metric for algorithmically discerning MLAA Bank’s HVCs.

Following standardisation, a series of scatter plots were first generated to visualise how these features interacted. In two of these plots distinguished and clear clustering was noticeable prior to running a K-Means test. By allocating a “most common category of purchase” tag to each entry, clear indications of interests began to group these clusters. A series of K-Means tests with varying K values were carried out, to determine the optimal number of clusters. Once determined, a data frame with these values was printed.

Section 4: Evaluation

[4.1] Results

Result for CLV:

The performance of each model using appropriate evaluation metrics such as mean squared error (MSE), root mean squared error (RMSE), and R-squared (R2) score was evaluated. Lower MSE and RMSE values and higher R2 scores indicate better model performance compared to the baseline performance metrics.

Baseline Performance Metrics:
Mean Squared Error (MSE): 35180725566.10877
Root Mean Squared Error (RMSE): 187565.25682041643
R-squared (R2) Score: -7.317909870963035e-06

1. Linear Regression Model
Mean Squared Error (MSE): 8.410641433390422e-16
Root Mean Squared Error (RMSE): 2.900110589855225e-08
R-squared (R2) Score: 1.0

2. Linear regression Model with polynomial features
Model Performance Metrics:
Mean Squared Error (MSE): 1.22446306052368e-13
Root Mean Squared Error (RMSE): 3.499232859533186e-07
R-squared (R2) Score: 1.0

3. Linear Regression Model, Hyper Parameters tuned with GridSearch
Best Hyperparameters: {'fit_intercept': True}
Mean Squared Error (MSE): 1.22446306052368e-13
Root Mean Squared Error (RMSE): 3.499232859533186e-07
R-squared (R2) Score: 1.0

4. Random Forest
Mean Squared Error (MSE): 0.0037291912391644744
Root Mean Squared Error (RMSE): 0.06106710439479241
R-squared (R2) Score: 0.999999999999894

5. Gradient Boosting model
Mean Squared Error (MSE): 563135.5159778388
Root Mean Squared Error (RMSE): 750.423557717799
R-squared (R2) Score: 0.9999839929498926

6. Gradient Boosting hyper parameters tuned using Grid Search
Best Hyperparameters: {'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 300}
Mean Squared Error (MSE): 564.3083168733981
Root Mean Squared Error (RMSE): 23.75517452837167
R-squared (R2) Score: 0.9999999839596132

Based on the evaluation results, a comparison was drawn related to the performance of the different models to identify the most effective model for CLV prediction. Based on these metrics, the model's rankings are:

1. Linear Regression Model with Polynomial Features
2. Linear Regression Model with Hyperparameters Tuned using GridSearch and Polynomial Feature
3. Random Forest Model
4. Gradient Boosting Model with Hyperparameters Tuned using GridSearch

The Linear Regression Model with Hyperparameters Tuned using GridSearch and Polynomial Feature appears to be the best-performing model for CLV prediction. It has the lowest MSE and RMSE values, indicating better accuracy in predicting CLV. Additionally, it achieves a high **R2 score of 0.9999999999974956**, indicating a very good fit to the data provided.

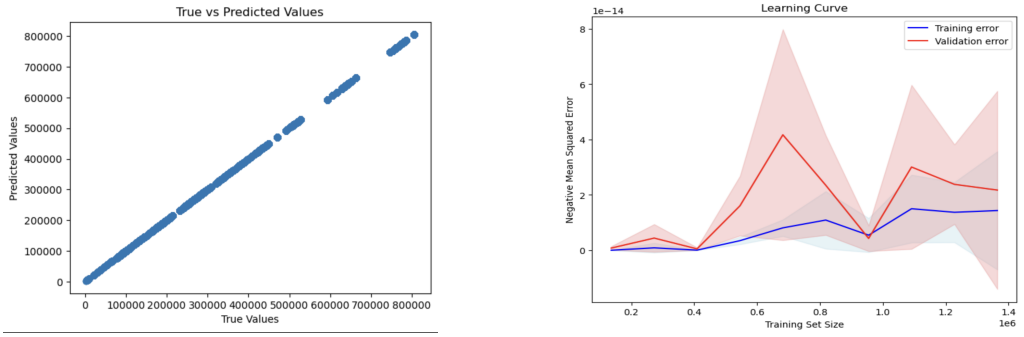


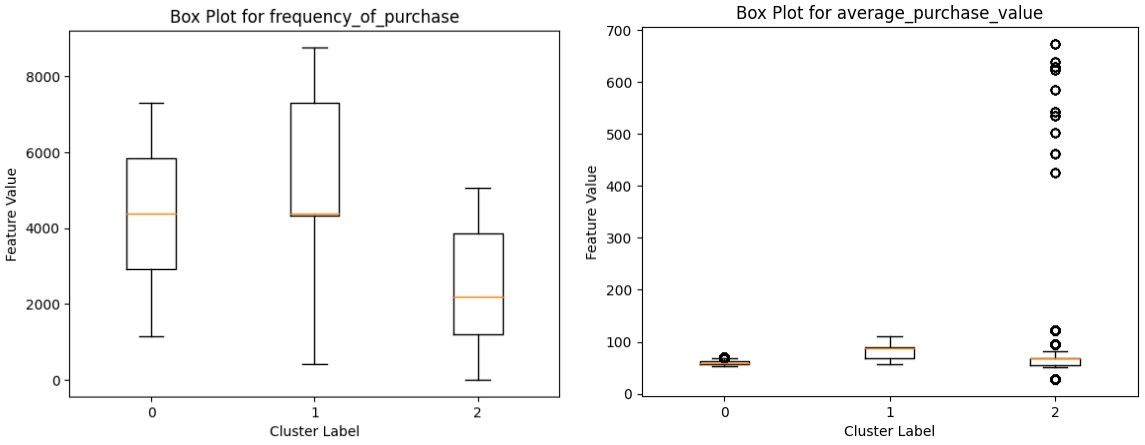
Figure (left): Plot comparing true value to predicted value.
Figure (right): Graph showing Learning curve for Linear Regression Model (Tuned using GridSearch and Polynomial Feature)

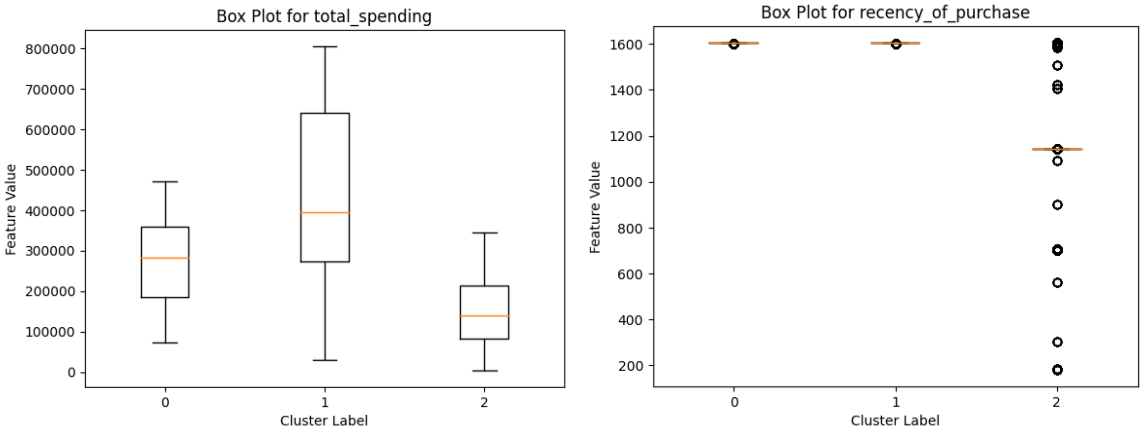
Result for k means clustering for customer segmentation:

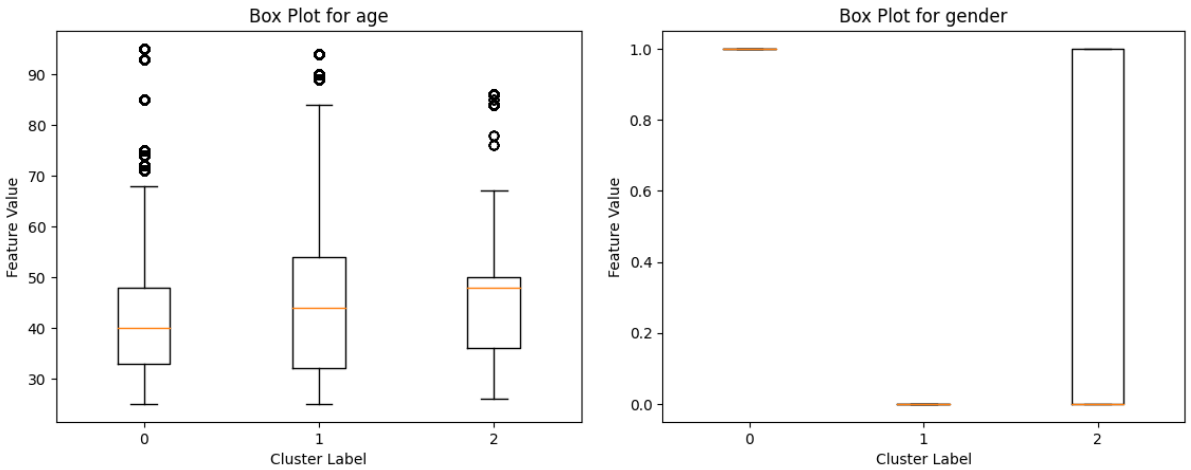
A performance metric namely Silhouette score is calculated to evaluate the performance of the clustering model. The Silhouette coefficient quantifies how well data points fit into their assigned clusters and how near they are to other clusters. It has a range of -1 to 1, with values closer to 1 indicating well-separated clusters, values close to 0 indicating overlapping clusters, and values near to -1 indicating improper cluster classification.

The analysis involved considering the top 100 points from each cluster generated, for measuring the euclidean distance between each point and calculating the silhouette score to check the performance of the model. The clustering result achieved a **Silhouette coefficient of 0.745**, indicating well-separated and cohesive clusters. This score suggests a good clustering performance with significant distances between data points within clusters and relatively closer distances to points in neighbouring clusters. This indicates that our model is a good fit for deployment.

Box-plots displaying the distribution of the segmented clusters of Gaussian Mixture Model (GMM).







Figure(s) 20: Box-plots representing the cluster labels for each feature value

The distributions are characterised by its mean values represented by a horizontal line in the box plots. The vertical line segment (whisker) that extends from the box to the lowest and highest values found in the ranges of Q1 - 1.5 * IQR and Q3 + 1.5 * IQR, are the upper and lower quartiles, respectively. Outliers are specific points that are outside of the whiskers.

Results for Anomaly detection using both supervised and unsupervised techniques.

After fitting the data in the models and training the data, classification is performed which provides an accuracy score of 0.9987, 0.9988 for isolation forest and local outlier factor models, which presents better performance and is a good fit for the data with higher accuracy in prediction.

When using classification techniques such as random forest or logistic regression, the performance provided odd results of full accuracy because of highly imbalanced data, which came a bit better with logistic regression after oversampling the imbalanced data with SMOTE technique (Synthetic Minority Over-sampling Technique) which is particularly designed to deal with the issue of unbalanced class distributions, in which one or more classes have a disproportionately small number of instances.

```
Isolation Forest: 1952
Accuracy Score :
0.9987598940065143
Classification Report :
              precision    recall  f1-score   support

           0       1.00      1.00      1.00   1573006
           1       0.02      0.02      0.02      1053

    accuracy                           1.00   1574059
   macro avg       0.51      0.51      0.51   1574059
weighted avg       1.00      1.00      1.00   1574059

Local Outlier Factor: 1772
Accuracy Score :
0.9988742480427989
Classification Report :
              precision    recall  f1-score   support

           0       1.00      1.00      1.00   1573006
           1       0.11      0.10      0.11      1053

    accuracy                           1.00   1574059
   macro avg       0.56      0.55      0.55   1574059
weighted avg       1.00      1.00      1.00   1574059
```

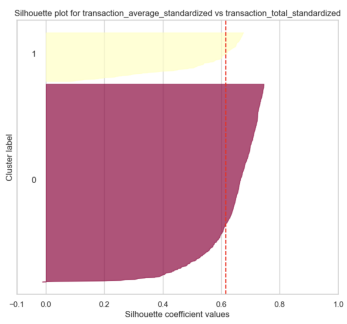
Discerning High-valued customers:

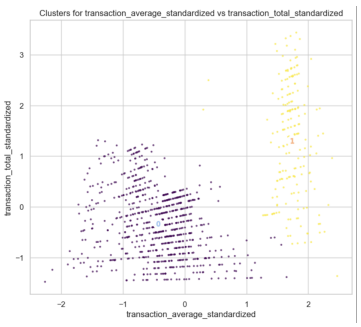
Following supplementary K-means analysis to determine HVC, multiple iterations were performed ranging between **2 to 6** centoids for clustering to take shape. Scatterplots bearing 2 centoids demonstrated the highest silhouette score, resulting where, out of three separate iterations gauging transaction counts, averages and totals, was **0.615**. While exercising use with 4 clusters provided visually cleaner distinctions, these silhouette scores yielded lower outcomes, at **0.417** (See figure below).

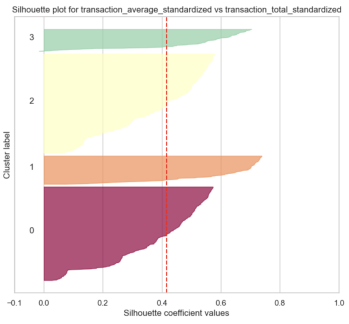
By sectioning the upper right quadrants of the 4 cluster K-Means analysis, a table detailing the customer details which offer the highest value in terms of transaction average and frequency, is determined (see right). These metrics can be adjusted and extrapolated further, depending on the business needs and boundaries that define “value”.

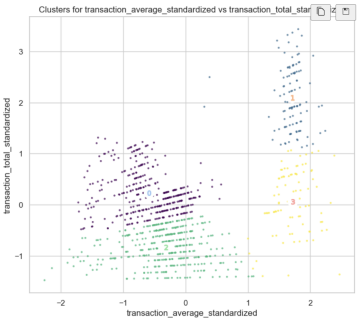
Table 2: MLAA Bank’s “HVCs.

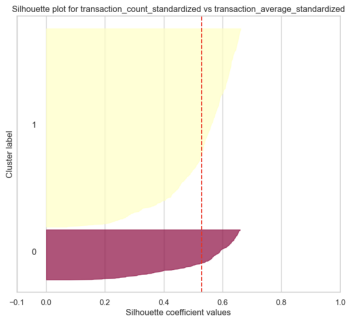
	first	last	city	job	dob
422341	Kristi	Moody	Madison	Surveyor, hydrographic	1983-06-21
614229	Lindsey	Murphy	Pittsburgh	Games developer	1988-05-26
615230	Lindsey	Murphy	Pittsburgh	Games developer	1988-05-26
410991	Amber	Butler	Ankeny	Technical sales engineer	1976-03-22
634421	Garrett	Chapman	Rosston	Fine artist	1993-04-26
358763	Michelle	Sellers	Elizabeth	Immigration officer	1978-11-29
578502	Sophia	Brown	Dunkirk	Financial manager	1995-10-08
705866	Tina	Olson	Clarks Summit	Music therapist	1992-06-17
224131	Caitlin	Becker	Rocky River	Artist	1983-03-01
435358	Kaitlin	Carter	Farmington	Nurse, mental health	1997-04-02
187487	Michelle	Williams	Sharpsville	Radio producer	1996-12-13
25056	Jordan	Martin	Rolling Meadows	Patent attorney	1978-12-26
94114	Stephanie	Kelley	Gardena	Arts development officer	1979-05-23
237232	Nicole	Beck	Beachwood	Engineer, chemical	1998-03-13
24480	Jordan	Martin	Rolling Meadows	Patent attorney	1978-12-26
173339	Kayla	Hess	Salina	Insurance claims handler	1976-04-20
499931	Teresa	Harris	Crownsville	Audiological scientist	1993-08-10
495733	Lisa	Moran	San Bernardino	Accounting technician	1974-10-02
624282	Samantha	Chavez	Crosswicks	Theatre director	1993-10-12
241659	Amy	Velasquez	Vandalia	Engineer, land	1991-07-20
639067	Paul	Thompson	San Simon	Careers adviser	1986-09-12
583700	Shelby	Brown	Girard	Education officer, museum	1985-01-10
269560	Jennifer	Murray	Chicago	Chemical engineer	1981-08-25
397481	Patricia	Roy	Cottage Grove	Health service manager	1996-11-12
136282	Melissa	Thompson	Ada	Producer, television/film/video	1990-08-21
30637	Jordan	Martin	Rolling Meadows	Patent attorney	1978-12-26
557185	Elizabeth	Miller	Westborough	Oceanographer	1990-10-03
131489	Monica	Knight	Ripley	Chief Financial Officer	1983-05-14
121722	Ashley	Porter	Eagle Rock	Teacher, adult education	1976-06-11
56275	Danielle	Mccoy	Wilson	Administrator, arts	1981-02-19
252728	Jennifer	Moore	Evanston	Surveyor, minerals	1988-04-17
713215	Eileen	Ross	Calexico	Garment/textile technologist	1989-05-17
688662	Maria	Howard	Houston	Commissioning editor	1988-05-26
195911	Lindsay	Cannon	Brooklyn	Camera operator	1983-06-01
645131	Jennifer	Murphy	Holland	Teacher, English as a foreign language	1993-04-28
33547	Lori	Perez	Turlock	Prison officer	1991-09-03
469940	Renee	Powell	Lancaster	TEFL teacher	1988-08-10
594835	Kim	Chan	Garden Grove	Jewellery designer	1976-03-31
547324	Jennifer	Hines	Bordentown	Medical technical officer	1992-07-19
54791	Danielle	Mccoy	Wilson	Administrator, arts	1981-02-19
632444	Garrett	Chapman	Rosston	Fine artist	1993-04-26
147868	Katherine	Davis	Big Rapids	Geographical information systems officer	1979-11-04
9340	Stephanie	Clark	Waterford	Data scientist	1978-08-18
382294	Julia	Friedman	Detroit	Chartered management accountant	1984-03-09
165977	April	Collier	Brooklyn	Patent examiner	1991-08-22
321382	Sandra	Stewart	Fairfield	Advertising account executive	1997-05-16
345198	Jennifer	Edwards	Oklahoma City	Chartered accountant	1994-06-22
60837	Danielle	Mccoy	Wilson	Administrator, arts	1981-02-19
594253	Kim	Chan	Garden Grove	Jewellery designer	1976-03-31
22238	Jordan	Martin	Rolling Meadows	Patent attorney	1978-12-26

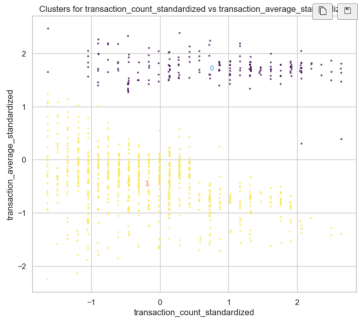


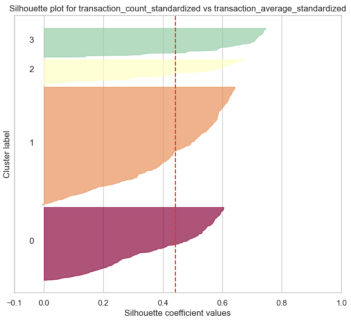


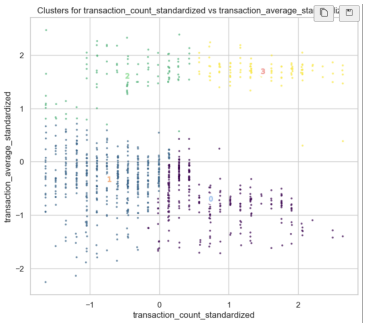


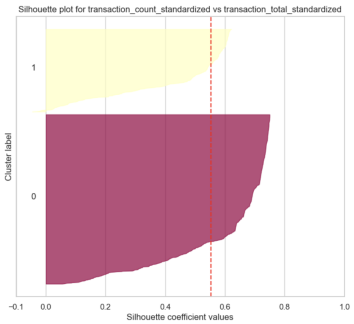


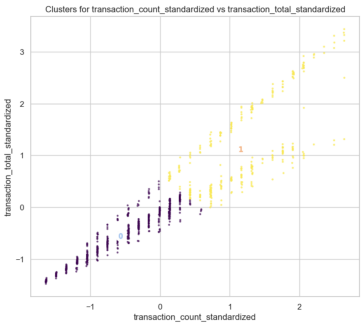


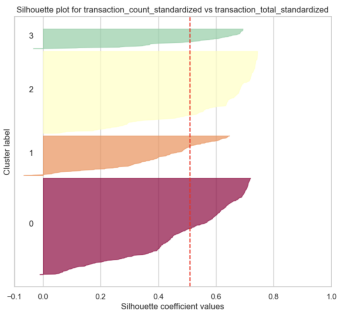


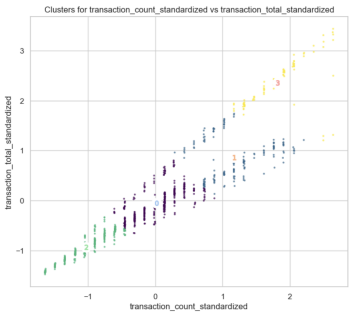












Figures 21 (left, tuplet): Silhouette scores and visualisations following K-Means analysis with 2 clusters.
Figures 22 (right, tuplet): Silhouette scores and visualisations following K-Means analysis with 4 clusters.

[4.2] Recommendation

Based on the analysis conducted on CLV predictions, the company can identify high-value customers and develop targeted marketing strategies. This may involve personalised offers, loyalty programs, and tailored communications to enhance customer retention, cross-selling, and upselling opportunities.

CLV prediction model achieved high accuracy, with low MSE, RMSE, and high R2 score, indicating its effectiveness in predicting customer lifetime value. The model can be used to estimate the future value of customers and inform marketing and retention strategies.

Customer segmentation analysis revealed distinct customer segments based on demographic and transactional characteristics. These segments can be utilised for targeted marketing campaigns and personalised customer experiences.

Identification of high-value customers using RFM analysis provided valuable insights into the customers who contribute significantly to the company's revenue. The company can focus on retaining and nurturing these high-value customers to maximise profitability.

[4.3] Deployment

The models and insights gained from this project can be deployed in the company's CRM (Customer Relationship Management) system to support decision-making, personalised marketing, and customer retention efforts. The trained models for CLV prediction, including the linear regression model, random forest model, and gradient boosting model, were saved for future use. These models can be deployed in production environments to predict CLV for new customers.

The CLV prediction model can be integrated into the company's existing systems to provide real-time estimates of customer value and guide resource allocation.

The customer segmentation analysis and identification of high-value customers can be utilised in targeted marketing campaigns and loyalty programs to enhance customer engagement and profitability.

Deployment action for targetting High-valued customers, following approval, is considered in the finalised table (table 2), detailed in this report.

Evidence of Collaborative Effort

The first week was exercised, attempting to reach everyone and consolidate a plan. Unfortunately, our fourth member, “Kalp Shah” didn’t respond. The department head data scientist “Anthony So” advised our team to proceed as a group of three.

Our Schedule / Meetings:

1st Meeting (4th May 2023):

- *Discuss and brainstorm possible use cases for the assignment.*
- *Assign tasks to explore and research potential use cases.*
- *Set a deadline for submitting the use cases.*

2nd Meeting (9th May 2023):

- *Review and discuss the submitted use cases.*
- *Select the final three business cases to focus on.*
- *Divide and assigning each team member to work on one business case.*

3rd Meeting (16th May 2023):

- *Share progress on the assigned business case.*
- *Discuss any challenges or roadblocks encountered.*

4th Meeting (23rd May 2023):

- *Continue working on the assigned business case.*
- *Share updated progress and findings.*
- *Discuss any modifications and refinements needed in the approach.*

5th Meeting (26th May 2023):

- *Finalise the contributions for the project.*
- *Consolidate the findings and prepare a report.*

Appendix

' transactions.csv ' listing information about relevant customer transactions:

category: *This feature contains the category of the payment, such as shopping_pos, grocery_pos etc*

amt: *This feature contains the transaction amount.*

is_fraud: *This feature contains a flag variable for a fraudulent transaction.*

acct_num: *This feature contains the account number of the customer.*

trans_num: *This feature contains a unique ID denoting the transaction.*

cc_num: *This feature contains the credit card number used in the transaction.*

merchant_name: *This feature contains the merchant name where the transaction took place.*

merch_lat: *This feature contains the merchant's latitude coordinate.*

merch_long: *This feature contains the merchant's longitude coordinate.*

unix_time: *This feature contains the time of the transaction.*

' customer.csv ' detailing information about each customer:

first: *This feature contains the first name of the customer.*

last: *This feature contains the last name of the customer.*

gender: *This feature contains the gender of the customer.*

ssn: *This feature contains the Social Security Number of the customer.*

street: *This feature contains the street name of the customer's address.*

city: *This feature contains the city name of the customer's address.*

state: *This feature contains the state name of the customer's address*

zip: *This feature contains the zip code of the customer's address.*

lat: *This feature contains the latitude of the customer's address.*

long: *This feature contains the longitude of the customer's address.*

city_pop: *This feature contains the city's population relative to the customer.*

job: *This feature contains the customer's job title.*

dob: *This feature contains the customer's date of birth.*

acct_num: *This feature contains the account number of the customer.*