ANALYTICS FOR CUSTOMER LIFETIME VALUE IN E-COMMERCE

By
Tam Nguyen & Uyen Tran

Introduction

In the fiercely competitive e-commerce landscape, understanding the Customer Lifetime Value (CLV) is paramount for businesses aiming for sustained growth and profitability. With the rising cost of acquiring new customers and the significant benefits of retaining existing ones, e-commerce companies must identify and invest in high-value customers to maximize profitability. The challenge lies in accurately forecasting the future value of customers from diverse segments, considering the dynamic nature of consumer behavior and market conditions. The escalating customers' expectations further necessitate personalized marketing strategies and continuous innovation in product offerings. In this light, CLV is a powerful marking measure and a financial improvement tool for any company (Gupta and Zeithaml, 2006). Historical CLV offers a retrospective view of customer behavior over time. E-commerce businesses can gain deeper insights into customer preferences, tendencies, and lifetime value evolution by analyzing past purchasing patterns, trends, and interactions. This understanding forms the foundation for informed decision-making and strategic planning. In conjunction with that, segmentation allows businesses to tailor their marketing efforts to different customer groups effectively. Therefore, segmentation based on CLV empowers e-commerce businesses to identify high-value customer segments that warrant specialized marketing campaigns. The project is driven by the critical need for e-commerce businesses to leverage data analytics for informed decision-making. Through an in-depth analysis of CLV segmentation techniques and their practical applications, we aim to provide valuable insights into how businesses can effectively leverage CLV to drive growth and achieve competitive advantage in the market. In this paper, CLV is calculated using the following formula:

CLV = Customer Value * Average Customer Lifespan

- Customer Value = Average Purchase Value * Average Frequency Rate

    - Average Purchase Value = Total Revenue over a time frame / Total number of purchases over the same time frame
    - Average Frequency Rate = Total number of purchases over a period / Total number of customers during the same period
- Average Customer Lifespan = Average number of days customers stay active / Total number of customers

Business Questions:

- Which customers will generate the highest profits for multi-category e-commerce companies?
- What is the best machine learning (ML) technique for customer segmentation in E-commerce?
- How can the analysis and modeling results be used for effective marketing strategies?

Our first objective is to optimize marketing spend by focusing on high-value customers. This means identifying and targeting customers who have the potential to generate revenue for our business. Moreover, we recognize the importance of nurturing long-term relationships with our customers and keeping them engaged with our brand. By understanding their preferences, behaviors, and needs, we can tailor our marketing messages and offers to resonate with each customer individually. Therefore, our second objective revolves around enhancing customer retention through personalized marketing strategies. To achieve these objectives, we will begin by profiling our customers based on their behaviors, demographic information, and preferences. Next, we will determine the most effective method by using parametric and non-parametric techniques for classifying customer segments for our e-commerce business.

In this report, we present the results of our analysis aimed at understanding customer segmentation and behavior in our e-commerce business. Through the application of cluster analysis, we have identified three distinct customer segments: Premium, Gold, and Silver. These segments exhibit varying levels of engagement, spending habits, and geographic distribution. Additionally, our analysis highlights the effectiveness of the Random Forest algorithm in accurately classifying customers and identifies key factors influencing segmentation. These insights provide valuable opportunities for targeted marketing, personalized experiences, and strategic decision-making to drive business growth and customer satisfaction.

Literature Review

With the advancement of digital technology, modern business is now characterized by an intensified emphasis on customer service and the cultivation of long-term relationships. The literature on Customer Lifetime Value (CLV) presents a compelling case for its critical role in e-commerce, driven by the need to understand and forecast customer behavior to inform business strategies. A study by Vanderveld et al., (2016) utilized CLV to emphasize the importance of Customer Relationship Management (CRM), in which CLV was employed to pinpoint the optimal target audiences for tailored promotional campaigns and personalized communication strategies. This process requires the calculation of CLV, which is defined as the sum of cumulative cash flows, discounted using the weighted average cost of capital, attributed to a customer throughout their entire lifetime with the firm (Kumar et al. 2004). Meanwhile, Pfeifer et al., (2005) employed indicator Customer Profitability (CP). CP represents the revenues generated during a specified period minus the costs associated with maintaining a mutually beneficial relationship with the customer. Jasek et al. (2017), on the other hand, used Customer Equity as the average Customer Lifetime Value (CLV) subtracted by acquisition costs. Other studies stated that RFM is one of the important measures for gauging CLV (Hiziroglu et al. 2018; Sun et al. 2023). Liu (2003) presented the AHP to determine the relative weighting of RFM metrics based on the subjective perceptions of decision-makers.

Probability models have been widely employed in predicting CLV. Nevertheless, with the advancement of technology and the rise of expansive e-commerce platforms has led to businesses gathering and storing vast amounts of data. Besides, recent research indicates that ML-based CLV models have the potential to surpass probability models in terms of predictive accuracy and performance. For example, Vanderveld et al. (2016) used the Random Forest model that was implemented to forecast the CLV for all users of e-commerce platforms, utilizing the same model across the board. Subsequent work by Dai (2022) extends the range of ML techniques for CLV prediction, including Linear Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Neural Network (NN). The finding suggested that RF is superior in terms of precision. Sun et al. (2023) utilized various machine learning algorithms, including but not limited to RF, NN, SVM, classification and regression trees (CART), generalized additive models (GAMs), multiple adaptive regression splines (MARS), to predict CLV under non contractual relationships. The results show Decision Tree achieved a better prediction performance. Similarly, Moro et al. (2013) employed LR, NN, decision trees (DTs), and SVM to predict the effectiveness of marketing initiatives. Their findings revealed that the NN results in highest predictive accuracy and overall performance. Martínez et al. (2016) compared logistic regression, limit learning machine, and gradient-boosting trees to forecast future purchases by customers. Their findings indicate that the performance of gradient-boosting trees was notably enhanced compared to the other methods.

Based on research above, across different datasets, the effect of machine learning algorithms varies, leading to deviations in predictions generated by individual methods. Hence, employing a comprehensive calculation approach using various techniques emerges as an optimal modeling approach characterized by robust versatility. In this context, concerning customer segmentation, this project integrates the cluster model with machine learning algorithms for data mining.

Data, Data Sources, and Characteristic

Data Gathering

For our project, the data-gathering process involves merging several key datasets to facilitate comprehensive analysis. The dataset is collected from Kaggle, comprising five separate datasets. However, to streamline our analysis and derive meaningful insights, we merged these datasets into one. We begin by merging the “CustomersData” with the “Online_Sales” dataset, linking these datasets using the common identifier “customerID”. Following this, we integrate the Tax_amount dataset into the merged dataset from the previous step, using “product_category” unique identifier for sales transactions. Next, we merge the “Marketing_Spend” dataset with the previously merged dataset, utilizing common identifiers such as “Transaction_Date”. Finally, we integrate the Discount_Coupon dataset into the merged dataset, linking discount information with sales transactions using “Product_Category” and “Month”. The final dataset consists of 52,924 observations and 20 variables. Given our business objective, we have organized our dataset to focus on individual customers. To achieve this, we have grouped our data by “CustomerID” and aggregated relevant information such as the duration of the customer’s relationship with us, the total number of transactions, the total invoice amounts, the total quantity items purchased, the customer's demographic information and the tenure of the customer's relationship with us.

Data Cleaning and Preprocessing

In our data analysis, we encountered missing values and duplicates within our dataset. To address missing values, we utilized Pandas' built-in functions such as `isnull()` and `duplicate()` to identify missing values and duplicates. The results show that missing values appear in "Coupon_Code" and "Discount_pct" variables. Besides, we use the function `describe()` to generate descriptive statistics. We can observe that "Avg_Price", "Quantity" and "Delivery_Charges" have a wide dispersion in the minimum and maximum values within our dataset. Such variability can suggest diverse ranges of values, potentially reflecting potential outliers that may require further processing to ensure the integrity of our analysis.

Treating Outliers

The presence of outliers was spotted in five variables, namely "Quantity", "Avg_Prices", "Delivery_Charges", "Offline_Spend" and "Online_Spend". For "Quantity" and "Delivery_Charges", a significant skewness of data was noticed, for which we applied a log transformation to reduce skewness. A long tail in "Quantity" indicates that most purchases involve a small number of items, however, there are also occasional bulk purchases. This could be due to a small number of customers placing large orders, or special events that lead to increased purchasing volume. Similar implications are applied for "Avg_Prices" and "Delivery Charges", where a long right tail suggests the existence of some instances with unusually high values. This could be explained by several factors such as express shipping, long-distance deliveries, large or heavy items that require special handling, or a combination of these. To deal with such outliers, we first defined a function to detect outliers using the Interquartile Range method and capped them with the corresponding lower and upper limit values. However, even after capping, 12,245 observations are still being flagged, suggesting significant variability in the dataset. As such, we have conducted a further investigation on assessing outliers in the context of e-commerce business, as well as a thorough review of outlier criteria to finally come to a conclusion of keeping these extreme values as they are indeed valid transactional information in the real-world business context. However, by keeping such values, it is essential to use modeling algorithms that are robust to outliers to avoid biased results.

Treating Missing Value

Since our missing values are categorical variables, we utilized mode imputation to replace missing values with the most frequent category, ensuring the preservation of the dominant characteristics of the data.

Data Dictionary

Attributes	**Description**
Number of Day	Number of days the customer stay active
Number of Transactions	Total number of transaction of each customer within the observed period
Total Purchase Amount	Total revenue acquired from each customer
Quantity	Total quantity purchased by each customer
Gender	Gender of the customer
Location	Location of the customer
Tenure (months)	Length of time that a customer has been actively engaging with a business or service, measured in month
AOV	Average purchase value of each customer
Churn Rate	The percentage rate at which customers stop buying product
Profit Margin	The amount of profit expected from each customer over the average customer lifespan

Table 1: Data dictionary

Data Visualization

We employed various data visualizations to gain insights into the distribution and patterns of key variables.

First, a correlation heatmap is created to understand factors tightly related to the increase in revenue to target for CLV enhancement (appendix 1). A notable correlation of 0.51 exists between “Quantity” and “Invoice”, which is expected as more items generally result in a higher total price. On the other hand, we observed a negative correlation between “Invoice and “GST”, suggesting that GST is more likely to decrease as the total purchase amount increases. The result is counterintuitive since GST is generally expected to increase in proportion to the total invoice. One explanation is the variability of products in an invoice, such that high-value invoices involve more product categories with lower GST rates (such as “Nest-USA” and “Office”). Another explanation is associated with the effect of discounts applied, which in turn affects the total GST collected. These insights help strategize pricing, discount, and sales approaches to optimize tax efficiency and revenue, which is crucial for maximizing CLV.

Next, we visualized the count of customers retained by each month using a bar chart (appendix 2). The chart suggests that certain times of the year are associated with increased purchases, as retention is shown to peak around July and start declining towards the end of the year. However, the cause of retention trends can vary, possibly affected by promotional campaigns, product seasonality, and economic factors. Hence it is important to carefully examine additional data to understand any external events that could impact buying behavior.

Similarly, a bar chart (appendix 3) was used to highlight the top five purchased product categories, identifying “office”, “apparel”, “drinkware”, “lifestyle” and “nest-USA” as having the highest demand. Moreover, we employed a pie chart (appendix 4) to visualize the distribution of average prices based on coupon usage. By comparing the proportions of spending with and without coupon usage and coupon clicks, we depicted the proportion of each group as about 32.8%, 34%, and 33.2% respectively. Moreover, we used a scatter plot (appendix 5) to examine the relationship between total purchase price and delivery charges. Our visualization revealed a negative correlation between these two variables, indicating that higher delivery charges tend to correspond with lower total purchase amounts.

A histogram (appendix 6) was used to show the distribution of customer tenure in months. The trend line indicates that most customers are actively engaged with the company in 30 months. This implies that there are critical tenure milestones where customers might require additional engagement to continue their relationship with the company. Lastly, using pie charts (appendix 7), we discovered that 62.4% of total customers are female, who are shown to spend 1.6 times more than male customers. This information can be pivotal for segmenting marketing strategies and product offerings. In particular, females can be considered the primary target demographic for the company.

Upon visualizing the data, we observed outliers in four significant features: "Avg_Price," "Quantity," "Delivery_Charges", “Online_Spend” and “Offline_Spend”. Utilizing box plots (appendix 8), we identified instances where data points deviated substantially from the overall trend.

Regarding the number of transactions, the bar chart (appendix 9) shows that Apparel & Net-USA products comprise a larger proportion of the total count of products in the dataset. Moreover, the pie chart (appendix 10) illustrates a significant portion of transactions is attributed to female customers, indicating their significant presence in our dataset. The map (appendix 11) highlights Chicago and California as key hubs of transactional engagement, suggesting strong market demand and customer engagement in these regions.

Methodology

Define Tasks

Based on our business questions, the three main tasks that need to be addressed include:

- Segmentation of customers based on the degree of CLV and RFM using K-means clustering.
- Build a classification model based on the segmentation result and determine the best model.
- Apply association rules to find product combinations preferred by each customer and their corresponding segment for targeted marketing efforts.

Technique Selection

For segmentation, we employ an unsupervised ML technique–K-Means clustering, using four variables: Recency, Frequency, Monetary value, and CLV, and profiling each segment based on behavior, demographic, and preference information. For model building, parametric methods such as Multinomial Logistic Regression, and non-parametric advanced machine learning algorithms such as Classification Tree, Random Forest, and XGBoost are utilized to determine the best model. The use of both parametric and non-parametric techniques for model building offers a comprehensive approach to capturing the underlying patterns and complexity within the data. Particularly, the parametric technique provides a greater degree of interpretability in terms of variable significance output with less computationally intensive, while non-parametric techniques are capable of handling complex and non-linear relationships between the target variable and features, hence yielding higher accuracy. A drawback of non-parametric techniques is that they are quite prone to overfitting and are computationally intensive. This is why we include XG-Boost, an advanced regularization form of Gradient Boosting, as one of our chosen algorithms which helps with over-fitting control and faster training parallelism across clusters.

Data Partition and Balancing

The dataset was partitioned into training (70%) and testing (30%) subsets for model evaluation. Upon partitioning, a substantial class imbalance was noted within the training dataset, in which Cluster 0 represented 15.94% of the total dataset, Cluster 1 accounted for 83.58%, while Cluster 2 constituted only 0.47%.

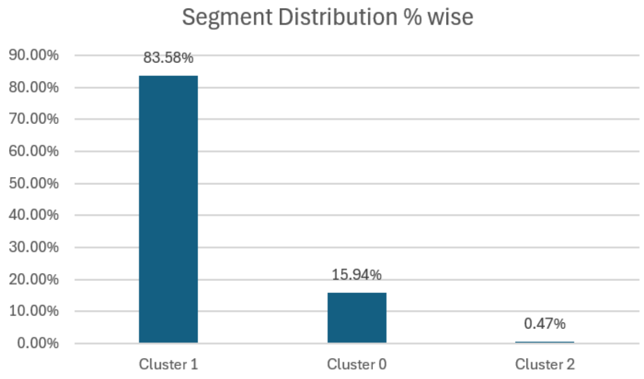


Figure 1: Segment distribution percentage wise

Therefore, SMOTE-Tomek was applied to the train set to balance all categories and ensure model predictability. SMOTE (Synthetic Minority Oversampling technique) works by creating synthetic instances that are similar to the existing minority class samples while Tomek links help remove pairs of instances that create ambiguous boundaries between two classes.

Applying Data Mining Techniques

For the first task of clustering, the elbow method was used to find the optimal number of clusters, which resulted in three clusters in our case. We then fit the K-Means model including four variables ‘Recency’, ‘Frequency’, ‘Monetary’, and ‘CLV’ with the chosen number of clusters, and thereby group the data frame by cluster number to examine the RFM and CLV characteristic in each segment (appendix 13).

Other variables such as ‘Gender’, ‘Location’, ‘AOV’, ‘Quantity’, and ‘Tenure’ were also added for the purpose of profiling customer segments.

For the second task of classifying customer segmentation and defining the best model, we created a new variable “Segment Label” based on the clustering results, in which Cluster 0, Cluster 1 and Cluster 2 are the three categories of the dependent variable. Meanwhile, independent variables such as “Gender”, “Location”, “Tenure”, “Product Category” and “Coupon Status” were incorporated to clarify their impact on customer segmentation. The model is written as:

P(Customer Segment | Gender, Location, AOV, Quantity, Tenure)

The model is then validated based on the test set using various ML techniques and chooses the effective model based on their accuracy. Lastly, we aim to examine the probability of products being purchased together, utilizing association rules. These rules analyze transactions to identify associations between items to spot pairs of items that are frequently purchased together.

Empirical Results

After conducting an analysis to determine the optimal number of clusters using K-Means, our findings indicate that three clusters provide the optimal segmentation (appenix 12).

Cluster	Quantity	Tenure	AOV	Recency	Frequency	Monetary	CLV	Location	Gender

No									
0	87	26	97	152	21	1968	153	California	F
1	2807	30	125	72	501	61713	198	Chicago	F
2	476	26	123	107	98	10809	194	California	F

Table 2: Cluster profiling result

- Cluster 0 - "Gold": a segment with strong revenue potential that has not yet been fully realized. Their AOV reflects a propensity for higher-value purchases. In terms of recency and frequency, they fall within the 'Semi-Recent' and 'Semi-Frequent' categories, respectively, indicating regular but not overly frequent interactions with our offerings. Despite their moderate monetary contributions, their CLV remains at an average level. A significant portion of this cluster is located in California, with a notable presence of female customers.

- Cluster 1 - "Silver": a segment of customers who engage with our products and services less frequently compared to other clusters. These customers exhibit lower levels of engagement over time. Their AOV has a tendency towards lower-value transactions. Classified as 'Least Recent' and 'Least Frequent', these customers demonstrate infrequent interactions with our offerings. Their monetary contributions and CLV fall within the low range, indicating a potential for targeted efforts to increase their engagement and loyalty. This cluster is located in California, with a predominance of female customers.

- Cluster 2 - "Premium": a segment with the most valuable and highly engaged customer segment. These customers demonstrate a strong and enduring commitment to our brand. AOV demonstrates a preference for premium products and services. Classified as 'Most Recent' and 'Most Frequent', these customers exhibit frequent and recent interactions with our offerings, indicating a strong and positive relationship between these customers and the products. Their monetary contributions and CLV are classified as high, underscoring their significant impact on our revenue and profitability. A substantial portion of this cluster is located in Chicago, with a predominant presence of female customers.

An interesting observation emerges from our analysis: across all three segments, the majority of shoppers are female. This trend aligns with instinctive expectations, given that the company's primary product offering revolves around “Apparel”.

In terms of model performance, Random Forest demonstrates superior performance in terms of predicting customer segments. The model results can be found in appendix 14. When comparing True Positive Rate versus False Positive Rate in the ROC Curve of Random Forest, the green curve representing Premium customers consistently outperforms the other curves, indicating its ability to accurately identify true positives. While the Gold curve shows good classification capability with an AUC of 0.87, the Silver curve has the lowest AUC at 0.85, suggesting it may be more challenging for the model to distinguish Silver customers accurately. Besides, the model suggests that ‘Quantity’, ‘AOV’ and ‘Tenure’ are most significant in segmenting customers, as shown in figure below.

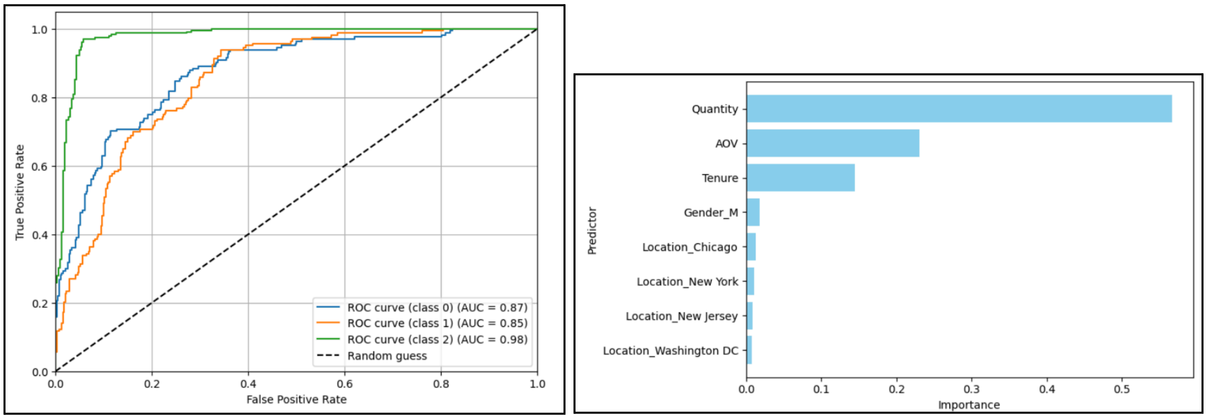


Figure 2: ROC curve and Variable Importance obtained from RF

Our analysis of association rules reveals interesting patterns in customer purchasing behavior (appendix 15).

Antecedents	**Consequents**	**Support**	**Confidence**	**Lift**
Drinkware	Apparel	0.045010	0.446910	1.377784
Apparel	Drinkware	0.045010	0.138762	1.377784

Lifestyle	Apparel	0.033079	0.484229	1.492836
Apparel	Lifestyle	0.033079	0.101981	1.492836
Office	Apparel	0.062128	0.441577	1.361343
Apparel	Office	0.062128	0.191536	1.361343
Office	Drinkware	0.046287	0.328985	3.266516
Drinkware	Office	0.046287	0.459588	3.266516
Office	Lifestyle	0.035114	0.249575	3.653381
Lifestyle	Office	0.035114	0.514019	3.653381

Table 3: Association rule

For instance, in the first rule, the likelihood of obtaining Apparel is 1.3 times higher compared to selecting randomly, while the probability of obtaining Apparel when using the rule is 0.4. Moreover, Office supplies demonstrate notable associations with both Apparel and Drinkware, with support values of 0.062 and 0.046, respectively, suggesting that these items are frequently purchased together.

Combo	**The customers who may buy these combo**
Drinkware + Apparel	ID 12356| Female | Chicago | Silver Member
Lifestyle + Apparel	ID 12347 | Male | New York | Gold Member
Office + Apparel	ID 12348 | Male | California | Silver Member

Table 4: Example of customer ID and itemsets integration

To derive actionable insights for the business, we integrate each customer's unique ID with their specific combination preferences. For instance, itemsets Drinkware and Appearel were previously bought several time by customer ID 12356, who is a female from Chicago and is a Silver Member. With the given information regarding each customer preferred itemsets and their member type, the company can fine-tune marketing strategies that target specific customer segments more effectively.

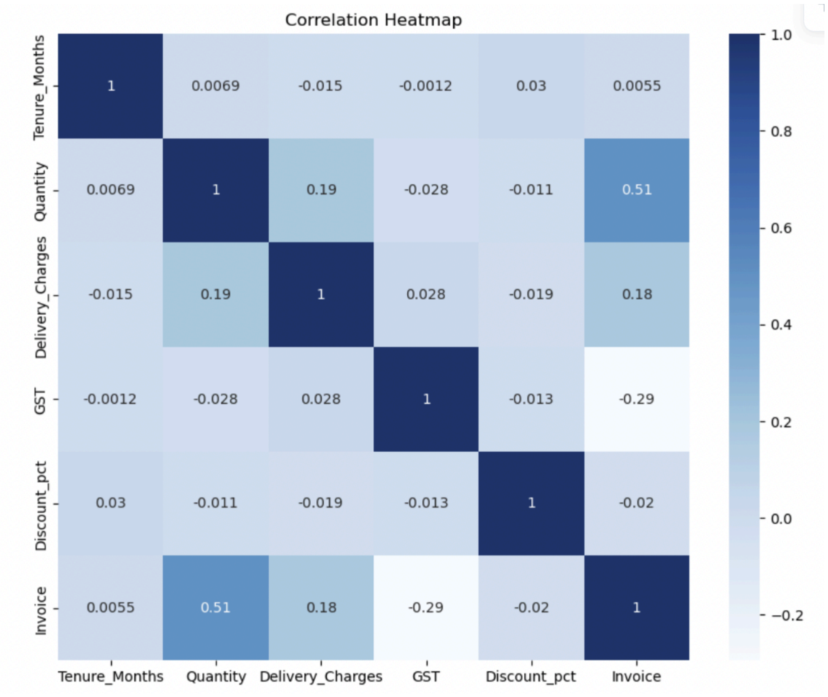
Conclusions and Recommendations

In conclusion, our analysis has provided valuable insights into customer behavior and preferences. The segmentation into Premium, Gold, and Silver clusters reveals distinct customer behaviors: Premium customers engage frequently, spending significantly. Gold customers shop less often but make valuable purchases and contribute boosting revenue. Silver customers shop infrequently and contribute less monetarily. Through segmentation analysis, we have identified these distinct customer segments that allow us to tailor our marketing efforts and customer retention strategies accordingly. Additionally, our analysis underscores the effectiveness of the Random Forest model, which yields high accuracy in customer segmentation. Furthermore, the model identifies “Quantity”, “AOV” and “Tenure” as the most important predictors in distinguishing between customer segments. Furthermore, our exploration of association rules has shed light on product associations and customer preferences, guiding us in optimizing product placement and targeted marketing campaigns. By integrating unique customer IDs with combination preferences, we have gained deeper understanding of specific customer groups, empowering us to personalize our marketing strategies effectively.

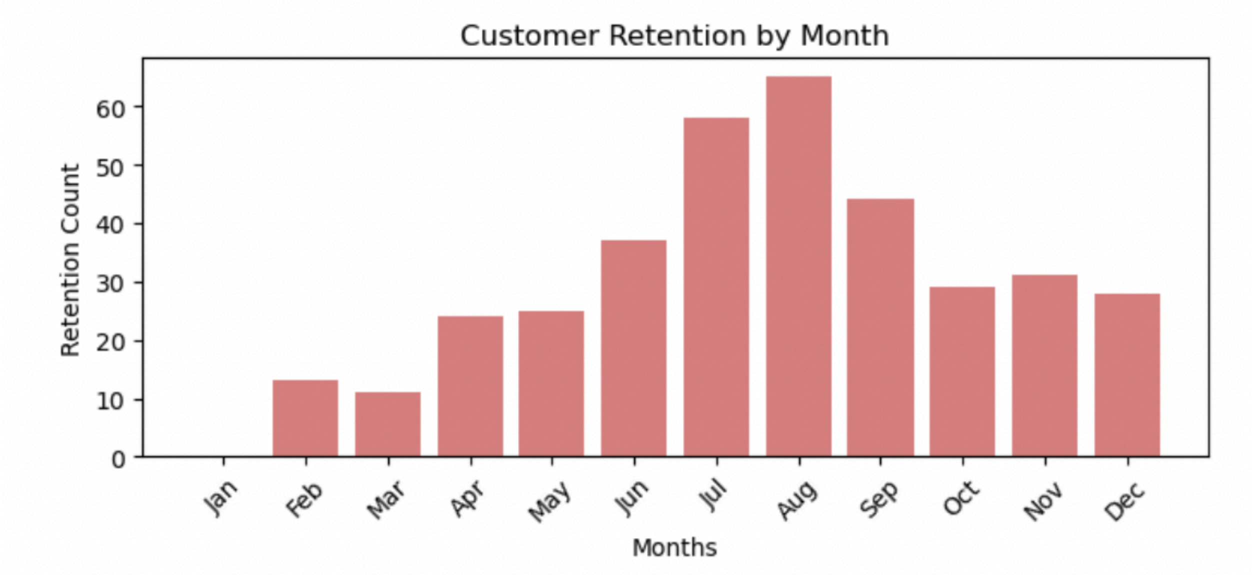
From the analysis, we recommend implementing several targeted strategies. The first step is to adopt the classification model to understand the type of customer, from which personalized advertisement campaigns can be implemented. This include tailoring email content and compelling ads showcasing relevant product combinations to their segment. By directing users to optimized landing pages, we can maximize conversion rates. Secondly, it is essential to enhance the dynamic website content; customizing the displayed content based on segment preferences will amplify user experience and drive conversions. Additionally, an approach to pricing strategies, such as premium pricing for high-CLV customers and competitive pricing for low-CLV, can optimize revenue generation. To foster customer loyalty, retention strategies should be diversified, including limited-time deals, bundle offers, or discounts for the Silver segment, while offering personalized experiences and exclusive benefits to the Premium group. Lastly, basket product development initiatives can be made, recommending combo products aligned with customer preferences, further enhancing their shopping experience and maximizing basket size.

Appendix

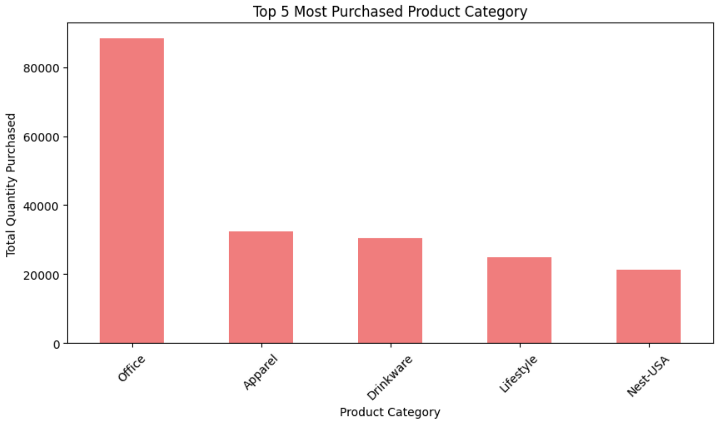
Appendix 1 - Correlation Heatmap



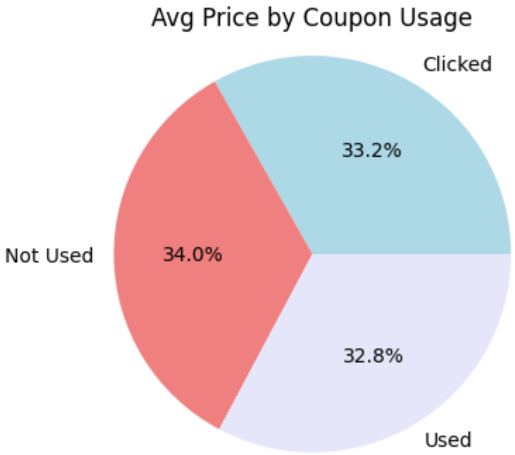
Appendix 2 - Customer Retention by Month



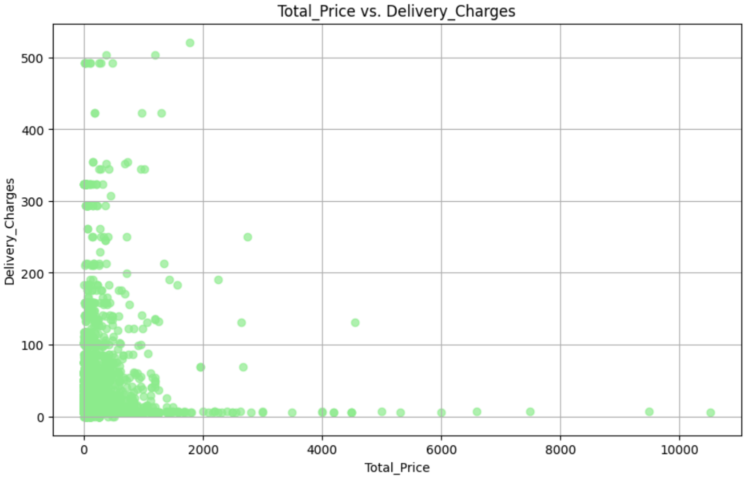
Appendix 3 - Bar chart of top five purchased product categories



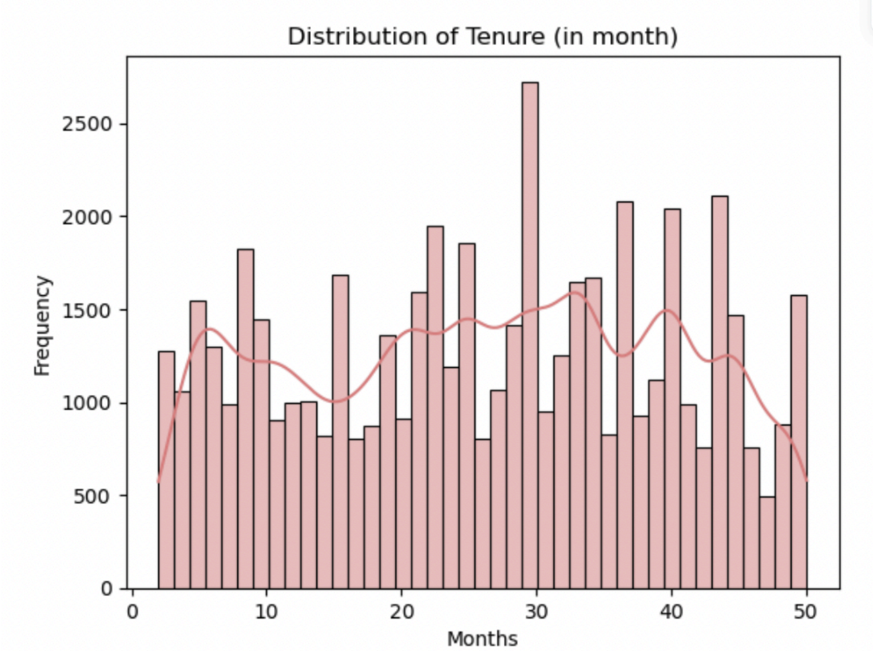
Appendix 4 - Pie chart of average prices based on coupon usage



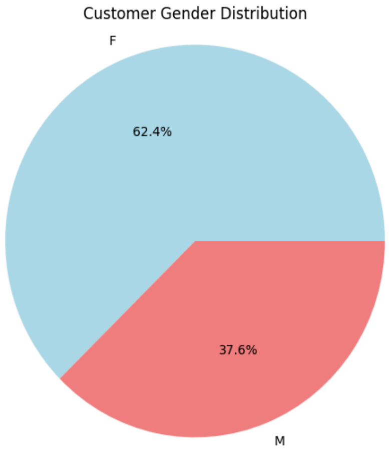
Appendix 5 - Scatter plot of total purchase price and delivery charges



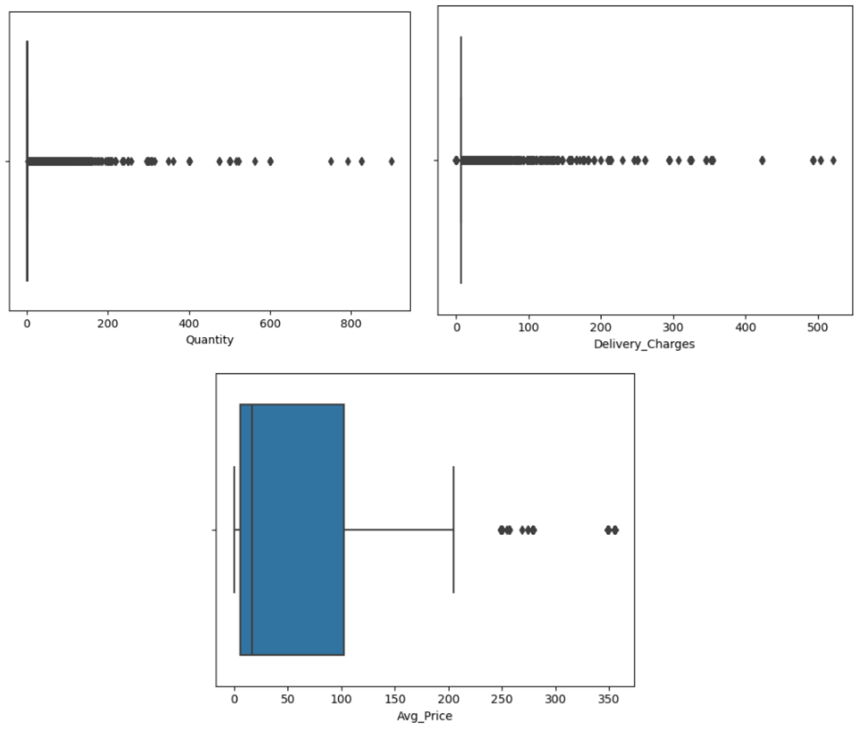
Appendix 6 - Customer Tenure in Month

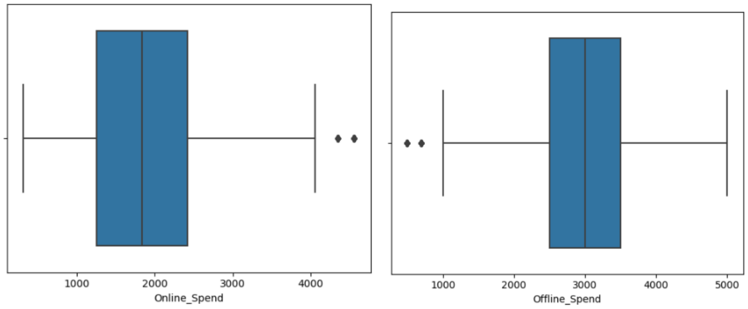


Appendix 7 - Gender Distribution

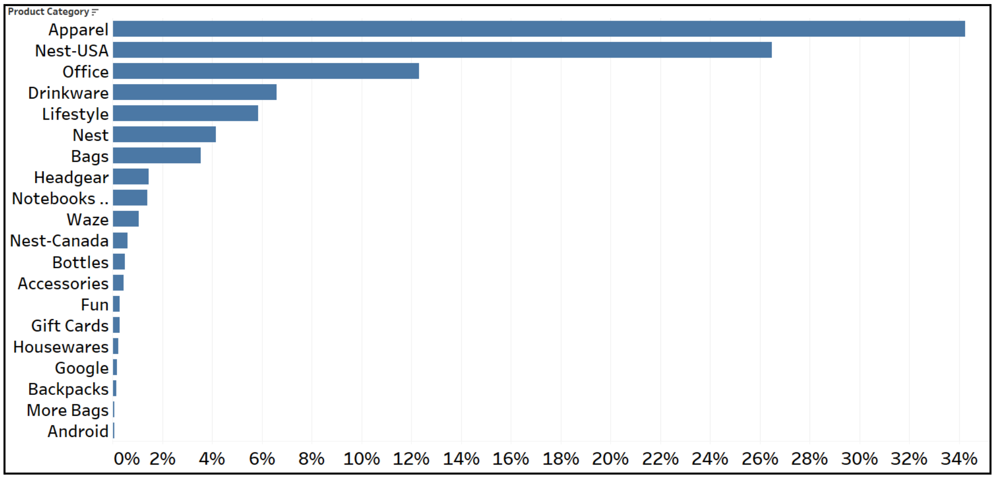


Appendix 8 - Box plot of "Avg_Price," "Quantity," "Delivery_Charges", “Online_Spend” and “Offline_Spend”

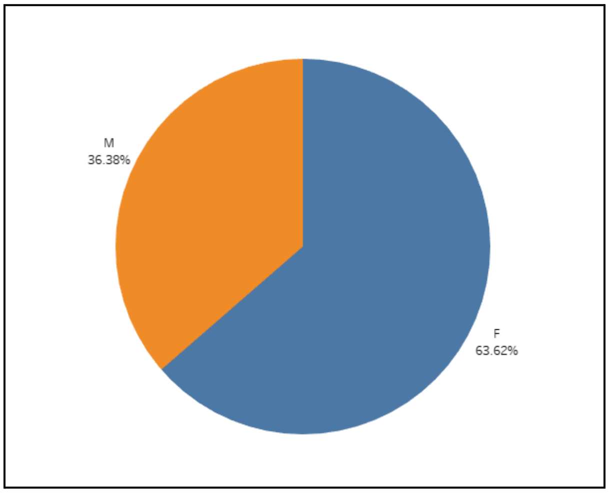




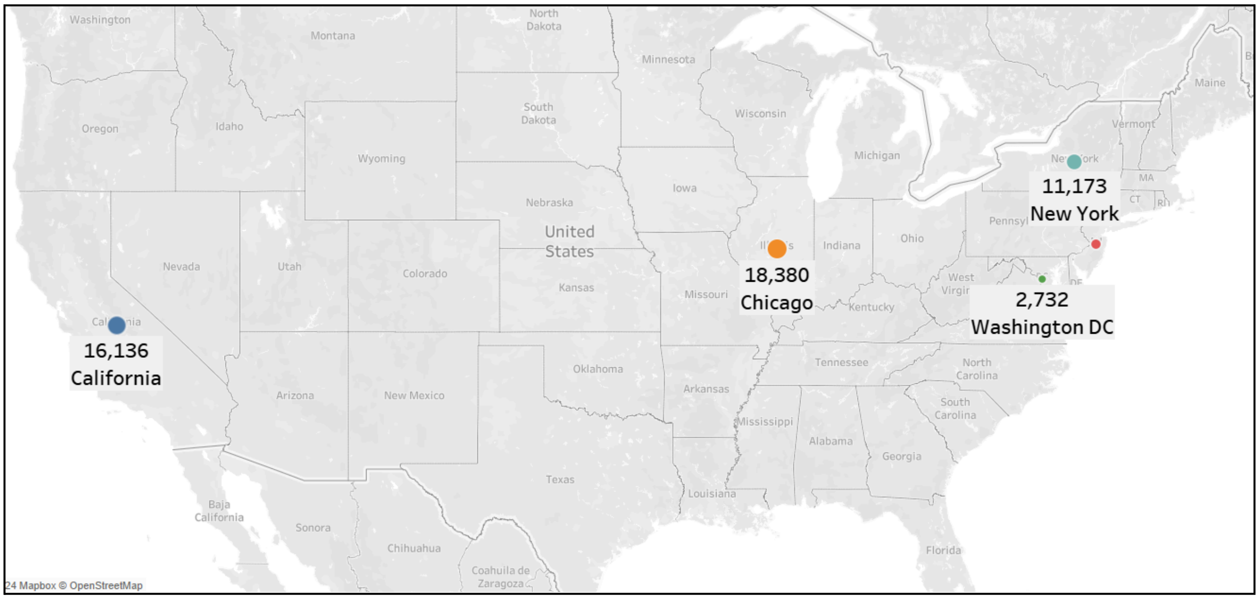
Appendix 9 - Distribution Transaction by Product Category



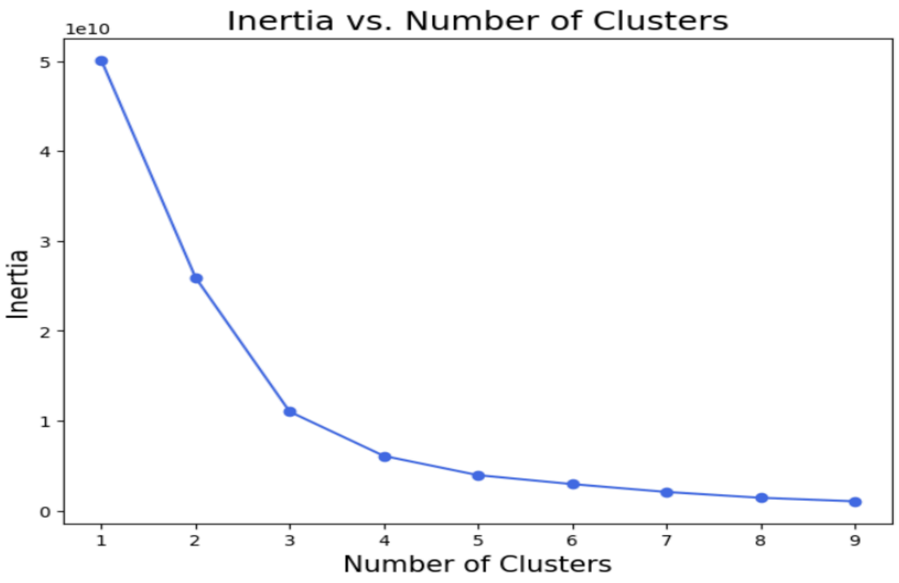
Appendix 10 - Distribution Transaction by Gender



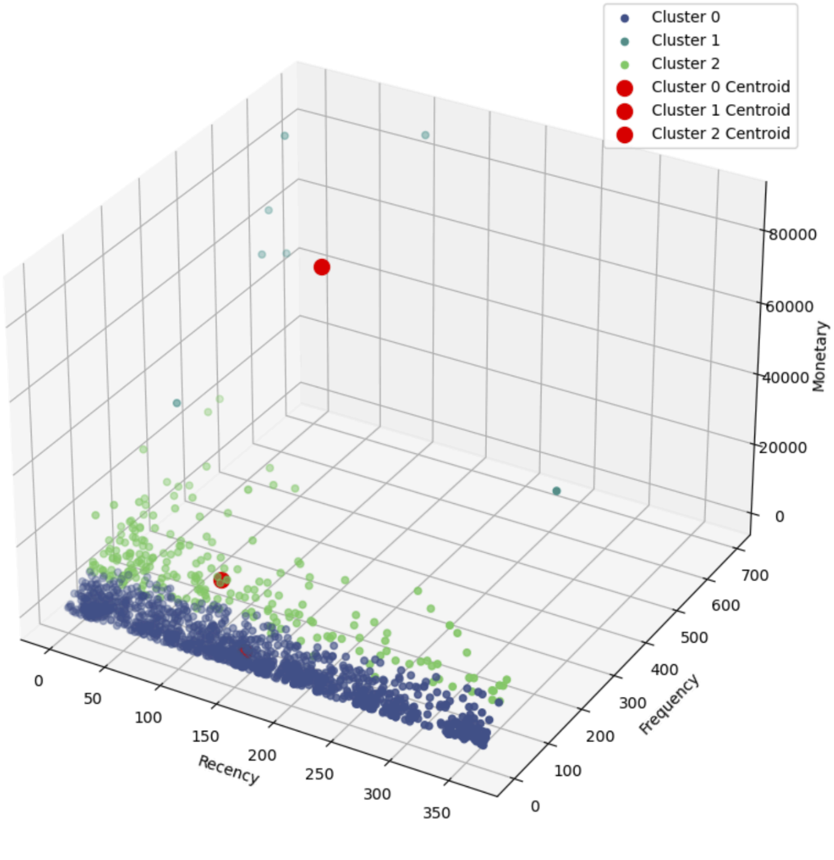
Appendix 11- Transaction by Location



Appendix 12- Elbow Curve plot

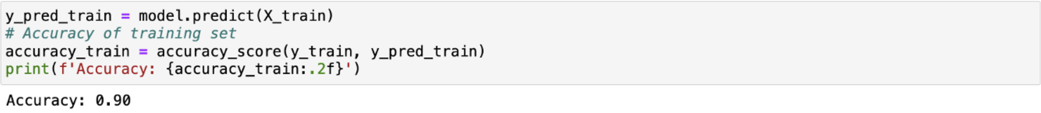


Appendix 13 - Visualization of Clusters



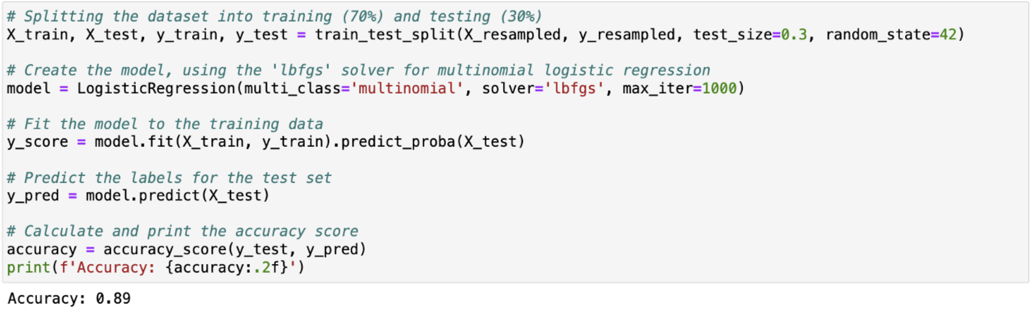
Appendix 14 - Model Accuracy

a) Multinomial Logistic Regression



```
y_pred_train = model.predict(X_train)
# Accuracy of training set
accuracy_train = accuracy_score(y_train, y_pred_train)
print(f'Accuracy: {accuracy_train:.2f}')
```

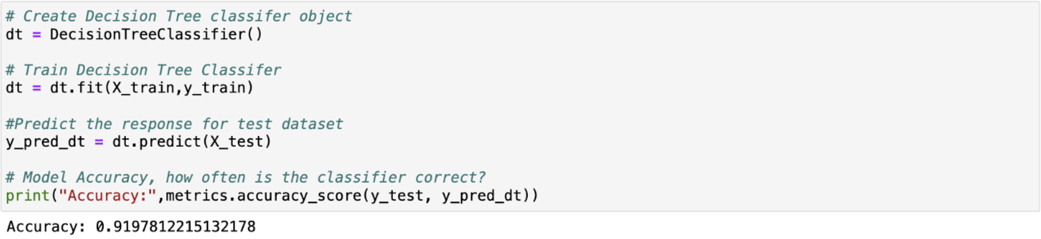
```
Accuracy: 0.90
```



```
# Splitting the dataset into training (70%) and testing (30%)
X_train, X_test, y_train, y_test = train_test_split(X_resampled, y_resampled, test_size=0.3, random_state=42)

# Create the model, using the 'lbfgs' solver for multinomial logistic regression
model = LogisticRegression(multi_class='multinomial', solver='lbfgs', max_iter=1000)

# Fit the model to the training data
y_score = model.fit(X_train, y_train).predict_proba(X_test)

# Predict the labels for the test set
y_pred = model.predict(X_test)

# Calculate and print the accuracy score
accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy: {accuracy:.2f}')
```

```
Accuracy: 0.89
```

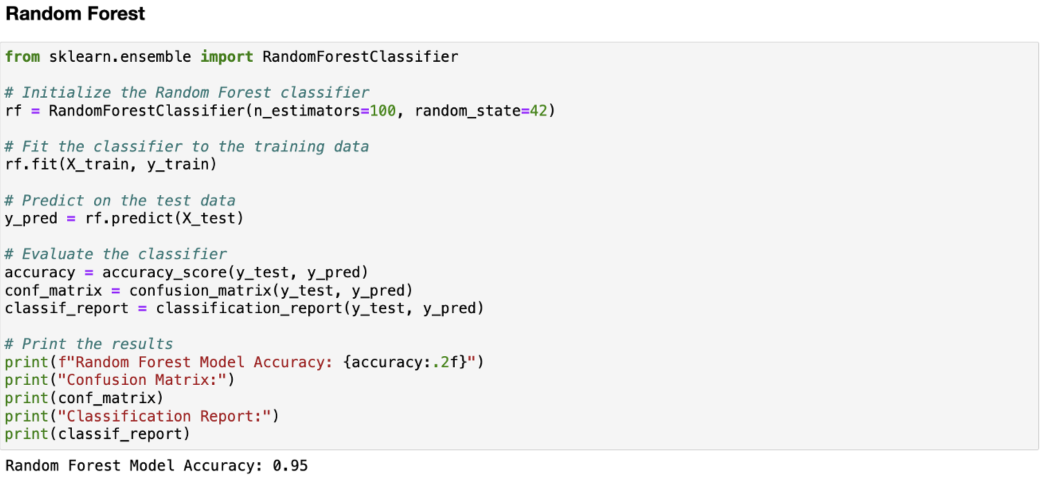
b) Classification tree



```
# Create Decision Tree classifer object
dt = DecisionTreeClassifier()

# Train Decision Tree Classifer
dt = dt.fit(X_train,y_train)

#Predict the response for test dataset
y_pred_dt = dt.predict(X_test)

# Model Accuracy, how often is the classifier correct?
print("Accuracy:",metrics.accuracy_score(y_test, y_pred_dt))
```

```
Accuracy: 0.9197812215132178
```

c) Random Forest

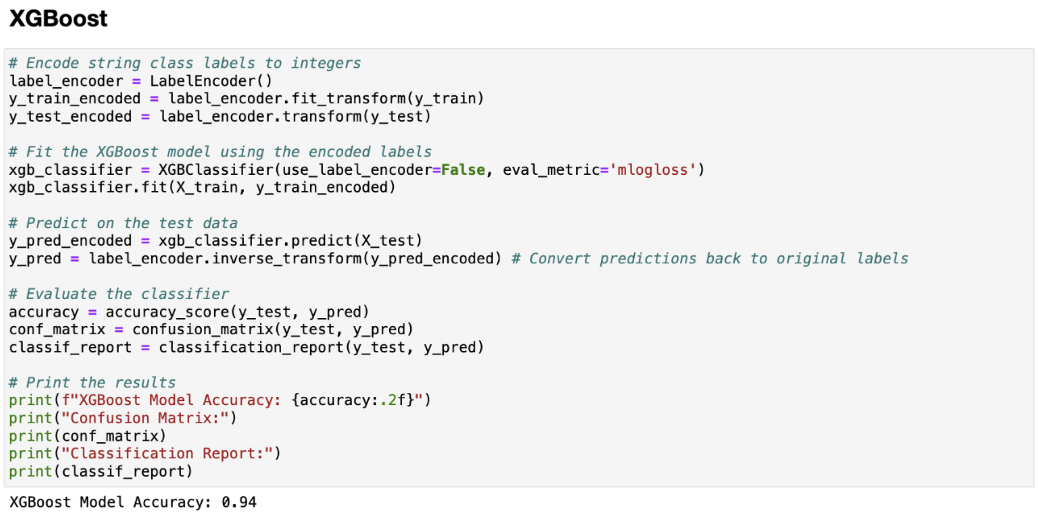


Random Forest

```
from sklearn.ensemble import RandomForestClassifier

# Initialize the Random Forest classifier
rf = RandomForestClassifier(n_estimators=100, random_state=42)

# Fit the classifier to the training data
rf.fit(X_train, y_train)

# Predict on the test data
y_pred = rf.predict(X_test)

# Evaluate the classifier
accuracy = accuracy_score(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)
classif_report = classification_report(y_test, y_pred)

# Print the results
print(f"Random Forest Model Accuracy: {accuracy:.2f}")
print("Confusion Matrix:")
print(conf_matrix)
print("Classification Report:")
print(classif_report)
```

```
Random Forest Model Accuracy: 0.95
```

d) XG-Boost

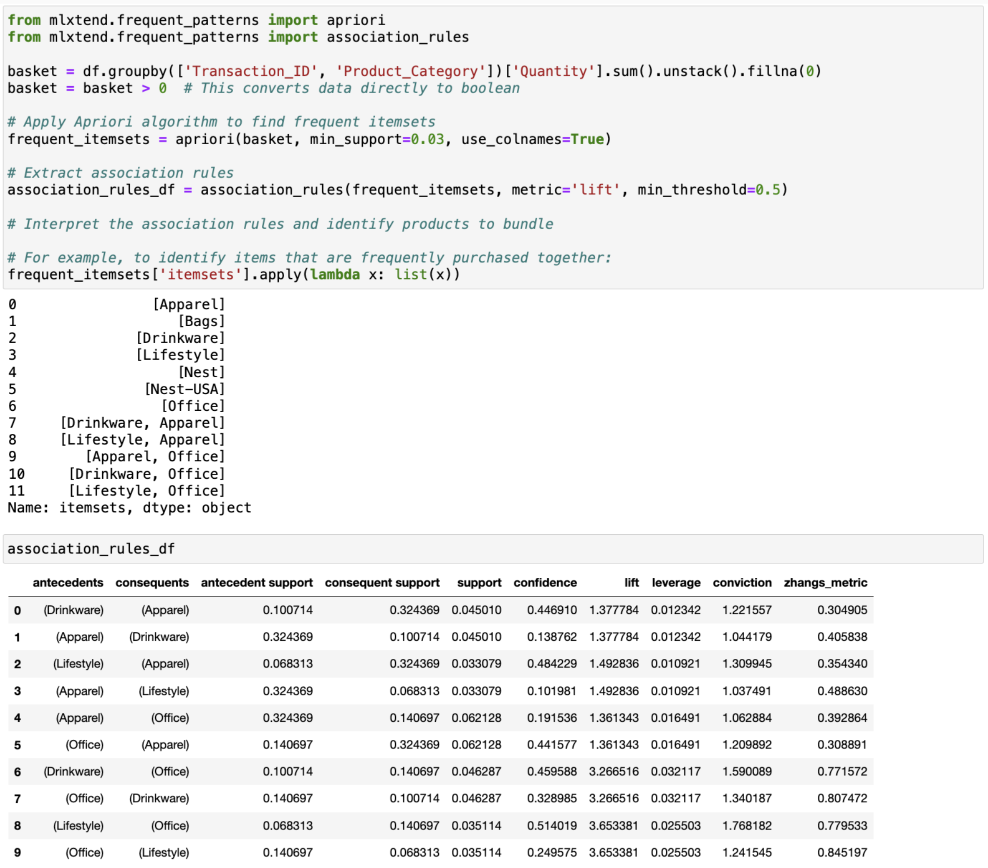


XGBoost

```
# Encode string class labels to integers
label_encoder = LabelEncoder()
y_train_encoded = label_encoder.fit_transform(y_train)
y_test_encoded = label_encoder.transform(y_test)

# Fit the XGBoost model using the encoded labels
xgb_classifier = XGBClassifier(use_label_encoder=False, eval_metric='mlogloss')
xgb_classifier.fit(X_train, y_train_encoded)

# Predict on the test data
y_pred_encoded = xgb_classifier.predict(X_test)
y_pred = label_encoder.inverse_transform(y_pred_encoded) # Convert predictions back to original labels

# Evaluate the classifier
accuracy = accuracy_score(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)
classif_report = classification_report(y_test, y_pred)

# Print the results
print(f"XGBoost Model Accuracy: {accuracy:.2f}")
print("Confusion Matrix:")
print(conf_matrix)
print("Classification Report:")
print(classif_report)
```

```
XGBoost Model Accuracy: 0.94
```

Appendix 15 - Association Rules



```
from mlxtend.frequent_patterns import apriori
from mlxtend.frequent_patterns import association_rules

basket = df.groupby(['Transaction_ID', 'Product_Category'])['Quantity'].sum().unstack().fillna(0)
basket = basket > 0  # This converts data directly to boolean

# Apply Apriori algorithm to find frequent itemsets
frequent_itemsets = apriori(basket, min_support=0.03, use_colnames=True)

# Extract association rules
association_rules_df = association_rules(frequent_itemsets, metric='lift', min_threshold=0.5)

# Interpret the association rules and identify products to bundle

# For example, to identify items that are frequently purchased together:
frequent_itemsets['itemsets'].apply(lambda x: list(x))
```

```
0                [Apparel]
1                   [Bags]
2              [Drinkware]
3              [Lifestyle]
4                   [Nest]
5               [Nest-USA]
6                 [Office]
7     [Drinkware, Apparel]
8     [Lifestyle, Apparel]
9        [Apparel, Office]
10     [Drinkware, Office]
11     [Lifestyle, Office]
Name: itemsets, dtype: object
```

```
association_rules_df
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(Drinkware)	(Apparel)	0.100714	0.324369	0.045010	0.446910	1.377784	0.012342	1.221557	0.304905
1	(Apparel)	(Drinkware)	0.324369	0.100714	0.045010	0.138762	1.377784	0.012342	1.044179	0.405838
2	(Lifestyle)	(Apparel)	0.068313	0.324369	0.033079	0.484229	1.492836	0.010921	1.309945	0.354340
3	(Apparel)	(Lifestyle)	0.324369	0.068313	0.033079	0.101981	1.492836	0.010921	1.037491	0.488630
4	(Apparel)	(Office)	0.324369	0.140697	0.062128	0.191536	1.361343	0.016491	1.062884	0.392864
5	(Office)	(Apparel)	0.140697	0.324369	0.062128	0.441577	1.361343	0.016491	1.209892	0.308891
6	(Drinkware)	(Office)	0.100714	0.140697	0.046287	0.459588	3.266516	0.032117	1.590089	0.771572
7	(Office)	(Drinkware)	0.140697	0.100714	0.046287	0.328985	3.266516	0.032117	1.340187	0.807472
8	(Lifestyle)	(Office)	0.068313	0.140697	0.035114	0.514019	3.653381	0.025503	1.768182	0.779533
9	(Office)	(Lifestyle)	0.140697	0.068313	0.035114	0.249575	3.653381	0.025503	1.241545	0.845197

References

1. Benk, G.Y.; Badur, B.; Mardikyan, S. (2022). A New 360° Framework to Predict Customer Lifetime Value for Multi-Category E-Commerce Companies Using a Multi-Output Deep Neural Network and Explainable Artificial Intelligence.

2. Dai, Xinqian (2022). Customer Lifetime Value Analysis Based on Machine Learning: Proceedings of the 6th International Conference on Information System and Data Mining. ACM Other Conferences.

3. Gupta, S. and Zeithaml, V. (2006). 'Customer Metrics and Their Impact on Financial Performance', Special 25th Anniversary Issue. Marketing Science, 25(6), pp.718–739.

4. Hızıroğlu, A., Sisci, M., Cebeci, H. İ., & Seymen, O. F. (2018). An Empirical Assessment of Customer Lifetime Value Models within Data Mining. Baltic Journal of Modern Computing, 6(4).

5. Jasek, P.; Vrana, L.; Sperkova, L.; Smutny, Z.; Kobulsky, M. (2017). Modeling and Application of Customer Lifetime Value in Online Retail.

6. Kumar, V., Ramani, G., & Bohling, T. (2004). Customer lifetime value approaches and best practice applications. Journal of Interactive Marketing, 18(3), 60–72.

7. Liu, D.; Shih, Y. (2003). Integrating AHP and data mining for product recommendation based on customer lifetime value. Information & Management

8. Martínez, A.; Schmuck, C.; Pereverzyev, S.; Pirker, C.; Haltmeier, M. (2016). A machine learning framework for customer purchase prediction in the non-contractual setting. European Journal of Operational Research.

9. Moro, S.; Cortez, P.; Rita, P. (2013). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems.

10. Pfeifer, E.P.; Haskins, E.M.; Conroy, R. (2005). Customer Lifetime Value, Customer Profitability, and the Treatment of Acquisition Spending. Journal of Managerial Issues.

11. Sun, Y.; Liu, H.; Gao, Y. (2023). Research on customer lifetime value based on machine learning algorithms and customer relationship management analysis model. Cell Symposia.

12. Vanderveld, A.; Pandey, A.; Han, A.; Parekh, R. (2016). An Engagement-Based Customer Lifetime Value System for E-commerce. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.