Customer Segmentation and Predictive Marketing of an Online Retail company in the UK

Viet Tuan Dinh 's Report

1. The Overview

In today's dynamic retail environment, the burgeoning realm of online shopping has propelled the accumulation of extensive customer data, sparking interest in leveraging machine learning for customer segmentation and predictive marketing. Although RFM-based customer segmentation and Customer Lifetime Value prediction have been explored across diverse sectors, their application within the UK online retail market still needs to be explored.

Previous research has predominantly focused on clustering methods for RFM-based segmentation, demonstrating the predictive capabilities of RFM variables for customer lifetime value. However, a notable gap exists concerning the prediction of the next purchase period, a pivotal aspect for data-informed decision-making and business expansion.

This MSc project endeavours to harness RFM for customer segmentation while developing predictive models for Customer Lifetime Value and the next purchase period within the UK online retail landscape. Employing K-means clustering for segmentation and supervised classification algorithms such as XGBoost, Support Vector Machine, and Random Forest for predictive modelling, this project seeks to deliver robust insights for businesses aiming to capitalize on machine learning for enhanced growth and profitability in online retail.

In the subsequent sections, the report will delve into the theoretical underpinnings of the algorithms utilized to accomplish these objectives.

1.1. RFM method

The Recency, Frequency, and Monetary (RFM) method is a powerful tool for customer segmentation and is widely employed in marketing and sales endeavours. RFM analysis enables businesses to tailor their marketing strategies effectively by categorising customers based on their purchase history. This report explores the application of the RFM method in predicting Customer Lifetime Value (CLV) and the next purchase period. Leveraging the simplicity and effectiveness of RFM, this study endeavours to provide actionable insights for businesses seeking to enhance customer engagement and drive growth in the online retail

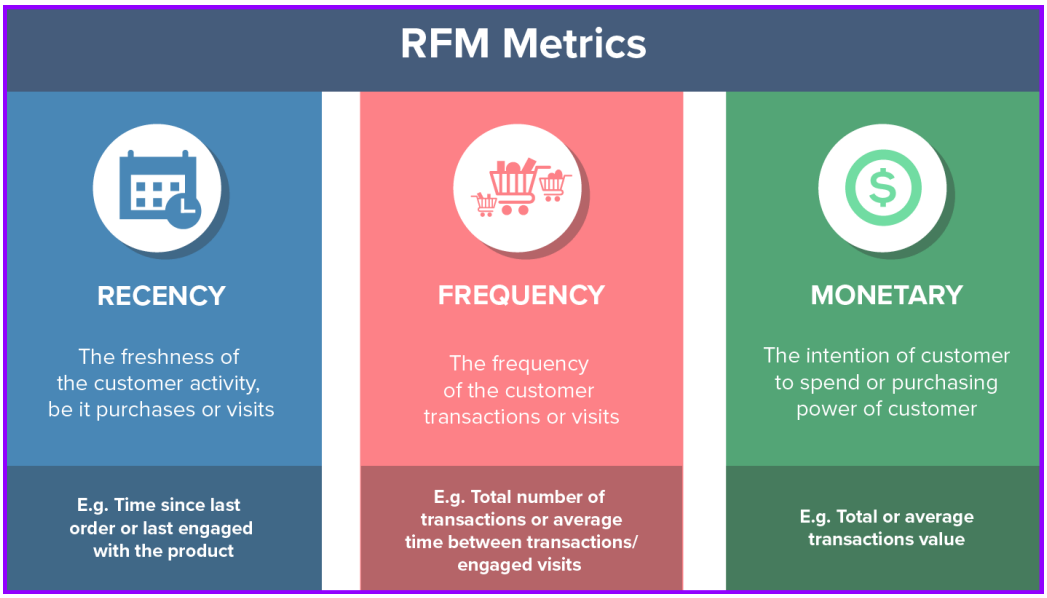
industry. Subsequent sections will delve into earlier segmentation studies utilising RFM and discuss the approach to addressing their limitations.

RFM is a customer segmentation model based on the Customer360 philosophy. Data on the transaction history between customers and businesses is collected and analyzed based on three key factors: Recency, Frequency, and Monetary.

1.2 RFM Metrics

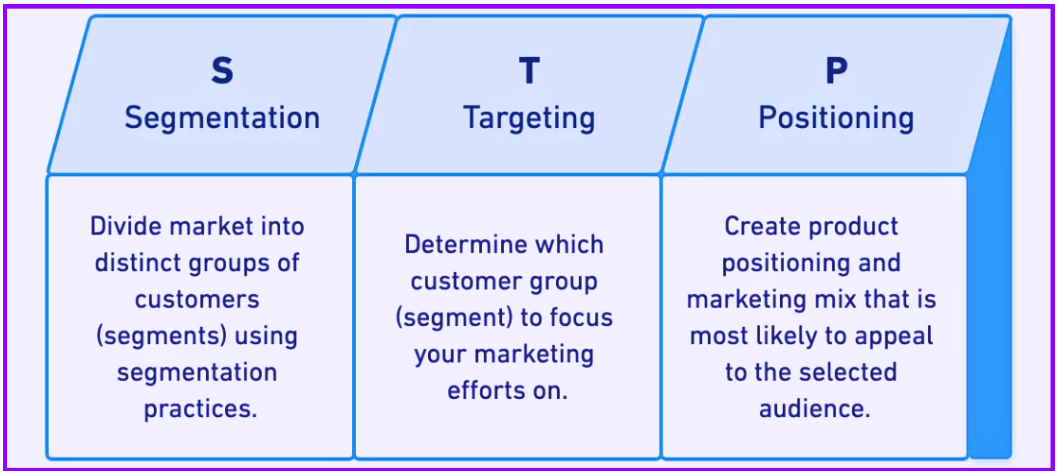
- Recency: The most recent time a customer purchased or used a service.
- Frequency: How often a customer makes purchases or uses services.
- Monetary: The amount of money a customer has spent on purchases or services.

Recency, Frequency, and Monetary are three crucial factors in quantifying customer behaviour and interactions with the business. Among these, Frequency and Monetary are determinants of the Customer Lifetime Value, while Recency influences Customer Retention.



1.3 Benefits of RFM Metris

The RFM model helps businesses enhance customer retention by aiding them in understanding fundamental marketing principles:



The RFM model allows businesses to gain crucial insights into customers that they can act upon, thereby shaping business strategies around these insights. This model enables companies to understand the significance of their brand to current customers, assisting them in managing customer perceptions and converting positive emotions into purchasing actions.

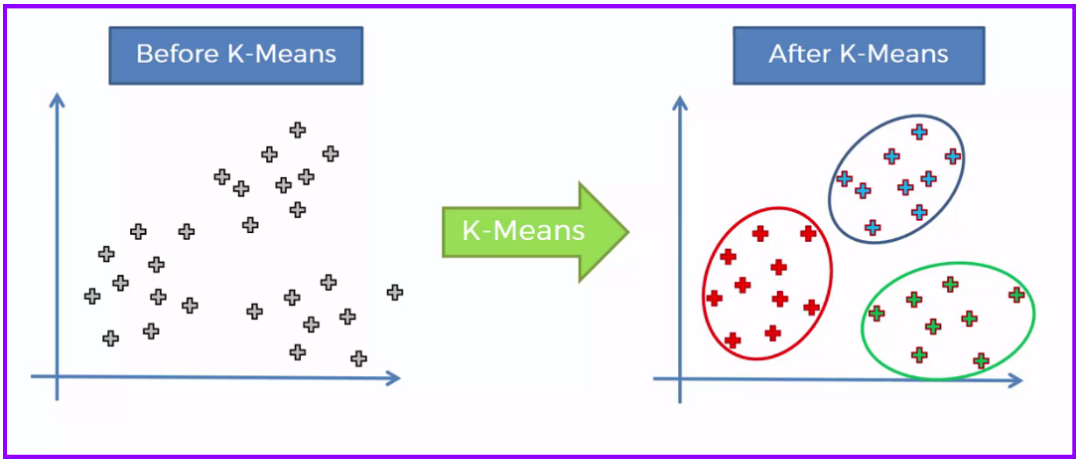
1.5 K-means algorithm

Picture a bustling marketplace, each shopper a data point waiting to be understood. Enter K-means, our guiding light in the realm of machine learning.

K-means begins by randomly assigning shoppers to clusters, laying the groundwork for organization. Iteratively, it calculates centroids, the heart of each cluster, refining them until minimal change is achieved.

This algorithm's simplicity and scalability make it a beacon in data analysis. Yet, its sensitivity to initial clusters and the need to predefine their number pose challenges. Computational costs rise with high-dimensional data.

Understanding K-means' allure and limitations is vital as we navigate our data landscape, uncovering insights to shape our analytical journey.



2.Exploratory Data Analysis

2.1 Navigating Data Exploration: Unveiling Insights

In our data exploration journey, we embark on a meticulous inventory of our dataset. This involves cataloging its origin, size, format, historical records, and variable descriptions. Moving forward, we delve into visual exploration, unveiling intricate patterns and trends to grasp the data's essence.

The dataset at hand, Online Retail II, chronicles **transactions spanning from January 1, 2010, to December 1, 2011, from a UK-based online retail company**. Specializing in gift-ware, the

company caters to a diverse clientele, notably including wholesale buyers.

Key attributes within the dataset include:

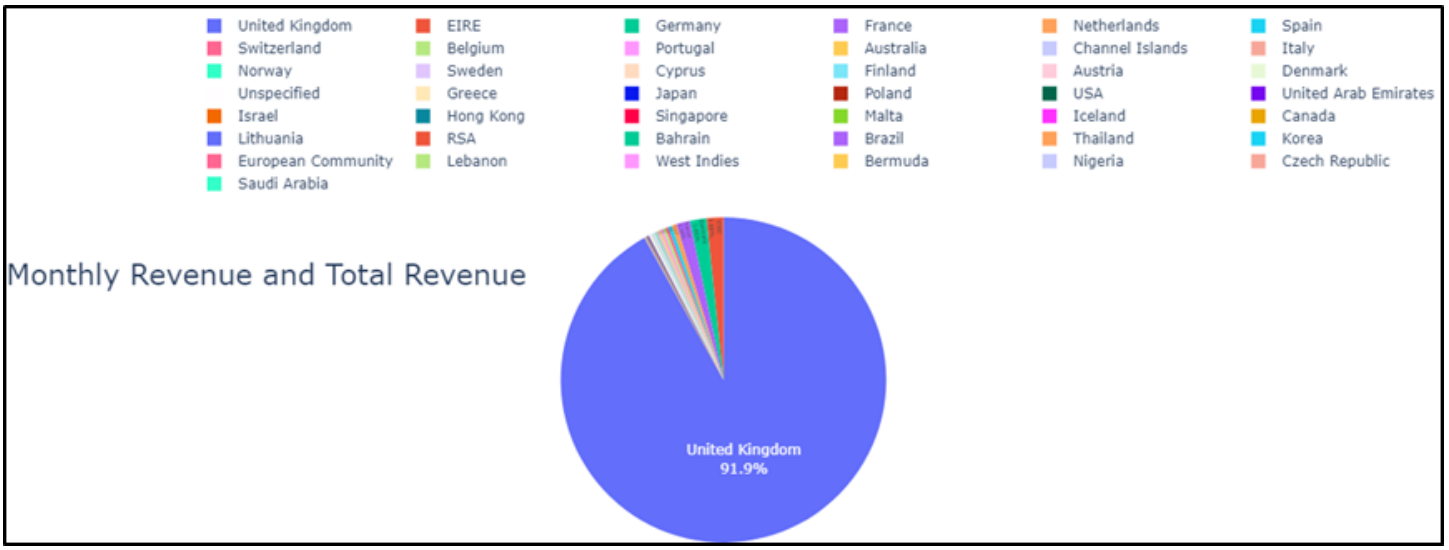
Variable	Description
InvoiceNo	Unique 6-digit integral invoice number
StockCode	A 5-digit integral number is assigned to each distinct product
Description	The name of the product or item being sold.
Quantity	The quantity of each product/item involved in a specific transaction.
InvoiceDate	The precise date and time when a particular transaction was generated.
Price	The price of a single unit of the product in British sterling (£).
CustomerID	A unique 5-digit integral number as each customer identification code.
Country	The name of the country where the customer resides.

2.2 Exploratory Analysis

In this section, we will dive into various aspects of our business, such as the monthly growth rate, the number of monthly active customers, the average monthly quantity of products sold, and more. By analyzing these metrics, we can better understand the factors that drive our business and identify opportunities for growth. Let us embark on this exciting adventure together and unlock the secrets hidden within our data.

2.2.1 Distribution of Sales by Countries

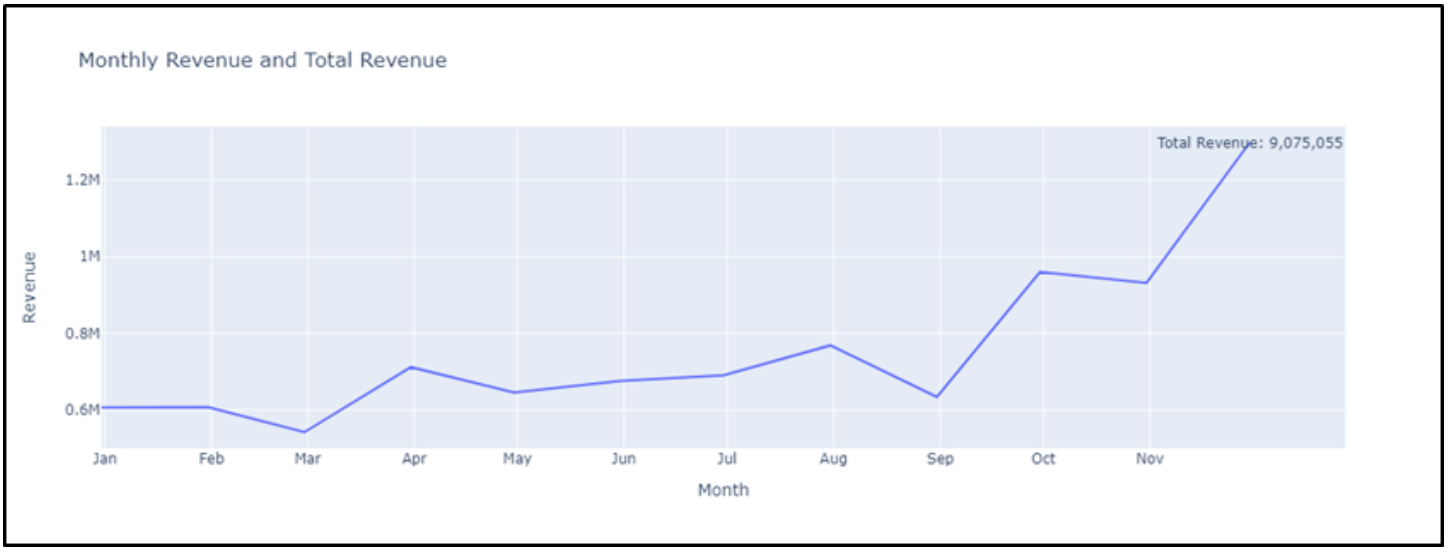
The **UK** dominates with **91.3% of total sales**, warranting focus on UK data for deeper analysis. This targeted approach enhances accuracy and reveals strategic insights for business decisions.



2.2.2 Revenue and monthly growth rate

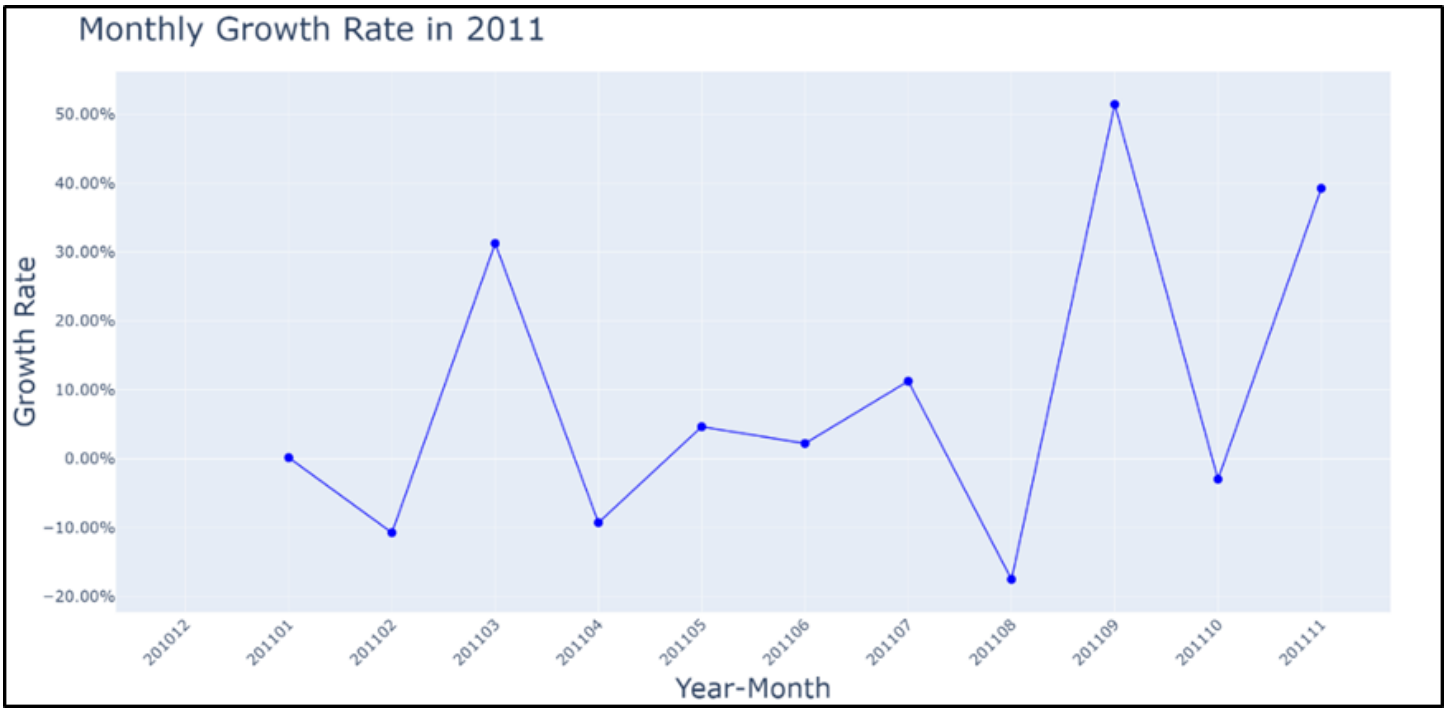
Revenue Analysis: UK Market

Total revenue in the UK from **December 2010 to December 2011** amounted to approximately **£9 million.** Monthly revenue **averaged between £0.6 to £0.8 million**, with a notable surge to £0.9 million between October and December 1st. November saw the **peak revenue of £1.3 million**, suggesting external factors like seasonal sales. Steady revenue growth was evident throughout the year, notably increasing leading up to the holiday season.



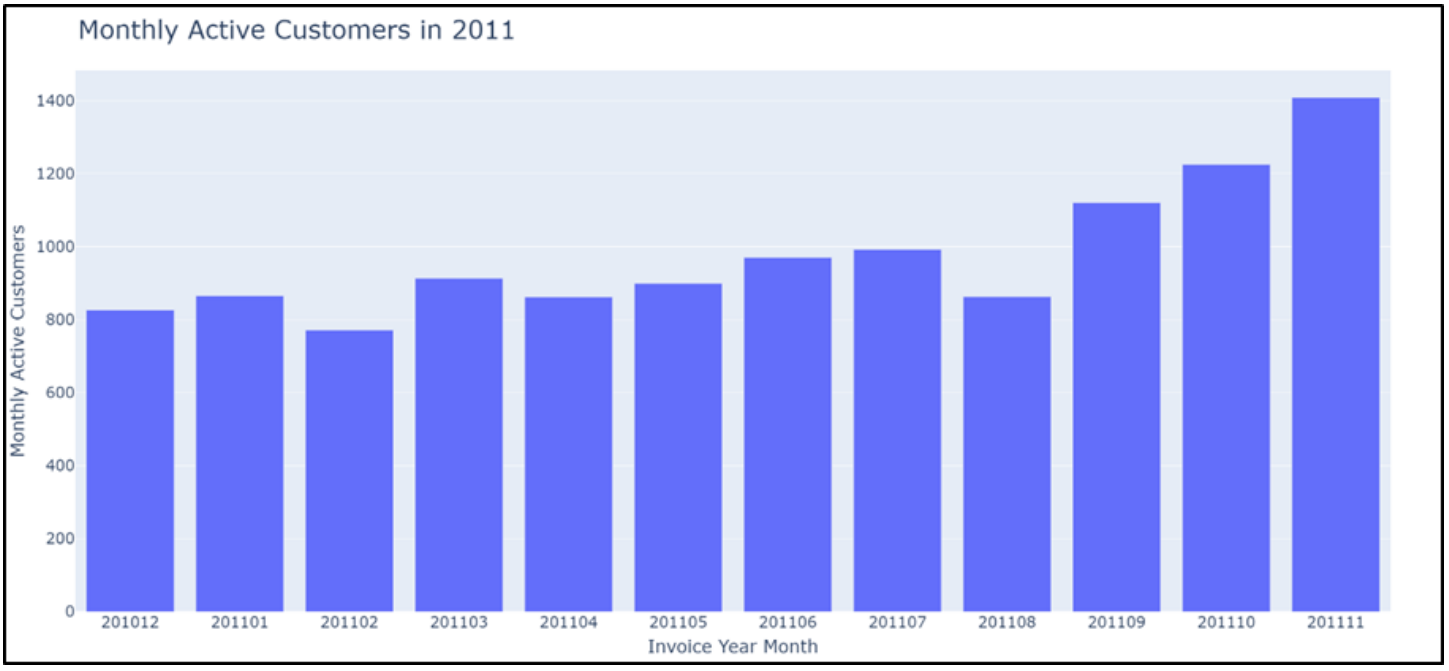
Monthly Growth Rate Insights

Monthly growth rates ranged from **-18% to 50%** over 12 months, with the **lowest in August** and the **highest in September**. Typically, growth rates varied between -10% to 15% in other months. Fluctuations may stem from shifts in consumer preferences, seasonality, economic conditions, or marketing strategies. For instance, September's high growth rate may be attributed to back-to-school shopping, while August's low rate could result from the holiday season or limited promotional activities.



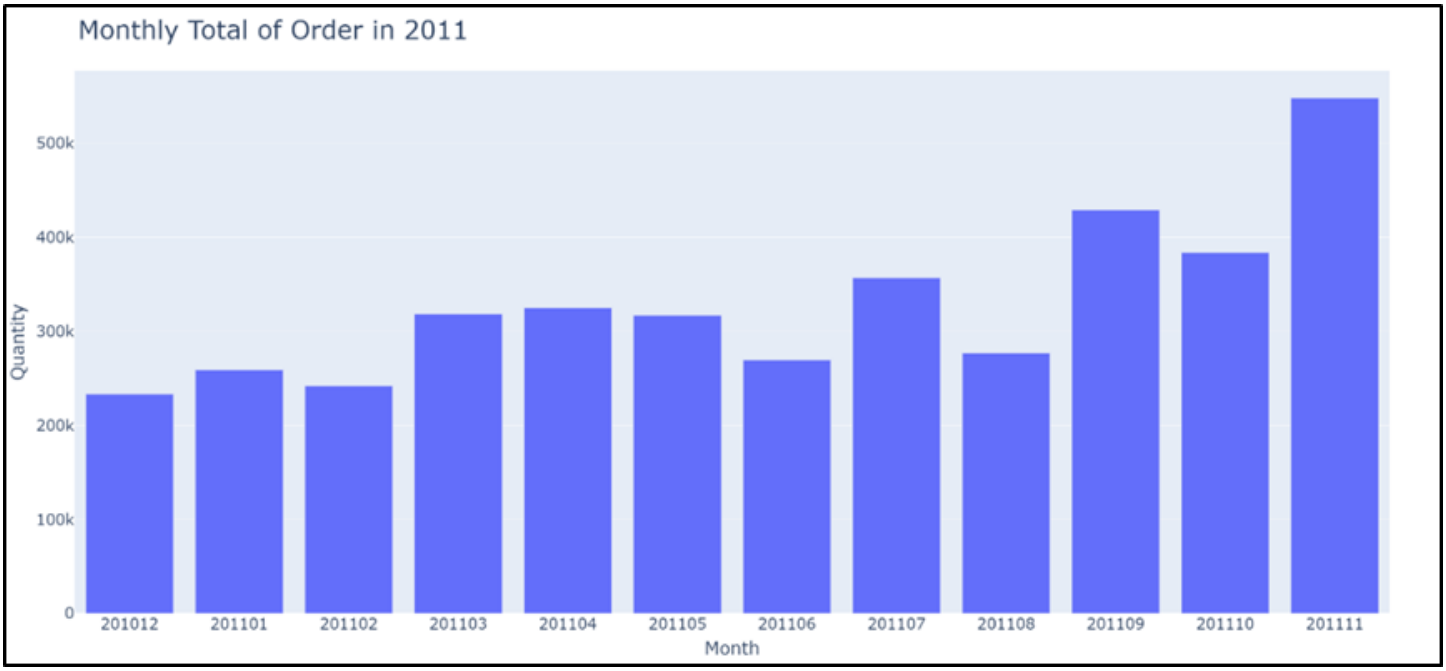
2.2.3 Monthly Active Customers

Our monthly **active customers** in 2011 ranged **from 750 in February to 1415 in November**, averaging 800-1000 monthly. These fluctuations provide valuable insights into customer behavior and the effectiveness of our marketing strategies. The low number of active customers in February may be attributed to seasonal trends, economic conditions, or external factors. Conversely, November's high number of active customers suggests effective holiday season marketing.



2.2.4 Average Quantity of Product sold monthly

In 2011, the business had an average monthly **quantity of products sold** ranging from around **250K to 450K**. The lowest sales occurred in **January at about 230K** and **the highest in November at about 580K**. These sales patterns are closely linked to our monthly active customers and revenue growth rate. The higher number of active customers in November could have contributed to the increased sales during that month. Similarly, the slower growth rate in August may have resulted in lower product sales.



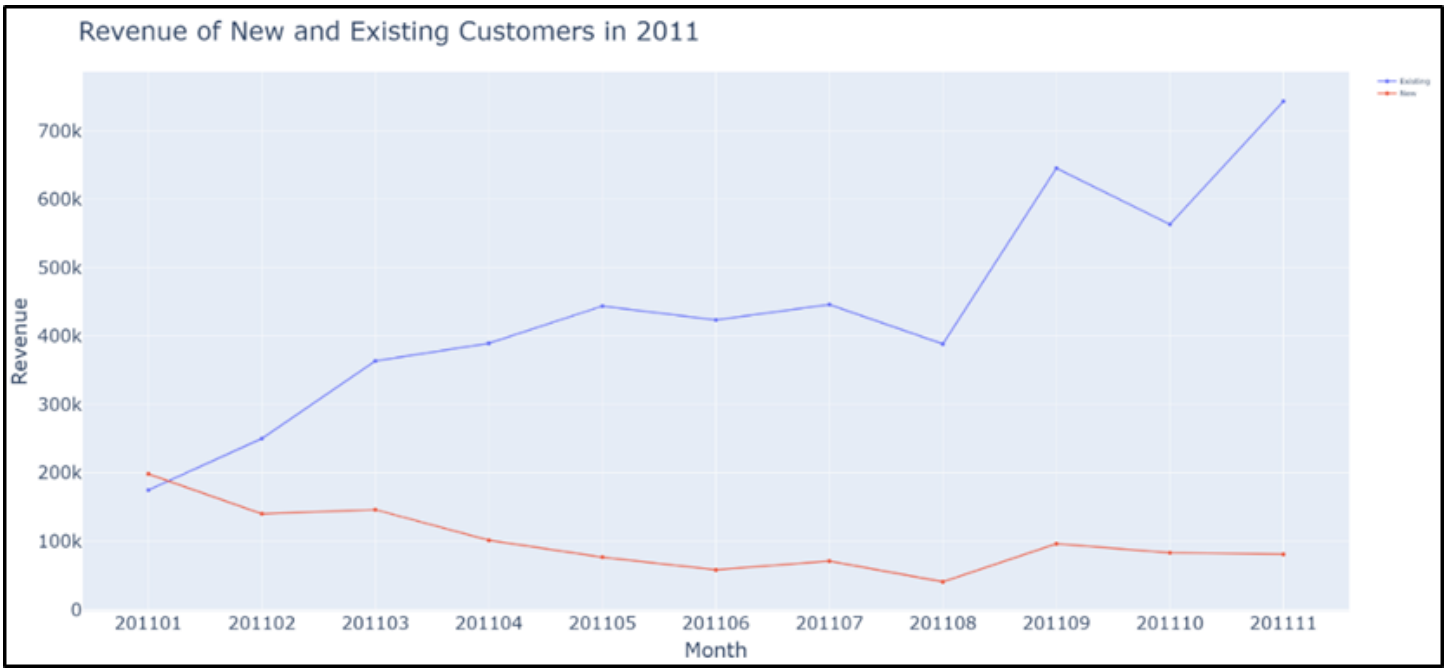
2.2.5 Revenue of New and Existing Customers

Based on the analysis of customer purchase behavior in 2011, it was determined that **95027 new customers** and **243202 returning customers made more than one purchase**. The **revenue** for **the existing customer group** was consistently **higher than that of the new customer group**, ranging from 250k to 750k pounds. Moreover, the revenue for the new customer group ranged from 50k to 125k pounds.

The idea of **customer lifetime value (CLV)** can **explain the disparity in revenue between the two groups**. **CLV is a metric that measures the net present value of the revenue a customer will generate over their entire relationship with a company**. Returning customers have already established a relationship with the company, which makes them more likely to make repeat purchases and spend more money, resulting in a higher CLV. In contrast, new customers have yet to establish trust and familiarity with the company, which makes them more hesitant to spend large amounts of money, resulting in a lower CLV.

Furthermore, the fact that the revenue for the existing customer group increased equally suggests that the company successfully retained its existing customers. This indicates that the company has a solid customer base and can maintain customer loyalty while expanding its customer reach, which is a positive sign for long-term revenue growth.

These findings highlight the importance of building long-term customer relationships to increase CLV and achieve sustained revenue growth. To accomplish this, companies should focus on providing exceptional customer service and products to establish customer trust and loyalty. By doing so, companies can create a loyal customer base that will continue to generate revenue for years.



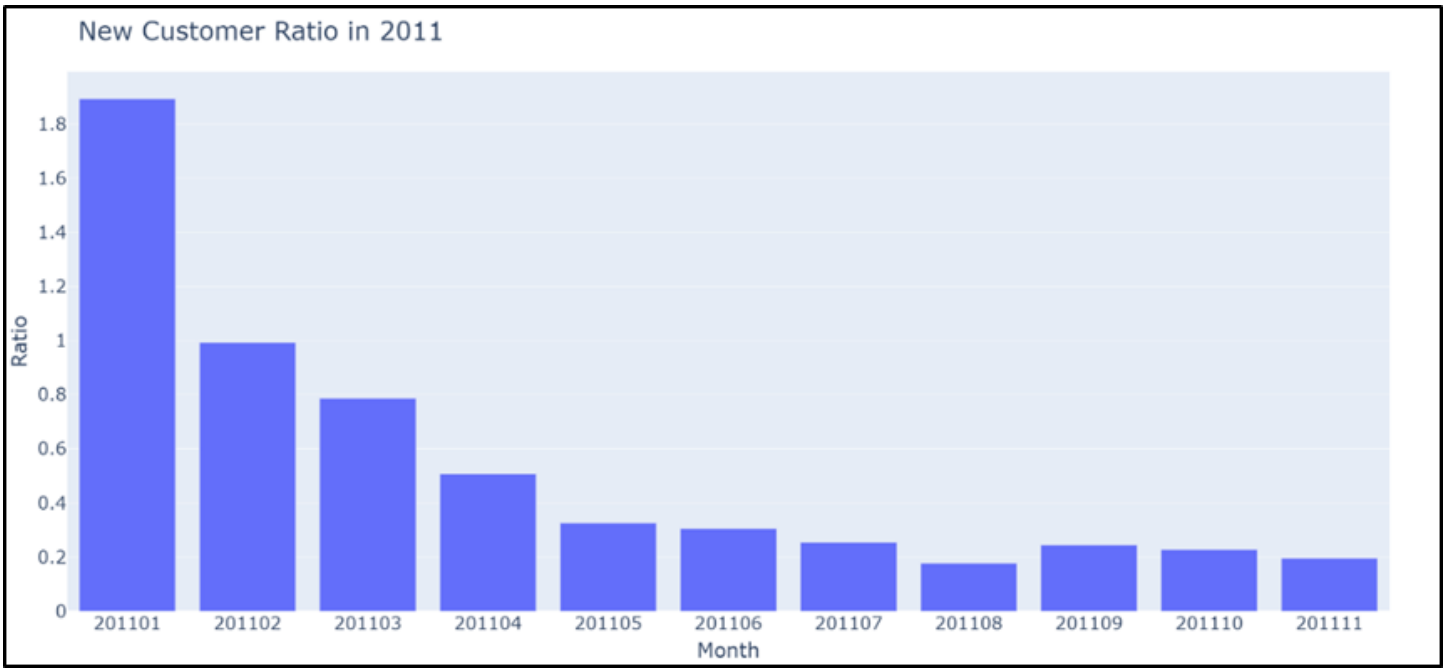
2.2.6 Monthly New Customer Ratio

New Customer Ratio is a metric that helps businesses understand the percentage of new customers compared to existing customers during a specific period. It is calculated by dividing the number of new customers by the number of existing customers.

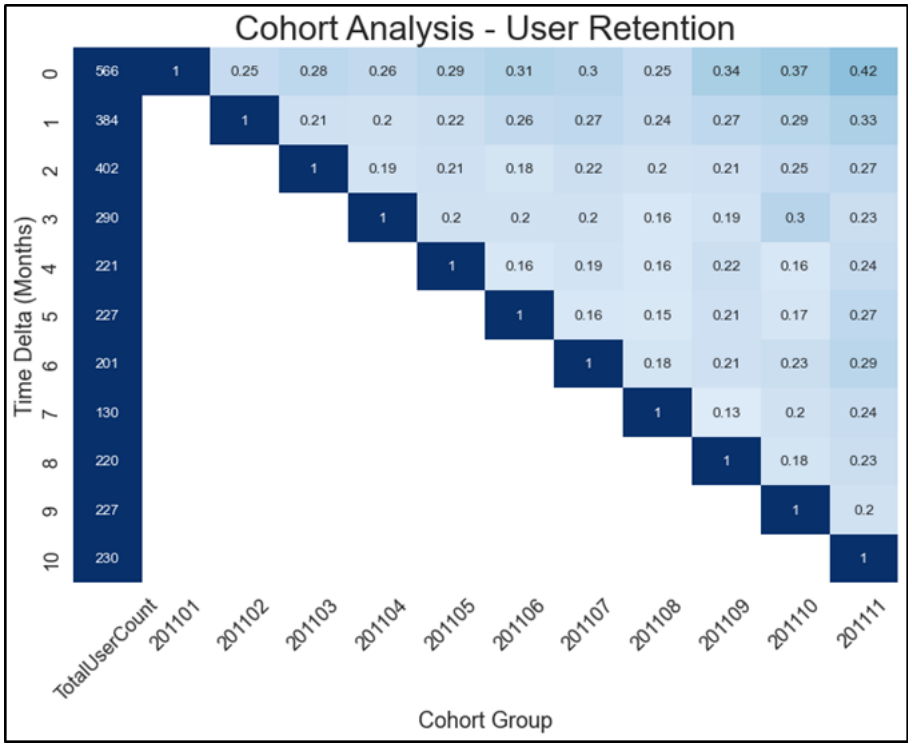
As we journey through the data landscape, let's unravel the narrative painted by the graph before us. In **January 2011, the New Customer Ratio stands tall at 1.8**, signaling a bustling influx of fresh faces into our business realm. It's a moment of promise and opportunity, where the air is filled with the excitement of new beginnings.

By **December 2011, the ratio dwindles to a mere 0.2**, marking a stark contrast to the vibrant dawn of the year. It's a sobering realization, hinting at challenges in retaining our new customers or perhaps a shortfall in acquiring new ones to maintain the balance.

But what lessons can we glean from this narrative? The fluctuations in the New Customer Ratio serve as a barometer for our customer acquisition and retention strategies. A declining ratio prompts introspection, urging us to refine our customer experience and amplify marketing efforts to allure and retain new patrons. Conversely, a rising ratio paints a picture of success, indicating our adeptness in expanding our customer base and fostering a flourishing ecosystem.



2.2.7 Cohort-Based Retention



3.Machine Learning Implementation and Model Evaluation

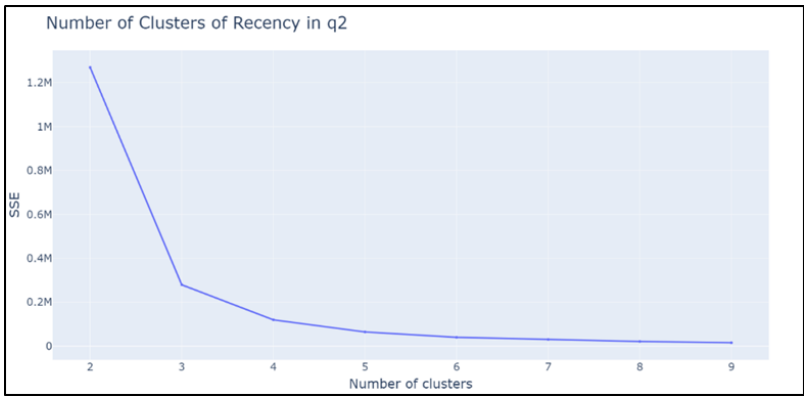
3.1 Clustering Models: CUSTOMER SEGMENTATION by RFM

To apply the RFM method, we must collect data on customer transactions, such as purchase date, purchase frequency, and monetary value. In this case, we will use the **data collected over three months**, from **01/03/2011 to 01/06/2011**, to calculate the RFM scores and

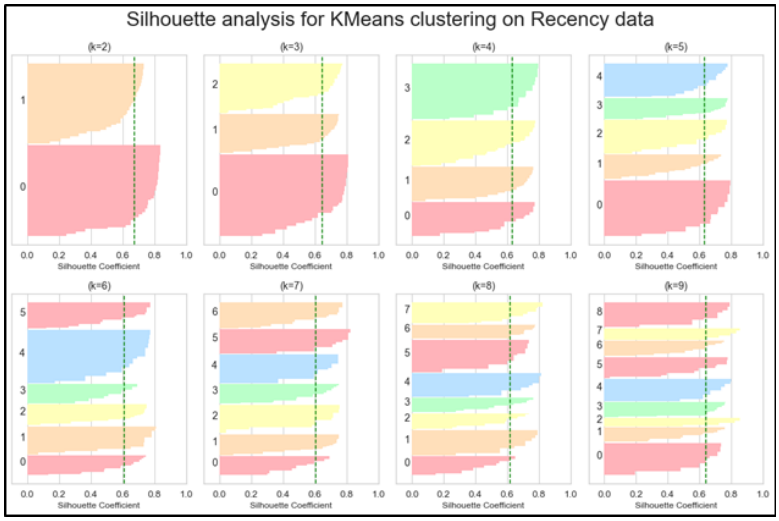
segment customers into groups. Using this data, we can gain insights into customer behavior and develop targeted marketing strategies based on their unique characteristics.

Once the RFM scores are calculated, businesses can cluster customers into groups based on their scores, creating segments such as high-value customers, loyal customers, and dormant customers. This segmentation enables businesses to develop targeted marketing strategies that cater to the specific needs of each segment, thereby increasing the effectiveness of their marketing efforts.

3.1.1 Recency Clustering



Based on the Elbow Curve method, the optimal number of clusters for a clustering model is often determined by identifying the "elbow point" on the WSS plot. In this case, the WSS plot showed a significant decrease in WSS up to k=4, beyond which the decrease became less significant. Therefore, we chose **k=4** as our model's optimal number of clusters. By choosing k=4, we aim to maximize the homogeneity within each cluster and maximize the heterogeneity between clusters, ultimately leading to more accurate and valuable insights from our clustering analysis.

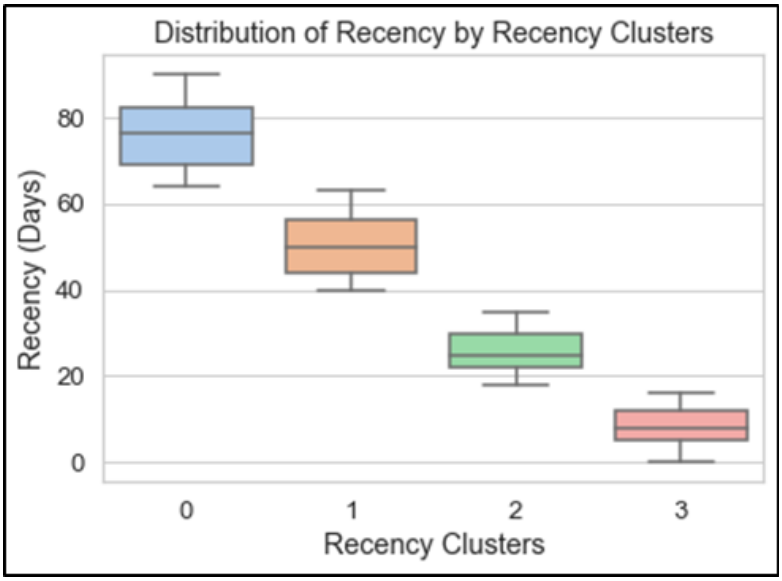


From the Silhouette Plot above, it can be observed that the average silhouette score is highest when k=4. This indicates that the data points are most similar to their assigned cluster

compared to other clusters when k=4. Therefore, using **4 clusters** in our KMeans clustering model is the optimal choice based on the Silhouette Method.

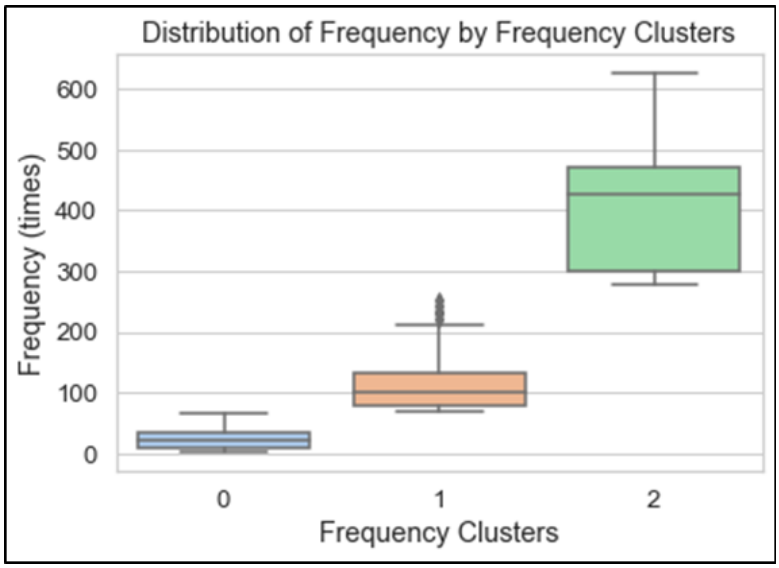
We will arrange the clusters in descending order **based on the mean recency value**, with cluster **0 having the highest and cluster 3 having the lowest**. This means that cluster 0 will contain customers who made their most recent purchase the farthest in the past, while cluster 3 will contain customers who made their most recent purchase the closest to the end of May (May 31, 2011).

Here is the Distribution of Recency by clusters:



3.1.2 Frequency Clustering and Monetary Clustering

Similarly, after splitting the frequency data using K-means into four clusters based on **k = 3** by K-means, we will arrange the clusters based on each cluster's mean of the frequency feature. The frequency feature measures how many times a customer has made a purchase. We will arrange the clusters in descending order **based on the mean frequency value**, with **cluster 2 having the highest and cluster 0 having the lowest**. This means that cluster 2 will contain customers who made the most purchases, with the mean is about 450 times, while cluster 0 will contain customers who made the least(mean =30 times). This ordering will provide valuable insights into customer behavior and preferences based on purchase frequency, allowing businesses to tailor their marketing and retention strategies accordingly.



3.1.3 Monetary Clustering

We thoroughly analysed the Elbow Curve and Silhouette Plot and found that **k=3** is the optimal number of clusters for Monetary data clustering. The Elbow Curve indicates that beyond k=3, adding more clusters does not significantly improve clustering performance. Our analysis shows a sharp decline until k=3, beyond which the curve levels off. Moreover, the Silhouette Plot supports our choice of k=3. This number of clusters has the highest Silhouette score, indicating that data points are well-separated and belong to their respective clusters.

We arranged the clusters in descending order of mean monetary value, with **Cluster 3 representing high-spending customers and Cluster 0 representing low-spending customers.**

3.1.4 Overall Segmentation Clustering

OverallScore = RecencyCluster + FrequencyCluster + RevenueCluster.

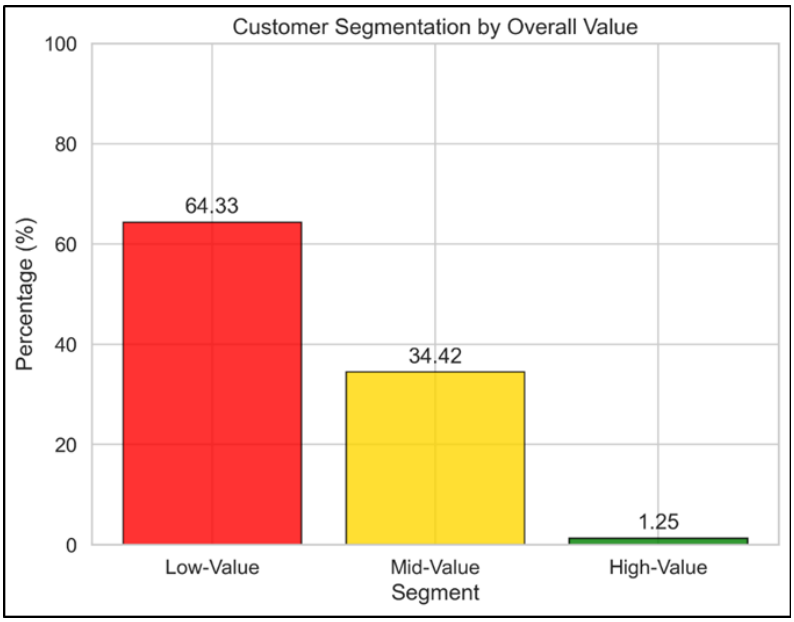
· With RecencyCluster from 0 to 3

· FrequencyCluster from 0 to 2

· RevenueCluster from 0 to 2

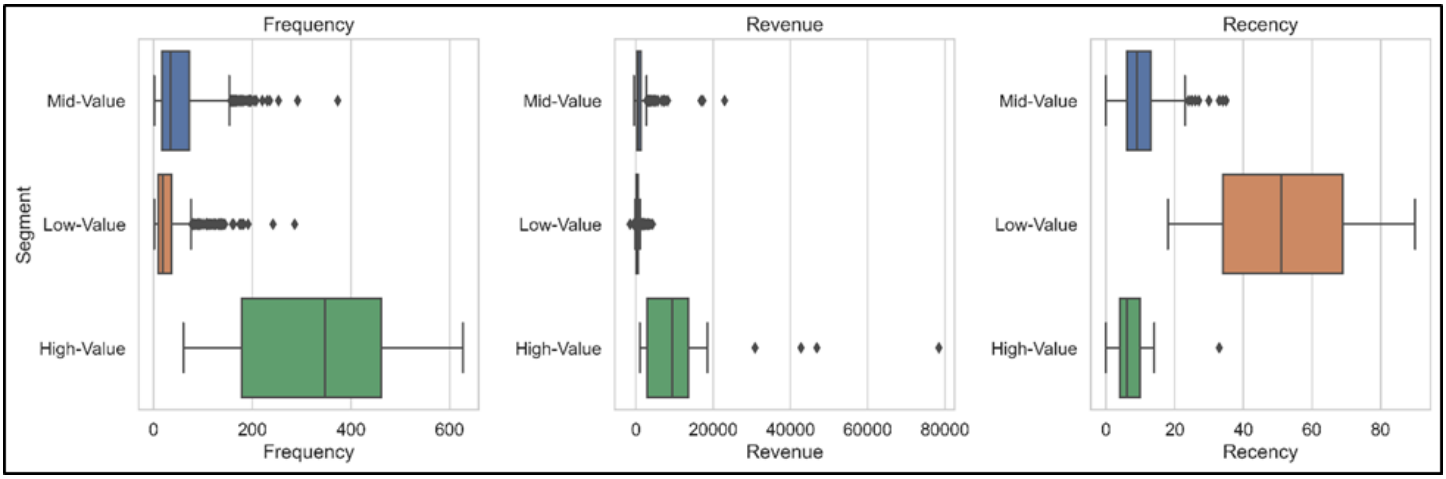
This formula allows businesses to segment their customers based on purchasing behavior, purchase frequency, and monetary value. By clustering customers based on their OverallScore, businesses can better understand their customers' needs and preferences, allowing them to tailor their marketing and retention strategies accordingly.

Here is the Distribution of 3 segments:



In this project, the segmentation of customers based on their OverallScore has been further enhanced by dividing them into three segments: High-value, Mid-value, and Low-value customers. **High-value customers account for only 1.25%** of the customer base, with an **OverallScore of 5-7,** indicating their importance due to frequent and high-value purchases. They represent the most valuable business segment and should be prioritized in marketing and retention efforts. **Mid-value customers account for 34.42%** of the customer base, with an **OverallScore of 3-5**. While they may have made less expensive purchases, they represent a valuable business segment, and retention efforts should be focused on them. **Low-value customers account for the most significant proportion, 64.33%**, with an **OverallScore of 1-3**. Although they may not represent the most valuable segment, they should still be addressed with efforts to encourage additional purchases and retain their loyalty.

Let's scrutinize the interplay of the Recency, Frequency, and Monetary features within the three customer segments that were performed:



Examination of the data reveals a marked disparity in the Frequency metric for the High-Value group compared to the remaining two groups. It is, therefore, advisable for enterprises to direct their attention towards customers who exhibit a Frequency greater than 200.

In contrast, the Low and Mid groups exhibit only marginal discrepancies in their Frequency and Monetary values. Instead, the salient factor that sets these two groups apart is Recency, where the Mid-Value group typically demonstrates a Recency index within the range of 5 to 15 days, while the Low-Value group tends to exhibit a Recency index ranging between 35 to 70 days.

3.2 Classification Models for Customer Lifetime Value

3.2.1 Data Preparation

To accomplish this part, we selected **three modelling nodes, including XGBoost model, Random Forest, and Support Vector Machine**, to fit their parameters and build the respective models. Using a dataset of customer purchase behavior from **March 1st, 2011, to June 1st, 2011, we predict the customer's lifetime value (LTV) for the following six months (June 1st, 2011, to December 1st, 2011).**

Customer Lifetime Value (CLV) is calculated from the revenue generated by customers over a certain period. To better understand and predict CLV, we first segment customers into different groups based on their revenue. In this project, we will split CLV into three segments: High-value, Mid-value, and Low-value. We will use K-means clustering, an unsupervised machine learning algorithm, to group customers based on their revenue. This will help us identify patterns and trends in the data and make more accurate predictions.

Here is he dataset after using Kmeans for LTVCluster

	CustomerID	Frequency	Revenue	Recency	RecencyCluster	FrequencyCluster	RevenueCluster	OverallScore	Segment	q34_Revenue	LTVCluster
0	17961.0	58	228.29	4	3	0	0	3	Mid-Value	413.42	0
1	14867.0	16	313.40	26	2	0	0	2	Low-Value	396.88	0
2	16841.0	17	308.76	81	0	0	0	0	Low-Value	1148.48	0
3	14239.0	22	297.26	8	3	0	0	3	Mid-Value	960.56	0
4	15299.0	5	2065.99	47	1	0	0	1	Low-Value	-113.91	0

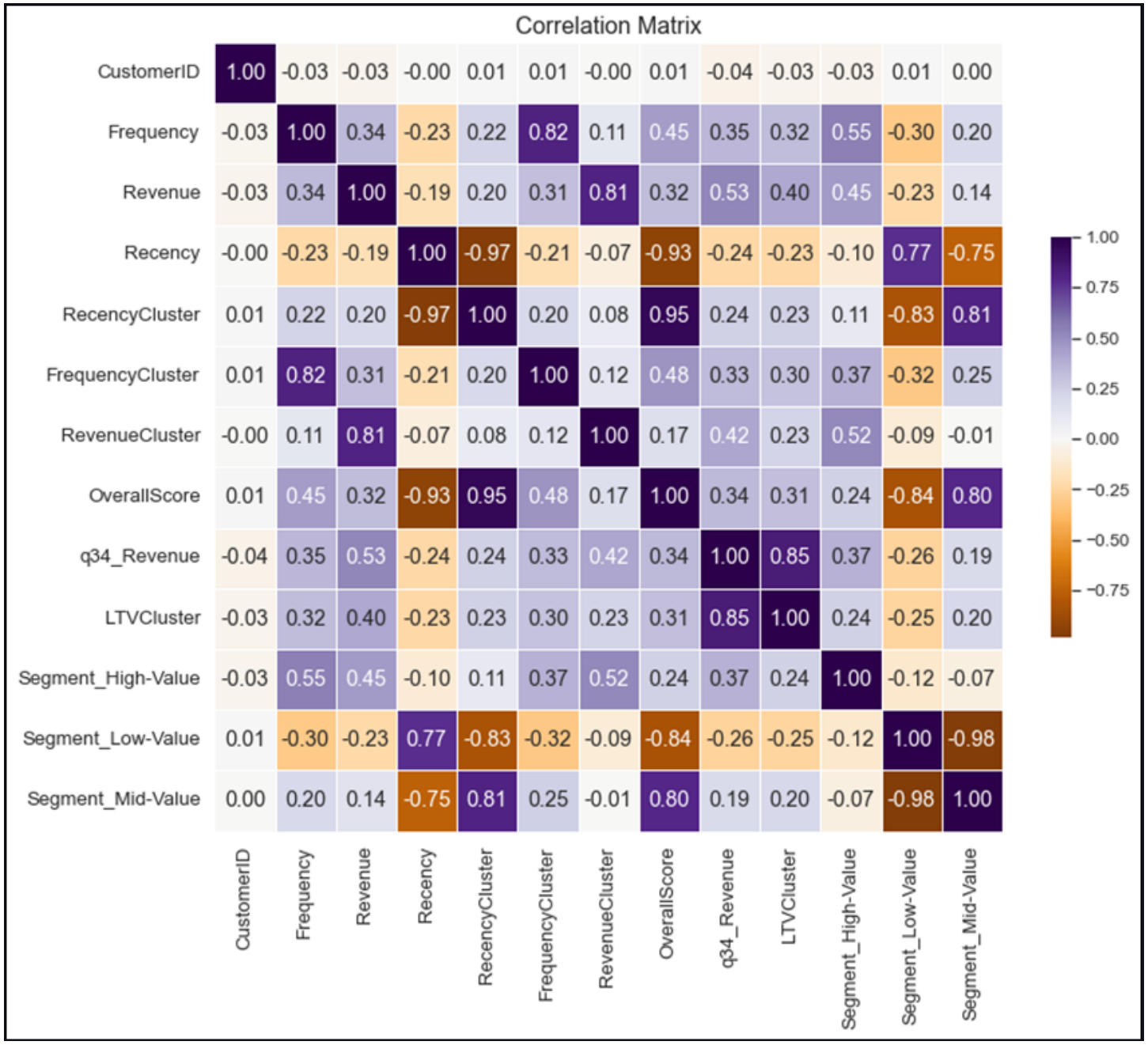
The data contains categorical variables, which need to be converted into numerical features for machine learning models. We will use the pandas library in Python to create dummy variables for categorical variables. Dummy variables are a common way to convert categorical variables into numerical features for machine learning models.

The dataset after using **Get Dummies** function:

	CustomerID	Frequency	Revenue	Recency	RecencyCluster	FrequencyCluster	RevenueCluster	OverallScore	q34_Revenue	LTVCluster	Segment_High-Value	Segment_Low-Value	Segment_Mid-Value
0	17961.0	58	228.29	4	3	0	0	3	413.42	0	0	0	1
1	14867.0	16	313.40	26	2	0	0	2	396.88	0	0	1	0
2	16841.0	17	308.76	81	0	0	0	0	1148.48	0	0	1	0
3	14239.0	22	297.26	8	3	0	0	3	960.56	0	0	0	1
4	15299.0	5	2065.99	47	1	0	0	1	-113.91	0	0	1	0

Before building our models, it's important to check the correlation between features in the data. Highly correlated features, such as overfitting, can cause problems in machine learning

models.



Based on the correlation matrix, we can see that LTVCluster has a **strong positive correlation with q34_Revenue (0.85)**, followed by Revenue (0.40), Frequency (0.32), and OverallScore (0.31). These results suggest that the customer's LTV is positively related to the total amount spent (q34_Revenue), as well as other indicators of customer value such as Revenue, Frequency, and OverallScore. Additionally, we can see that there is a negative correlation between LTVCluster and Recency (-0.23), indicating that customers who made purchases more recently have lower LTVs.

However, we also see that q34_Revenue has a very high correlation with LTVCluster, which can lead to multicollinearity issues when building a model to predict LTVCluster. Therefore, it is a good practice to **remove q34_Revenue from the model** and only include variables that are not highly correlated with each other.

After dropping the q34_Revenue and LTV Cluster:

X data set:

	CustomerID	Frequency	Revenue	Recency	RecencyCluster	FrequencyCluster	RevenueCluster	OverallScore	Segment_High-Value	Segment_Low-Value	Segment_Mid-Value
0	17961.0	58	228.29	4	3	0	0	3	0	0	1
1	14867.0	16	313.40	26	2	0	0	2	0	1	0
2	16841.0	17	308.76	81	0	0	0	0	0	1	0
3	14239.0	22	297.26	8	3	0	0	3	0	0	1
4	15299.0	5	2065.99	47	1	0	0	1	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...
1815	17084.0	162	2949.75	50	1	1	0	2	0	1	0
1816	14354.0	3	75.84	49	1	0	0	1	0	1	0
1817	13596.0	18	285.44	49	1	0	0	1	0	1	0
1818	13153.0	15	486.22	49	1	0	0	1	0	1	0
1819	15773.0	10	635.68	49	1	0	0	1	0	1	0

1820 rows × 11 columns

y dataset: LTVCluster

In order to assess the efficacy of our machine learning models, it is necessary to partition the data into distinct subsets for training, validation, and testing purposes. The data will be partitioned into three sets, with **80% allocated for training and 10% each for validation and testing purposes**. The data will be split using the train_test_split function provided by scikit-learn. This step helps us build a robust and accurate model by preventing overfitting or underfitting. **The test size for the train-test split is set to 0.1, and the test size for the validation set is set to 0.125.**

3.2.1 Result for Customer Lifetime Value Classification model :

- XGBoost Model after GridSearchCV:

```
Classification Report of XGBOOST after tuning for Test Data:
              precision    recall  f1-score   support

           0       0.88      1.00      0.94       153
           1       0.71      0.19      0.30        26
           2       1.00      0.33      0.50         3

    accuracy                           0.87       182
   macro avg       0.86      0.51      0.58       182
weighted avg       0.86      0.87      0.84       182
```

- Support Vector Machine after GridSearchCV:

```
Classification Report of SVM after-tunning parameter:
              precision    recall  f1-score   support

           0       0.87      0.99      0.93       153
           1       0.50      0.12      0.19        26
           2       1.00      0.33      0.50         3

    accuracy                           0.86       182
   macro avg       0.79      0.48      0.54       182
weighted avg       0.82      0.86      0.81       182
```

- Random Forest after GridSearchCV:

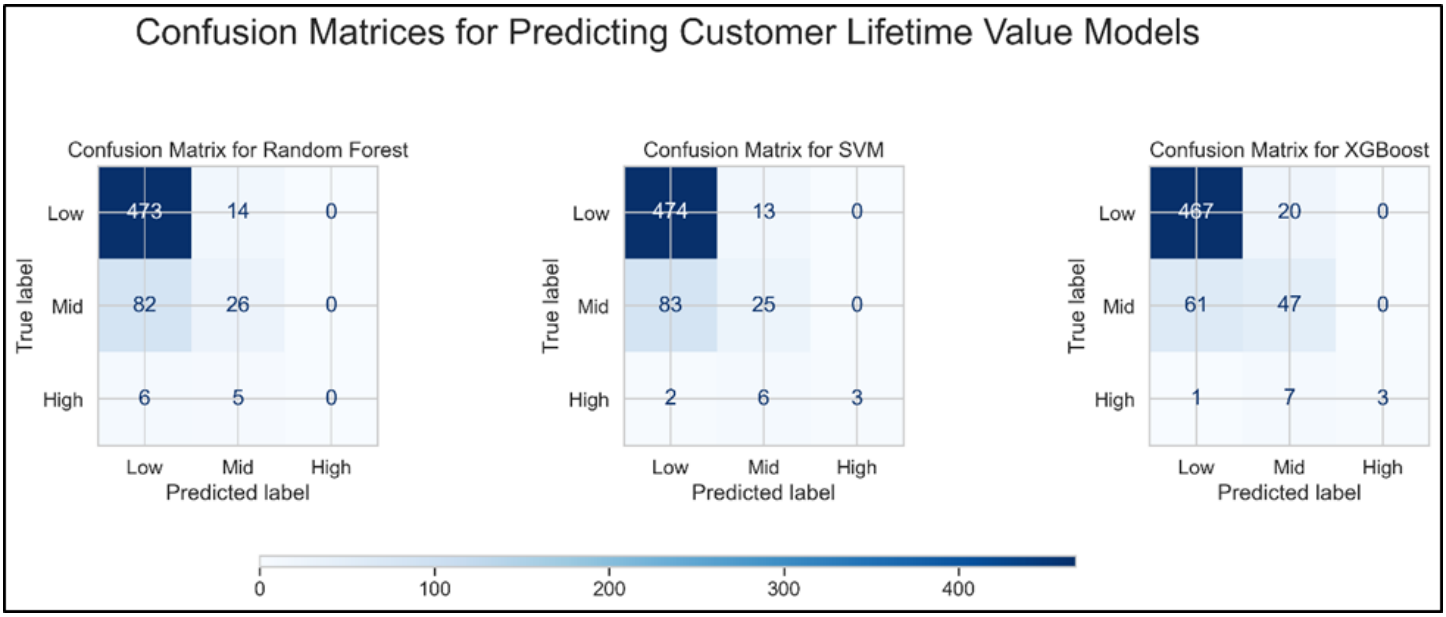
Classification Report of Random Forest after tunning parameter:

	precision	recall	f1-score	support
0	0.87	1.00	0.93	153
1	0.83	0.19	0.31	26
2	1.00	0.33	0.50	3
accuracy			0.87	182
macro avg	0.90	0.51	0.58	182
weighted avg	0.87	0.87	0.84	182

- Customer Lifetime Value Models Assessment:

	Accuracy	Precision	Recall	F1
XGBoost	0.853	0.841	0.853	0.839
Support Vector Machine	0.828	0.801	0.828	0.795
Random Forest	0.823	0.781	0.823	0.786

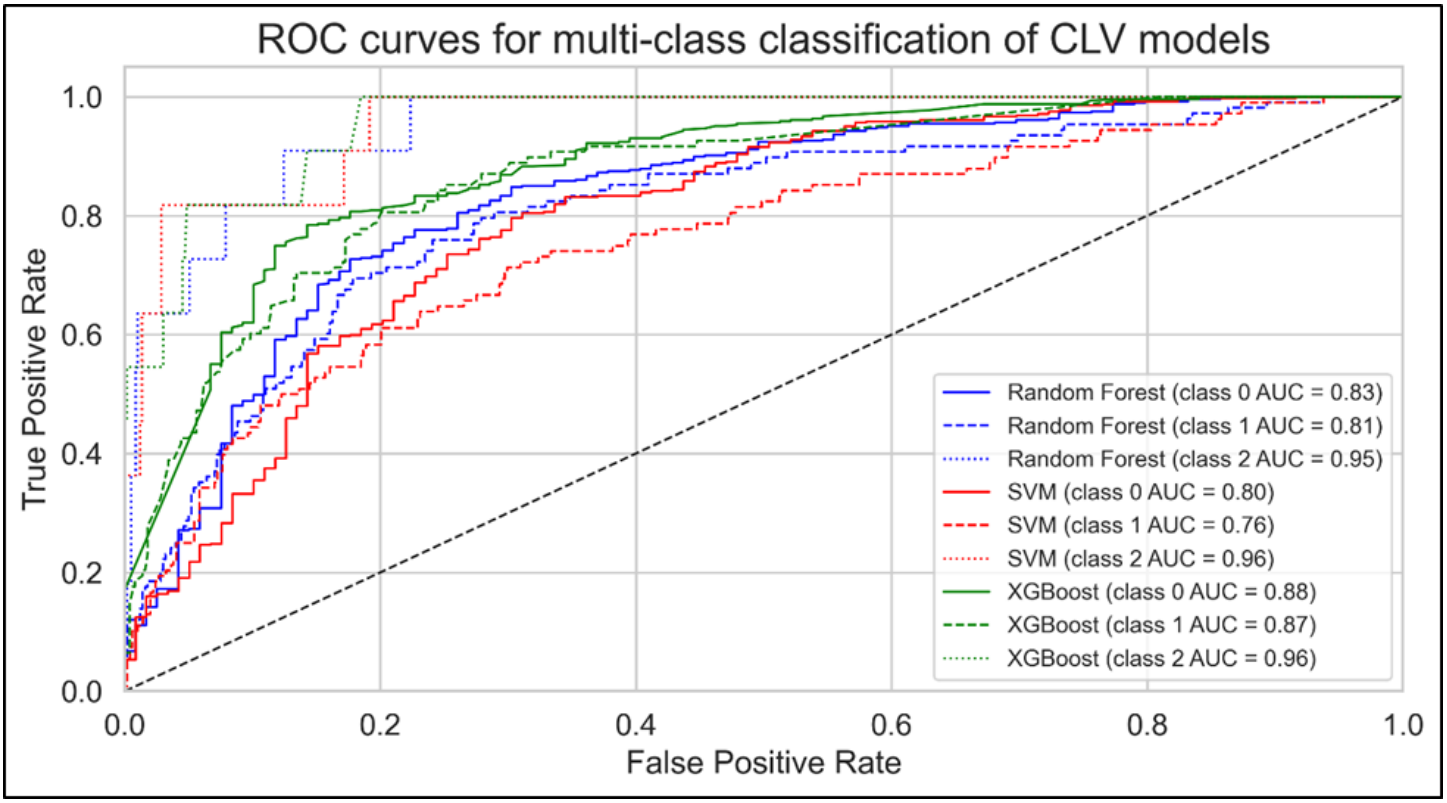
- Confusion Matrices



Based on the confusion matrices, it is apparent that all three models perform reasonably well in predicting the Low Class (Low Customer Lifetime Value), with the Support Vector Machine model demonstrating the best performance in this category, but similar to others. However, all three models encounter difficulties in predicting the Mid Class and High Class, with XGBoost exhibiting the highest accuracy. Given that these models enable businesses to predict different customer segments accurately, we prioritize the accuracy metric of the model. In summary, after evaluating the four assessment scores and confusion matrices,

XGBoost appears to be the best model. Nevertheless, ROC and AUC should also be considered to ensure their reliability.

- ROC and AUC



The ROC curve and AUC scores indicate that XGBoost model has the highest AUC score for all three classes (the closer to 1, the better). When combined with the evaluation results from other methods, it can be concluded that the business should select the XGBoost model for predicting Customer Lifetime Value. **With an accuracy score of 85.3%, precision score of 84.1%, recall score of 85.3%, and F1 score of 83.9%, XGBoost model has the best overall performance. The AUC scores provide additional support for this conclusion.**

3.3 Prediction models of Customer Next Purchase Period

3.3.1 Data Preparation

Now we need to add more features to the model of predicting **The Next Period Purchase**

	CustomerID	Recency	RecencyCluster	Frequency	FrequencyCluster	Revenue	RevenueCluster	OverallScore	DayDiff	DayDiff2	DayDiff3	DayDiffMean	DayDiffStd	Segment_High-Value	Segment_Low-Value	Segment_Mid-Value	NextPurchaseDayRange
0	12346.0	133	2	35	0	-146.73	0	2	202.0	204.0	283.0	73.000000	92.761343	0	1	0	0
1	13821.0	69	2	188	0	1150.59	0	2	102.0	132.0	265.0	89.000000	43.166345	0	1	0	1
2	16792.0	117	2	102	0	623.46	0	2	74.0	279.0	308.0	102.666667	91.434858	0	1	0	0
3	14477.0	105	2	6	0	821.10	0	2	139.0	207.0	231.0	80.000000	47.756326	0	1	0	1
4	15712.0	95	2	122	0	2439.40	0	2	92.0	199.0	238.0	41.571429	41.355601	0	1	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1693	15311.0	4	3	2494	2	56756.97	1	6	2.0	4.0	9.0	3.982301	3.499955	1	0	0	2
1694	13089.0	5	3	2222	2	74257.11	1	6	2.0	3.0	10.0	4.520408	3.653073	1	0	0	2
1695	14298.0	33	3	2101	2	67309.32	1	6	1.0	7.0	9.0	13.580645	11.339530	1	0	0	1
1696	15039.0	6	3	1990	2	29264.93	1	6	9.0	15.0	29.0	8.296296	4.752873	1	0	0	2
1697	18102.0	14	3	739	1	371466.82	2	6	1.0	9.0	27.0	9.239130	11.407960	1	0	0	2

1698 rows × 17 columns

Here is the training dataset:

The training dataset is filtered from **“2010-03-01” to “2011-06-01” (15 months), used to predict the test dataset from “2011-06-01” to “2011-12-01” (the next 6 months).**

This project uses the RFM (Recency, Frequency, Monetary) method to segment our customers and predict their Next Purchase Day Range. We **added five new features to our model: DayDiff, DayDiff2, DayDiff3, DayDiffMean and DayDiffStd**, representing **the number of days between a customer's current purchase and their previous three, the mean and standard deviation of them for each customer**, which will be used as additional features in our model.

After adding these features, we converted our categorical variables into binary/dummy variables and assigned values to the 'NextPurchaseDayRange' column for all rows in our dataset. We then split the 'NextPurchaseDayRange' into three categories: 0, 1, or 2. The following scale is used to determine the likelihood of customer purchases within specific timeframes:

· **A value of 0 signifies customers who are unlikely to make a purchase within the next 180 days.**

· **A value of 1 indicates customers who are likely to make a purchase between 31 and 180 days.**

· **A value of 2 corresponds to customers who are likely to make a purchase within the next 30 days.**

We separated the features (X) and target variable (y) from our dataset to train our classification model. The features are all columns except for 'NextPurchaseDayRange', the target variable. We can then use various machine learning algorithms to train our model and make predictions on new data. By predicting a customer's Next Purchase Day Range, we can better understand their behaviour and tailor our marketing efforts to improve customer retention and increase revenue.

3.3.2 Result for Customer Next Purchase Period models

- XGBoost Model after GridSearchCV:

Classification Report of XGBOOST after tuning for Test Data:

	precision	recall	f1-score	support
0	0.98	0.67	0.79	63
1	0.71	0.98	0.82	86
2	0.89	0.38	0.53	21
accuracy			0.79	170
macro avg	0.86	0.67	0.72	170
weighted avg	0.83	0.79	0.78	170

- Support Vector Machine after GridSearchCV:

Classification Report of SVM after-tunning parameter:

	precision	recall	f1-score	support
0	0.83	0.71	0.77	63
1	0.70	0.86	0.77	86
2	0.70	0.33	0.45	21
accuracy			0.74	170
macro avg	0.74	0.64	0.66	170
weighted avg	0.75	0.74	0.73	170

- Random Forest after GridSearchCV:

Classification Report of Random Forest after tunning parameter:

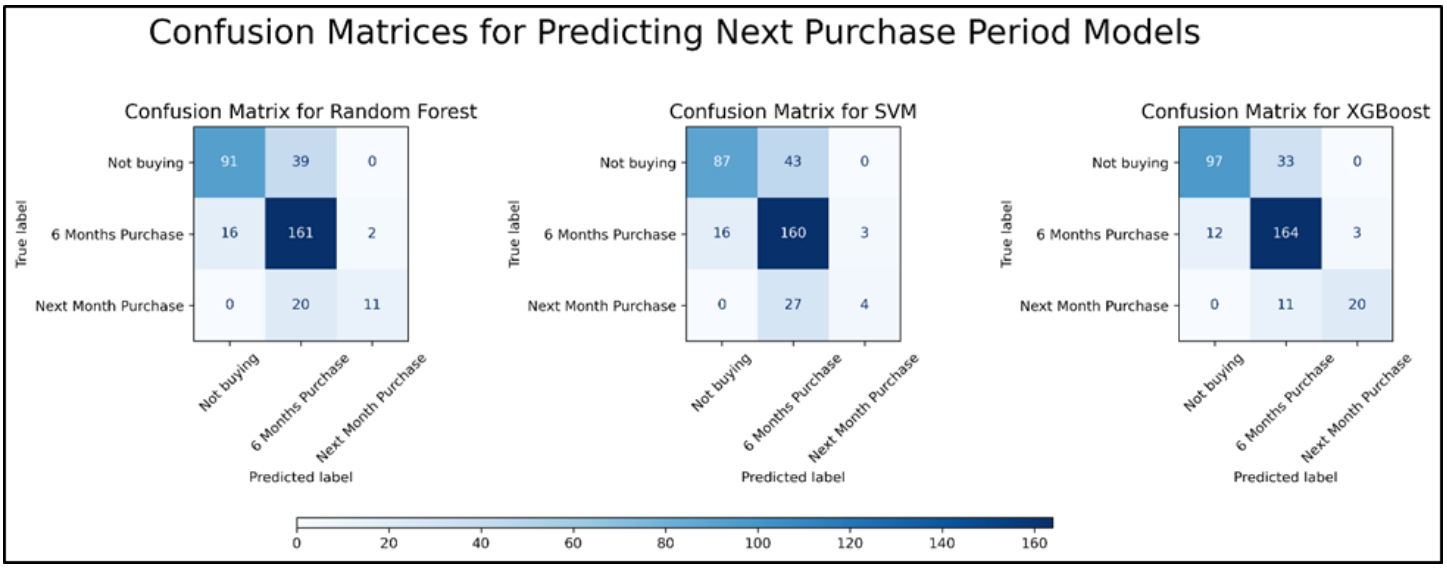
	precision	recall	f1-score	support
0	0.89	0.65	0.75	63
1	0.69	0.91	0.78	86
2	0.73	0.38	0.50	21
accuracy			0.75	170
macro avg	0.77	0.65	0.68	170
weighted avg	0.77	0.75	0.74	170

- Predict Next Day Purchase Models Assessment:

	Accuracy	**Precision**	**Recall**	**F1**
XGBoost	0.788	0.797	0.788	0.784
SupportVector Machine	0.738	0.741	0.738	0.717
Random Forest	0.774	0.788	0.774	0.764

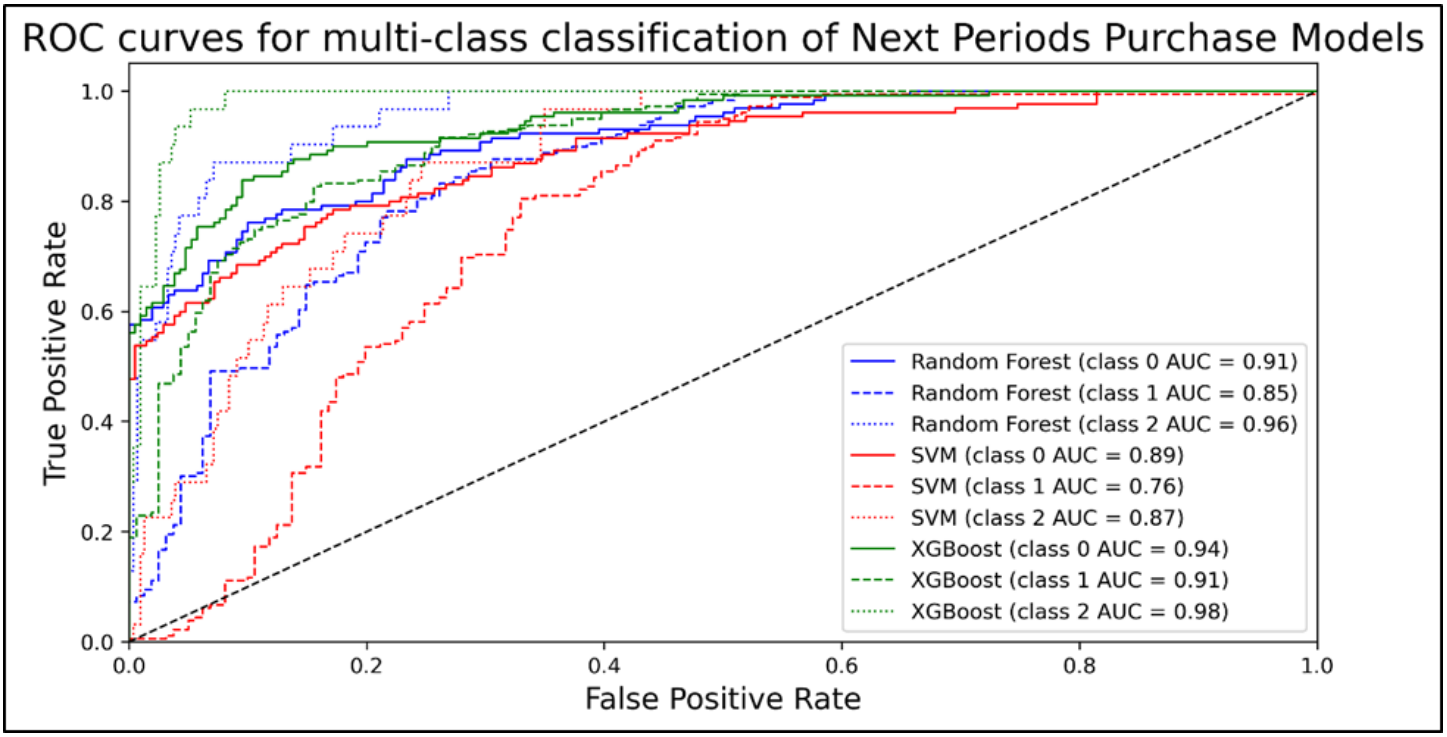
Similarly to evaluating the Customer Lifetime Value models, it can be observed that **XGBoost consistently has the highest performance** across all three models when predicting the next purchase period. The scores nearing 0.8 indicate that this is a reliable model. For a detailed analysis, refer to the confusion matrices below.

- Confusion Matrices



It is evident that all three models predict with high accuracy (>=90%) whether a customer will make a purchase within the next six months. The models also accurately predict the majority of customers who have no intention of purchasing in the next six months, with approximately one-third of this group being wrongly predicted as likely to make a purchase. However, there is clarity between these wrongly predicted customers and those expected to purchase within the next month. For the "Next Month Purchase" class, **XGBoost proves to be the best prediction model**, as it can accurately predict about two-thirds of customers who are certain to make a purchase in the next month, with all remaining incorrect predictions being predicted as making a purchase within the next six months. Since the business is more concerned with accurately predicting customers who will make a purchase in the next one and six months rather than predicting non-purchasing customers in order to offer appropriate stimulus packages, XGBoost is the model of choice.

- ROC and AUC



The ROC curve and AUC scores indicate that XGBoost model has the highest AUC score for all three classes (the closer to 1, the better). When combined with the evaluation results from other methods, it can be concluded that the business should select the **XGBoost model for predicting The Next Purchase Period. With an accuracy score of 78.7%, precision score of 79.7%, recall score of 78.8%, and F1 score of 78.4%, XGBoost model has the best overall performance. The AUC scores provide additional support for this conclusion.**

4.Findings

The data discoveries and models presented above provide several important insights that can support businesses in their decision-making process. Firstly, there is a tendency for an increase in customer volume during the year-end period. Therefore, businesses should prepare appropriate communication and promotional programs starting from August to December. Additionally, businesses should pay attention to the large customer base in the UK market.

By dividing **customers into three segments - high-value customers (7.5%), medium-value customers (31.05%), and low-value customers (61.45%**) - businesses can provide suitable incentives and allocate corresponding resources to each customer group. For instance, customers with the highest recency tend to have medium value rather than belonging to the other two groups.

The **XGBoost Model includes a customer lifetime value prediction model that accurately predicted 87% of cases**, demonstrating its reliability for businesses to effectively care for both existing and new customers. Furthermore, **the next purchase period prediction model also performed well, with an accuracy score of 78.7%, precision score of 79.7%, recall score of 78.8%, and F1 score of 78.4%. The XGBoost model exhibited the best overall performance.**

Based on these findings, businesses can promptly develop solutions for customers with **high potential for purchasing within the next 21 days or the next 6 months while saving costs by reducing efforts towards customers with low potential for purchasing in the coming months.** By combining all three methods, this research provides a comprehensive overview and facilitates timely decision-making for businesses.

5.Appendix

PLEASE CLICK ON THE LINK HERE TO READ FULL CODE.

THANK YOU !

Contact Details

Viet Tuan Dinh

Please do not hesitate to contact me!

SE1 LONDON, UK

PORTFOLIO

+44 8498123569