Machine Learning Algorithms and Applications

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.

In an effort to meet the sixth proponent of the marking criteria:

“Appropriateness of communication style to audience.”

This report has been tailored to communicate specific data science terms to the recipients and/or employed teams within the bank, such as data analysts, marketing personnel and board members.

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 about 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 into the customer's lifetime value to Bank MLAA may be understood. From a business perspective, the objective is to recognise customers which frequently demonstrate high value and formulate marketing strategies to enhance customer retention to ultimately increase profitability. Predicting CLV offers Bank MLAA the capacity to allocate resources and customise personalised offers to optimise customer satisfaction and loyalty.

Identifying Customer Segments

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 their 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.

Identifying HVC’s

A HVC is an individual which 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, a 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 HVC’s, as the dependence on the bank's service may indicate longevity through a dependance on the service provided, or provide marketing application. Overall, a HVC establishes more overall business in the long run.

Identifying HCV's 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 category of interest, or promotions from the bank itself with tailored account plans for specific deposit thresholds at intervals. Such an 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 Segments - RAFIA
   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 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 into the spending habits of MLAA Bank's spending habits in the USA, and thus is subjective 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 which typically pays high, a suburb which 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 a bank's database. There were about 131 csv files holding detailed information of the transactions of the customers, as well as the client's demographic information stored in a separate csv file. All the data are then merged to form one single file which stores all the transactional and demographic details of the clients, namely 'merged_data.csv'.

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

See the appendix for a complete list of the features.

The first goal was to understand the variables in relation to the business objective.

By inspecting each variable's contents, key themes were identified:

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

Transaction-centric data:
such as the transaction number, account number, product categories, amount and unix time..

Dealership-centric data:
such as the number of merchant's name, merchant's latitude coordinator, longitude coordinator,

[2.2] Data Preparation

Prior to modelling, data must be organised by applying a range of standard cleaning and manipulation techniques.

Data Cleaning: To assure the quality, accuracy, and applicability of the data, data cleaning and preprocessing are done in the first stage. The information and description of data, as well as the occurrence of null values, outliers were identified where missing values were imputed and outliers which may skew the clustering

results were eliminated. Moreover, categorical values such as customer's gender were encoded into numerical values through label encoding. In order to prevent any characteristic from predominating the clustering procedure, standardisation is also performed to scale numerical features to a standard limit.

Feature Engineering: Additional features such as transaction frequency, total transaction amount, and average transaction amount to capture relevant information for CLV prediction were created. To perform the analysis of clustering, some necessary features have been formed based on the relevant features provided in the dataset. The demographic information of the customer such as his age is derived from the customer's date of birth. Besides, the average purchase value, frequency of purchase, total spending and recency of purchase of each customer are calculated based on their account number, amount, transaction number and unix time. Therefore, Recency, Frequency, and Monetary Value, or RFM, analysis has been performed where recency is the most recent purchase a consumer made, frequency is how frequently they make purchases, and monetary value is the total amount they spent along with the time till customer's first transaction. These necessary features are required as the transactional information for clustering.

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

Table 1 Data frames applied for analysis and modelling.

[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 showcases a bar chart illustrating the most popular states where transactions took place. This visualisation allows us to identify the states with the highest transaction counts, providing valuable information about the geographic concentration of transactional activity.

Moving to Figure 3, we have another bar chart that focuses on cities. It displays the cities where the highest transaction counts occurred, offering a more granular view of transactional activity at the city level.

Figure 4 presents a bar chart that highlights the most frequently visited merchants. By analysing this chart, we can identify the merchants that attract the most transactional activity, enabling the identification of key players in the market.

In Figure 5, we revisit the heatmap to observe the regions with the highest transaction volumes, similar to Figure 1. This repetition allows us to compare and validate the findings from both visualisations.

Figure 6, like Figure 2, provides a barchart that represents the most popular states where transactions occurred. This serves as another perspective to reinforce the understanding of transaction distribution across different states.

Figure 7 offers a bar chart focused on cities, reiterating the information presented in Figure 3. By presenting the data in multiple visual formats, we gain a comprehensive understanding of transaction activity at the city level.

Figure 8 complements Figure 4 by showcasing a bar chart that highlights the most frequently visited merchants. This allows us to validate and further analyse the significance of these merchants in driving transaction volumes.

Moving to Figure 9, we have a graph that displays the top 30 spending overall based on SSN (Social Security Number). This graph provides insights into the customers who contribute the most to overall spending based on their unique SSN.

Figure 10 presents a graph depicting the top 30 spending monthly based on SSN. This graph allows us to identify the customers who have the highest spending patterns on a monthly basis, providing insights into their purchasing behaviour over time.

Figure 11 showcases a graph illustrating spending based on categories. This graph allows us to identify the categories that drive the highest spending, enabling the company to focus their efforts and resources accordingly.

Moving to Figure 12, we have a graph that compares fraudulent transactions. By analysing this graph, we can identify patterns and trends related to fraudulent activity, aiding in the development of fraud prevention strategies.

Figure 13 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.

Figure 14 offers a plot that shows transaction counts based on gender per category. By visualising this data, we can identify any gender-specific preferences or trends in transaction behaviour across different categories.

Figure 15 showcases a plot illustrating the average transaction amount versus CLV. This plot helps us understand the relationship between transaction amount and customer lifetime value, providing insights into the value of individual transactions.

Figure 16 presents a plot showing the relationship between customer age and transaction amount. By analysing this plot, we can identify any correlations or patterns between age and spending behaviour, aiding in targeted marketing strategies.

Finally, in Figure 17, we have a plot displaying the transaction frequency graph. This plot helps us understand the distribution of transaction frequency, highlighting any spikes or variations in transaction activity over time.

Overall, these visualisations provide valuable insights into transactional patterns, customer behaviour, and fraud detection, enabling data-driven decision-making and strategic planning for the company.

First - spending volume was determined by Location and Popular Merchants

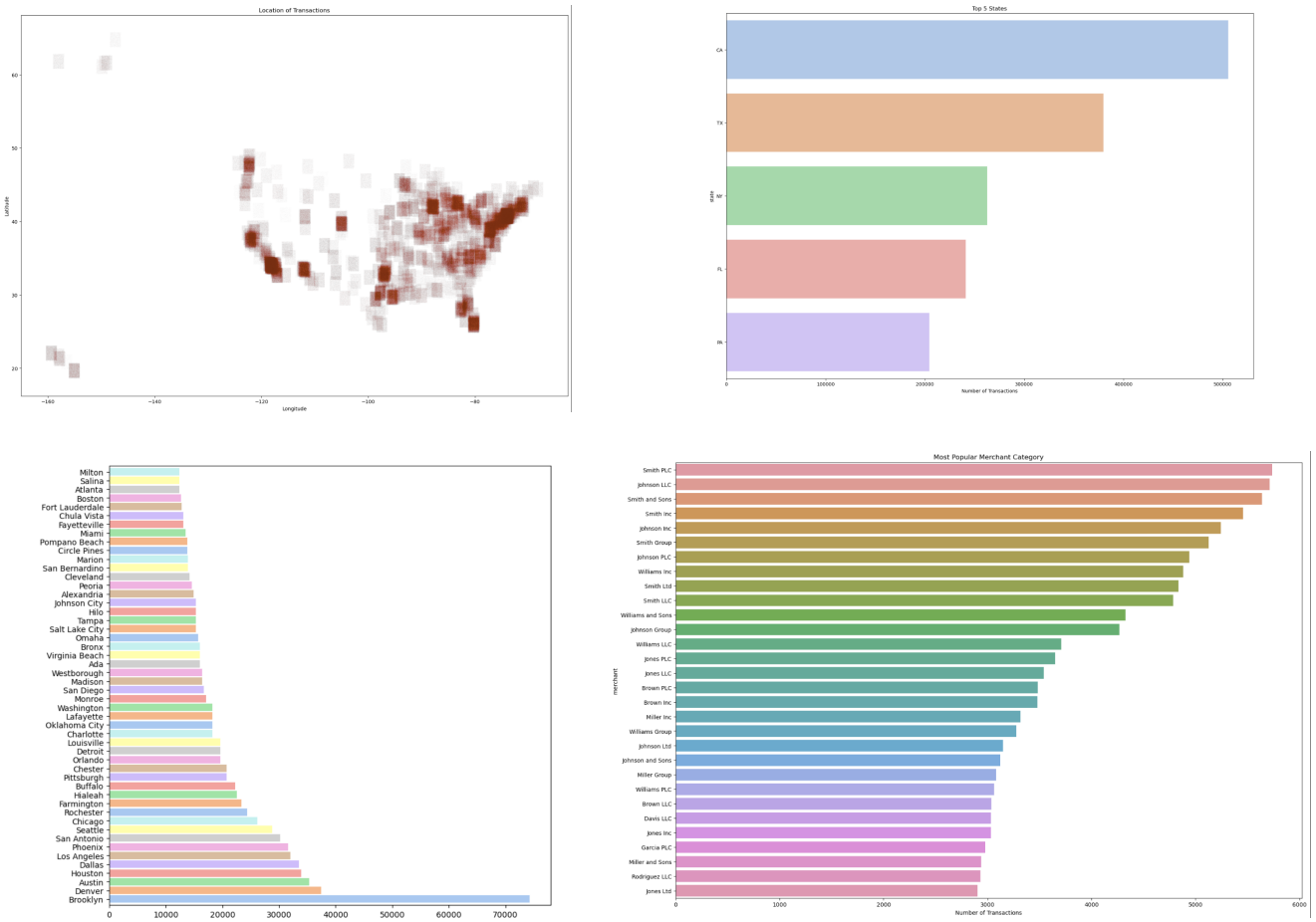
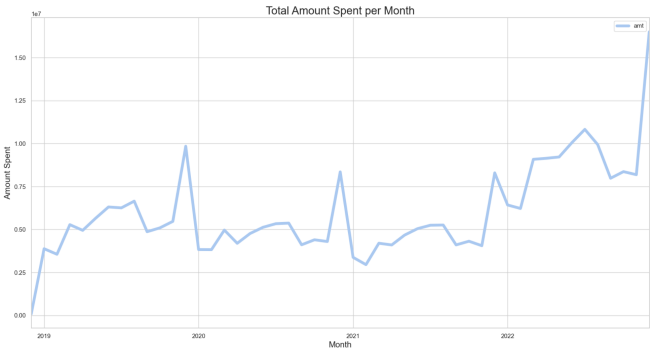
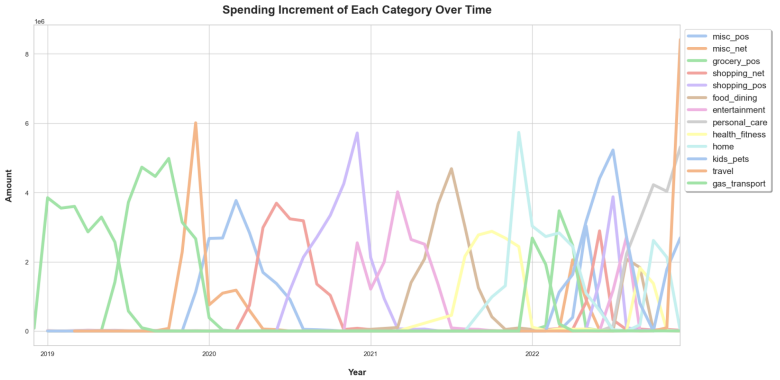
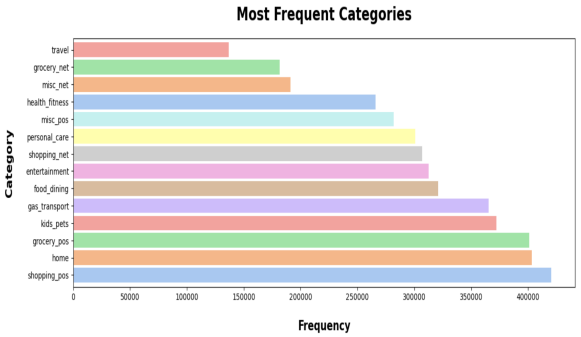


Figure 1 (top left): Heatmap, visualising the regions with the highest transaction volumes.
Figure 2 (top right): Barchart, visualising the most popular states where transactions took place.
Figure 3 (bottom left): Bar chart, visualising the cities where the highest transaction counts took place.
Figure 4 (bottom right): Bar chart, visualising the most frequently visited merchants.







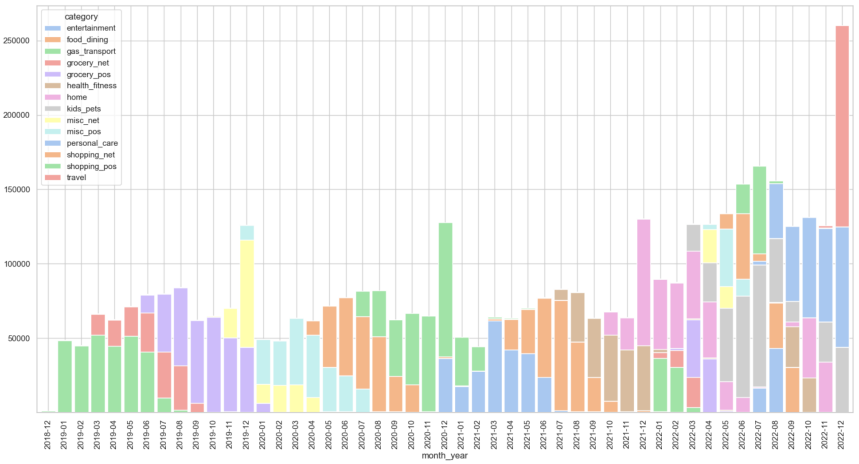
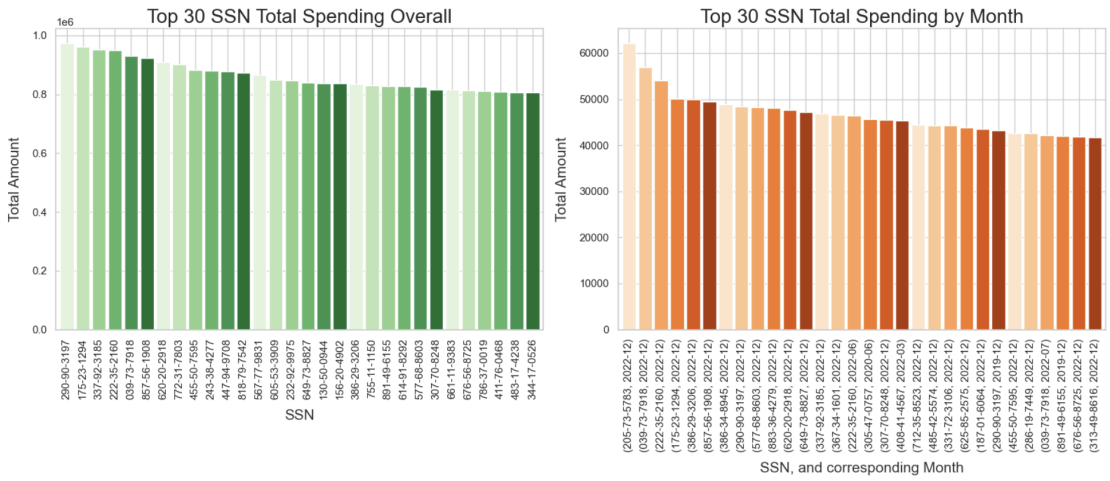
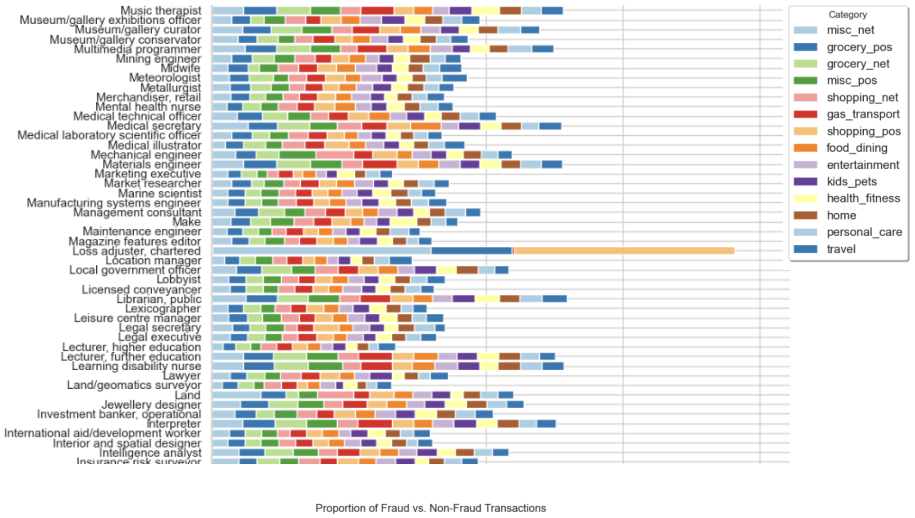


Figure 5 (top left): Heatmap, visualising the regions with the highest transaction volumes.
Figure 6 (top right): Barchart, visualising the most popular states where transactions took place.
Figure 7 (bottom left): Bar chart, visualising the cities where the highest transaction counts took place.
Figure 8 (bottom right): Bar chart, visualising the most frequently visited merchants.





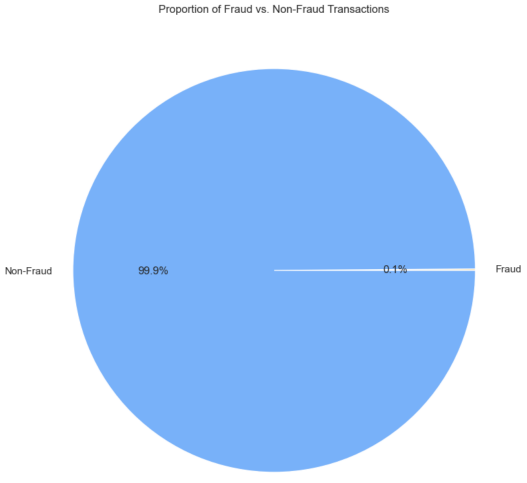
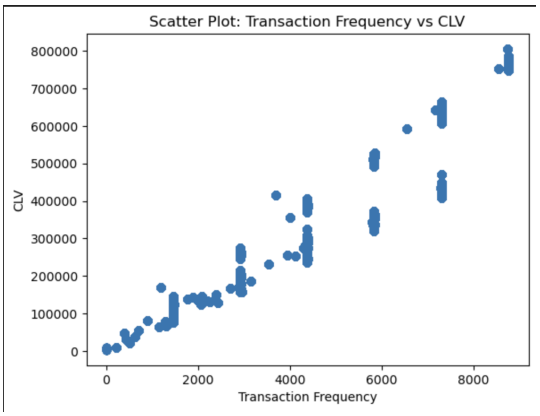
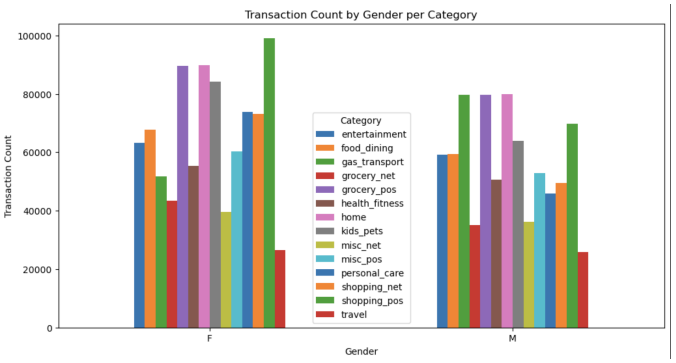
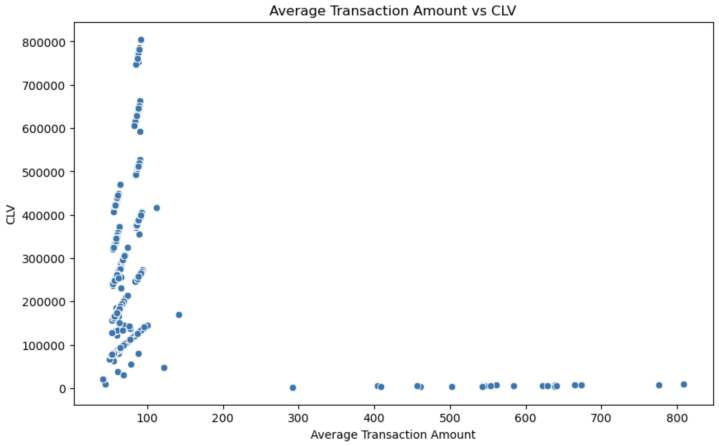
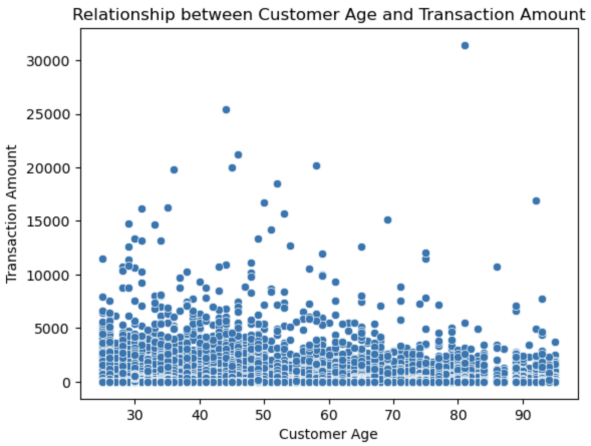


Figure 9 (top left): Graph showing top 30 spending overall based on SSN.
Figure 10 (top right): Graph showing top 30 spending monthly based on SSN.
Figure 11 (middle): Graph showing spending based on category.
Figure 12 (bottom): Graph showing comparison of fraudulent transactions.









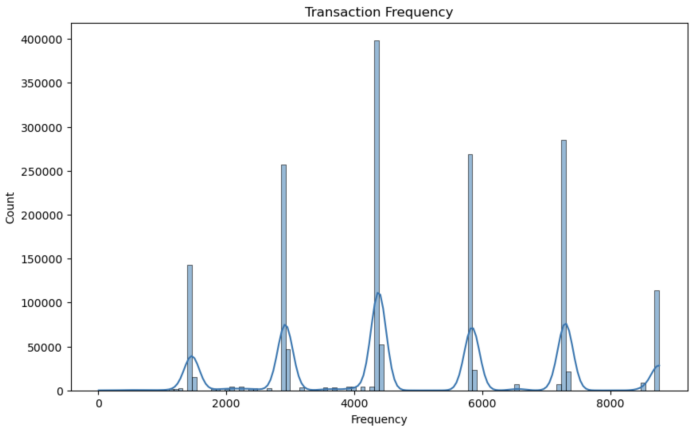


Figure 13 (top left): Plot showing Transaction Frequency vs CLV.
Figure 14 (top right): Plot showing Transaction count based on gender per category.
Figure 15 (middle): Plot showing Average Transaction Amount vs CLV
Figure 16 (bottom left): Plot showing Relation between customer age and transaction amount.
Figure 17 (bottom right): Plot showing Transaction frequency graph.

Section 3: Modelling

[4.1] Applying Techniques

Models for Predicting "Customer Lifetime Value"

For 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 2
"Identifying Customer's Segments based on transactional and demographic behaviour."

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.

KMeans 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:

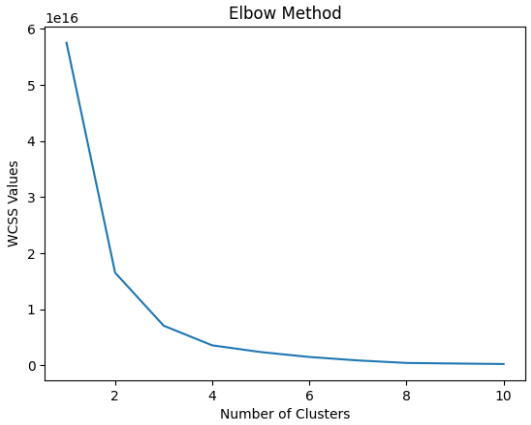


Figure: Elbow Method

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.

Model 3

K-Means | Identifying "High-value Customers”

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 models 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.

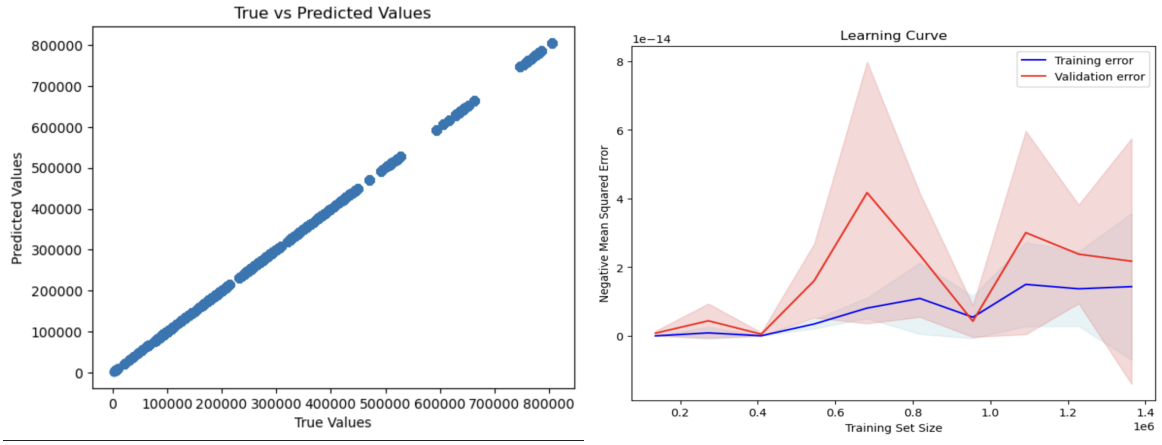


Figure (left): Plot comparing true value to predicted value.
Figure (right): Graph showing Learning curve for Linear Regression Model (Tuned using GridSearch and Polynomial Feature)

[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.

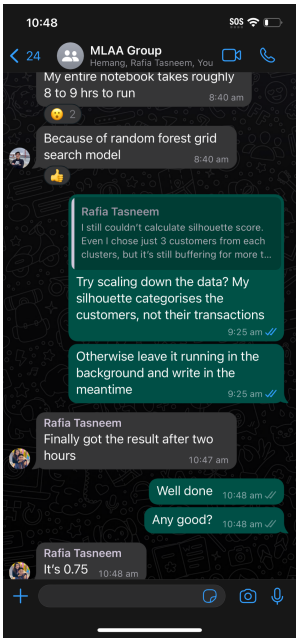
Evidence of Collaborative Effort

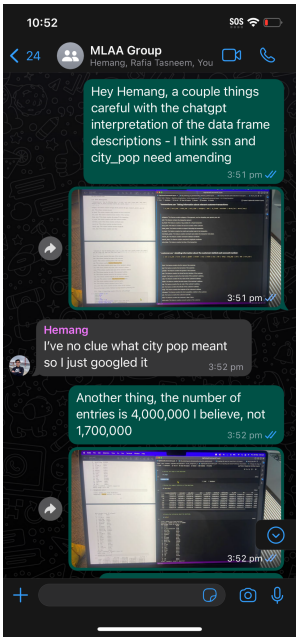
The first week was exercised, attempting to reach everyone and consolidate a plan. Unfortunately, our fourth member, “Kalp Shah” didn’t respond. The department head data scientist “Anthony So” advised our team to proceed as a group of three.

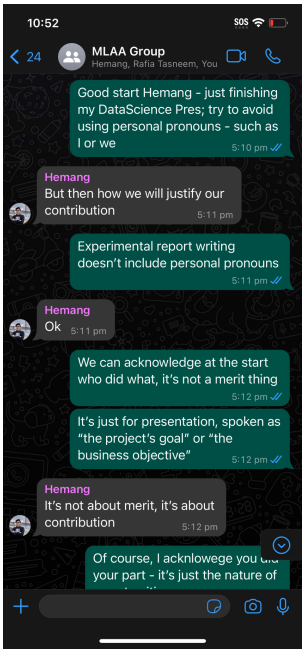
Adhering to a schedule: *This feature*

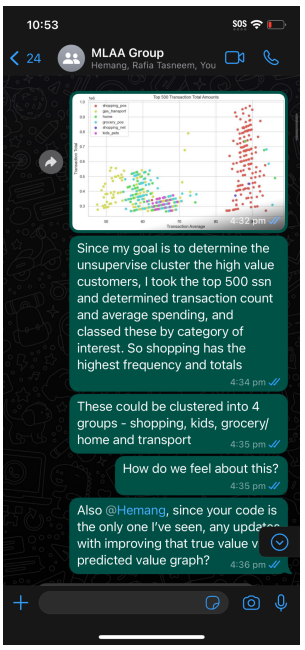
Meeting 1-2 a week: *This feature*

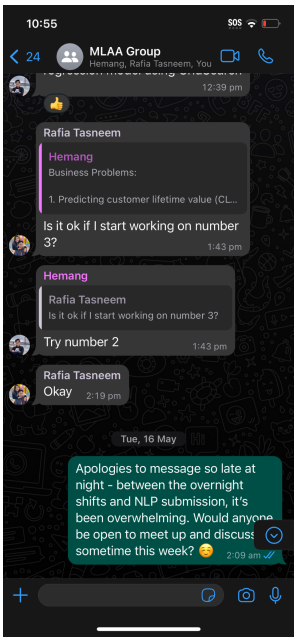
Whatsapp Chat: *This feature*

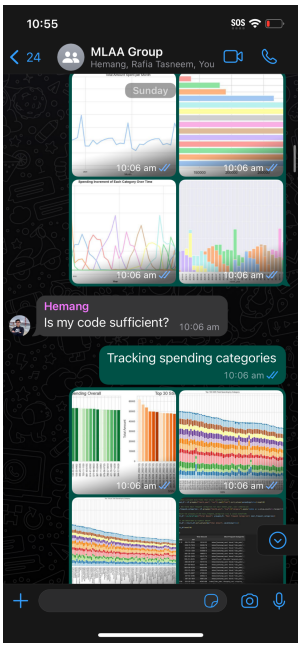


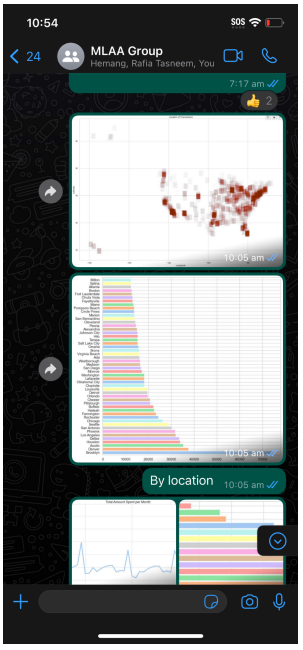












Pooling Resources & Assisting Coding blocks: *This feature*

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.*