REPORT SUMMER PROJECT 2023

"Large Scale Predictive Market Segmentation using Machine Learning & Data Analysis"

Submitted By –

Divyanshu Singh Bisht

PRN: 22070243018

Under The Supervision Of –

Dr. Vidya Patkar

Submitted to –

Symbiosis Institute of Geoinformatics

Symbiosis International (Deemed University)

5th Floor, Atur Centre, Gokhale Cross Road,

Model Colony, Pune – 411016

CONTENTS

ACKNOWLEDGEMENT

I want to extend my heartfelt appreciation to Dr. Vidya Patkar, Deputy Director of Symbiosis Institute of Geoinformatics, for her invaluable support and guidance throughout the completion of my project titled “Large Scale Predictive Market Segmentation using Machine Learning & Data Analysis”. Your expertise and encouragement have been instrumental in my success, and I am genuinely grateful for the opportunities you have provided me during this project.

I would also like to express my deepest gratitude to all the Symbiosis Institute of Geoinformatics faculty members for their continuous support and encouragement, especially Dr. Vidya Patkar, for their unwavering dedication and mentorship throughout the project. Their guidance, valuable insights, and constructive feedback have been immensely beneficial, and I cannot thank them enough for their time and efforts. Their expertise and encouragement have shaped me professionally and personally, and I am forever grateful for their mentorship.

I want to emphasize that this project was completed solely by me, and I take full responsibility for its content and execution. I have put in my best efforts to ensure its accuracy and quality.

Once again, I extend my deepest gratitude to all those mentioned above for their support, guidance, and contributions to the successful completion of my project.

Sincerely,

Divyanshu Singh Bisht

List of Figures

List of Tables

7	BIC and AIC scores for 2-8 clusters
8	Average silhouette score for 3,4 and 5 cluster solution
9	Comparison of supervised learners

Abbreviation list

Abbreviation	Stands For
RFM	Recency, Frequency, and Monetary Value
POS	Point Of Sale
MBA	Market Basket Analysis
EM	Expectation Maximisation
PCA	Principal Component Axis
ARIMA	Autoregressive Integrated Moving Average
RARM	Rapid association rule mining algorithm
FP-Tree	frequent pattern tree algorithm
AIS algo	Artificial Immune System algorithm
SVM	Support Vector Machine

Large-scale predictive market segmentation using machine learning & data analysis

Abstract

Market segmentation is splitting a market along likeness, kinship, or familiarity. In other words, a market segment's participants have something in common. Segmentation focuses marketing efforts and resources on the subdivision (or market segment) to acquire a competitive edge. Marketing and sales managers frequently make "best guesses" when identifying and categorizing market categories. As a result, several presumptions about the nature and significance of consumer variables and related data are made. The era of "big data" now offers the chance to "let the data tell you" Or to start with the data and move backward. In this study, we develop an approach based on three purchase characteristics (Recency, Frequency, and Monetary value - RFM) and product attributes to identify segments in the customer dataset. Using an unsupervised learning approach, we use a large set of point-of-sale (POS) data to segment a retail market. The product association is identified using Market Basket Analysis (MBA) on the POS data to find the frequent and cross-selling items.

Keywords: Market segmentation, market segment, big data, RFM, product attributes, segments, unsupervised learning, point-of-sale (POS) data, Market Basket Analysis, product association.

1. INTRODUCTION

Nearly every organization seeks compatibility between themselves and the surroundings in which they operate. This entails balancing the firm's market offering with market demands. However, this approach is complex without a precise aim. To solve this problem, marketing managers have historically used market segmentation to choose several groups of people with similar interests and purchasing patterns as the foundations for their marketing activities. Market segmentation is a crucial aspect of marketing strategy for businesses worldwide, allowing them to divide a heterogeneous market into smaller, homogeneous segments based on similar characteristics. These characteristics can include demographic, geographic, psychographic, and behavioral factors. Recent market segmentation strategies use inductive techniques to try to solve these problems. Most segmentation analyses are built on various

kinds of "cluster analysis," which is a collection of well-defined statistical methods that classify people based on how close together their ratings are.

This paper builds a market segmentation method based on market-based analysis and unsupervised learning. Unsupervised learning is a machine learning approach for drawing inferences from datasets having input data but no labeled answers. One of the popular techniques for identifying customer buying habits is market basket analysis (MBA), also known as association rule mining. Co-occurrences may be extracted from transactional databases in retail outlets using MBA.

The study is an example of using soft computing approaches to satisfy decision-maker's business requirements.

2. RELATED WORK

A. Customer Segmentation

Most individuals instantly consider psychological characteristics, lifestyles, values, habits, and multivariate cluster analysis techniques when "market segmentation" is mentioned. However, market segmentation is a far more significant idea affecting worldwide company operations. At the most basic level, "market segmentation" is based on similarity, closeness, and inherent characteristics. (Thomas, 2019)

The type of segmentation can be broken down into the following types. Geographic Segmentation, Distribution Segmentation, Media Segmentation, Price Segmentation, Demographic Segmentation, Occasion-Based Segmentation, and Lifestyle or Psychographic Segmentation. Most segmentation analyses are built on various kinds of "cluster analysis," which is a collection of well-defined statistical methods that classify people based on how close together their ratings are. (Thomas, 2019) also discusses the steps involved in conducting a segmentation analysis, common mistakes to avoid, the benefits of segmentation for businesses, and tips on avoiding common mistakes in segmentation studies, such as segmenting a segment, creating too many segments, and confusing the results.

Clustering is a commonly utilized method for creating or identifying database market segmentation. Over the years, cluster analysis techniques have been used to look for consistent groupings of customers. (Dolnicar, 2002) Methodology-based approaches employ

tools including fuzzy sets, genetic algorithms, and statistical methods to group data into homogenous clusters. A study by (Iromi R Paranavithana, 2021) proposes a two-stage process in which the first stage involves segmenting the retail market data, and the second consists of determining particular segments purchasing probabilities on a POS (Point of sale) data. Stage one uses an unsupervised learning technique to determine customer categories, which considers three purchasing factors (Recency, Frequency, and Monetary value - RFM) and product features. In stage two, Market Basket Analysis (MBA) estimates the likelihood of each segment's purchasing patterns. The study compares the performance of different clustering algorithms and machine learning techniques, such as K-Means, Ward's Minimum Variance, Expectation Maximization, Support Vector Machine, Random Forest Classifier, and Naive Bayes, and shows that the optimal algorithm depends on the dataset characteristics. The proposed approach helps retailers to identify high, medium, and low-value customers, their purchase behavior, and the placement of products according to customer levels. The approach is also useful in failure classification in retail stores or supermarkets.

(Ms. G. Nathiya, 2010) provides an analysis of the partition method clustering techniques, specifically the Expectation Maximization (EM) algorithm, K-means algorithm, and K* Means algorithm, on the HeartSpect dataset. Moreover, a comprehensive review of types of clustering algorithms is provided: **Exclusive Clustering**; data is organized distinctly; thus, if a particular datum exists and belongs to one specific cluster, it cannot be included in another cluster, Kmeans algorithm, for instance. Data is clustered using fuzzy sets using the **overlapping clustering** method. For instance, each point may fall into multiple clusters with varying degrees of membership, the Fuzzy C-means algorithm. **Hierarchical clustering,** the basis of this model, is the union of the two closest clusters in the provided dataset. The starting condition is achieved by designating every data point as a cluster. After a few more repetitions, the desired final clusters are reached—for example, the Agglomerative clustering algorithm. The last sort of clustering is **probabilistic clustering**, which relies on probabilities—for example Mixture of Gaussian algorithms. The paper also highlights the importance of clustering in various fields such as information retrieval, medicine, and archaeology, and suggests that there is still much room for improvement in clustering research.

RFM (Recency, Frequency, and Monetary Value of the Transaction) research has been used to segment prospective clients for over for segmenting prospective clients for more than 50 years. Market segmentation practices are based upon segmentation, positioning, and targeting

against potential opponents by breaking down heterogeneous markets into small homogenous markets. A study by (Deepali Kamthania, 2018) proposes a model that can leverage the user's location and interests for formulating business strategies. The proposed model used PCA (Principal Component Axis) to reduce the attributes, and the K mode clustering algorithm was used to segment the customers. Subsequently, the clusters formed were evaluated using the Silhouette Score Evaluation.

In the current market, catering to the customers' needs is essential. Enhancing customer relationship management is crucial, and it may be done by breaking out the market utilities based on variables such as supply, demand, and weekly sales. (Muthalagu, 2020) proposes a study that aims on enhancing Customer Relationship Management by segmenting the market utilities using various factors such as weekly sales, demand, and supply. The paper's objectives are to find the top profitable products in the market, understand customer behavior, and forecast sales using ARIMA. The authors use k-means clustering, a popular unsupervised learning model, to cluster customers based on their weekly sales behavior. They also apply the moving average technique to understand market behavior and use ARIMA to forecast future sales. The ARIMA model was chosen because it works well on stationary data with a single feature. The paper concludes that data analysis techniques are vital in the market industry to obtain patterns from available data and predict the future of the market to make profitable decisions.

A study (Dolnicar, 2002) highlights the unquestioned standards in cluster analysis. The study aims to investigate three areas after reviewing 243 data-driven segmentation studies- 1. Whether and which cluster analysis technique is surfaced in the market segmentation literature. 2. To contrast these application criteria with the theoretical understanding of clustering techniques. 3. Propose modifying the clustering techniques. As a result, it was found that 1. The cluster analysis is not practiced exploratively. 2. The characteristics of the algorithm are not studied, which leads to improper results; the size of the sample with the number of variables is not questioned. 3. Hyperparameter tuning is missing from most studies, which inhibits the algorithm from working beyond its default settings. The clustering method selected should consider factors such as the number of variables, data size, sample size, data format, and association measure. Using k-means clustering and the ARIMA model in this research can help markets maximize their profits.

B. Association Rules

The extraction of knowledge from vast amounts of data is known as data mining. Using association rule mining, linkages between many data points are found. Numerous businesses are worried about mining association rules from their databases since enormous amounts of data are continuously gathered and maintained in databases. Market basket analysis is an illustration of association rule mining in action. This strategy evaluates consumer purchasing habits by analyzing relationships between different things customers put in their shopping baskets. (Savi Gupta, 2014) The authors provide a survey of existing data mining algorithms for market basket analysis, with a particular focus on the apriori algorithm. They suggest that this algorithm can be modified to improve time complexity and accuracy. The paper concludes by highlighting the importance of data mining tools in today's data-driven world and the potential for further research in this area.

(Manpreet Kaur, 2016) discusses Market Basket Analysis (MBA) as a data mining technique used in various fields like marketing, bioinformatics, education, and nuclear science. The paper proposes a new MBA algorithm, which mines static data and captures changes in data over time. The proposed algorithm can help find interesting patterns from large amounts of data, detect fraud, predict future association rules, and find outliers. The study also mentions that periodic mining is a new approach in data mining that is gaining significance due to needs in different applications and limitations of data mining, and it may enhance the power of existing data mining techniques.

In the study done by (Qiankun Zhao, 2003), various association techniques such as the Apriori algorithm, RARM algorithm (rapid association rule mining algorithm), FP-Tree algorithm (frequent pattern tree algorithm), and AIS algorithm has been studied. Apriori, however, is the easiest to use and represents the most advancement over earlier algorithms. (Loraine Charlet Annie M.C., 2012) describes how market basket analysis is utilized by retail organizations to improve customer satisfaction and profit by determining the placement of goods and designing sales promotions for different customer segments. The focus is on a leading supermarket, ‘Anantha Stores’, which uses frequent itemset mining and the K-Apriori algorithm to generate highly informative frequent itemsets and association rules. The paper emphasizes the importance of market basket analysis for retail organizations and highlights the effectiveness of the K-Apriori algorithm in generating informative association rules.

3. METHODOLOGY

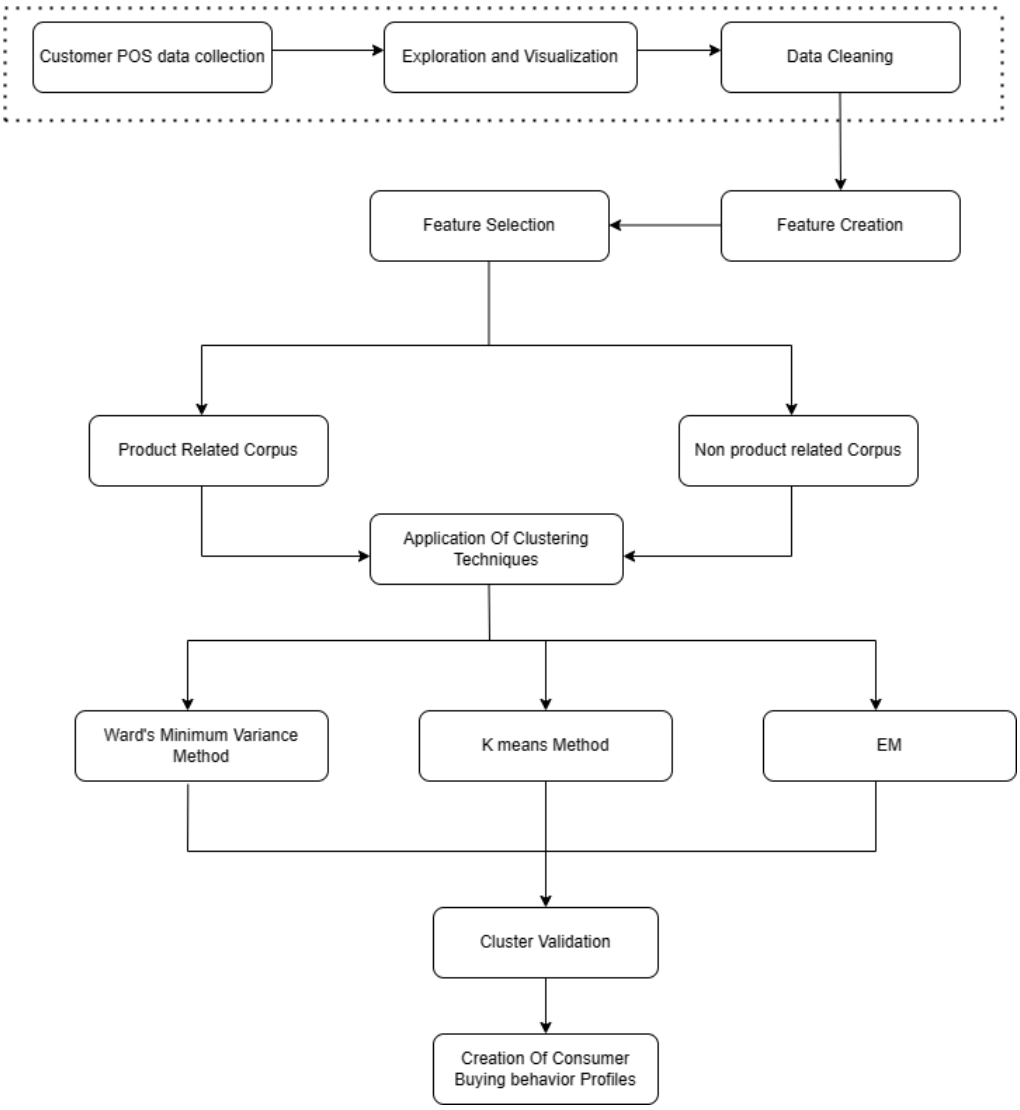

Fig1: Overview of the methodology

A. Objective

Marketing has always relied on market segmentation to tailor marketing strategies to specific customer groups. Traditionally, based on limited data, marketers have made educated guesses about market categories and customer characteristics. However, with the advent of big data, it is now possible to let the data guide the identification of market segments, allowing for more precise targeting and improved marketing outcomes. This project uses machine learning and data analysis to segment a retail market based on three purchase characteristics (Recency, Frequency, and Monetary value - RFM) and product attributes. Specifically, unsupervised learning techniques were applied to a large point-of-sale (POS) dataset to identify consumer

groupings based on their purchasing behavior profiles. This project aims to provide marketing and sales managers with a reliable approach to segment their customers according to their business requirements, thus allowing them to tailor their marketing efforts to specific customer groups and gain a competitive advantage.

Qualitative and quantitative data is required to implement market segmentation, consisting of segmentation variables and several descriptor variables. The data used in this study is secondary data.

B. Data Collection methods

The data collection method used for this project is secondary data collection, which involves collecting data from existing sources. In this case, the dataset used for the study was obtained from the retail company's internal database. The dataset consists of transactional data, including product descriptions, the number of goods bought, prices, and the date of purchase.

The dataset used for this study is a historical transnational dataset with transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The dataset features real and integers and contains multivariate, sequential, and time-series characteristics. There are 541,909 instances in the collection, each with eight properties. The following briefly describes the attributes in the dataset: Invoice number: Nominal, a characteristic 6-digit number explicitly issued for each transaction. This code denotes a cancellation if it begins with the letter "c." Stock Code: Product or item number, a 5-digit integral number known as the nominal, is issued to each unique product. Description: Name of the product (item), nominal. Quantity: the number of each item (product) in a single transaction, numeric. Invoice Date: the day and time each transaction was created, expressed as a number. Unit Price: Numeric, sterling price per unit of the product. Customer ID: Nominal, a five-digit integral number issued to every client separately. Country: Nominal, the nation’s title in which each client is domiciled.

C. Tools Used:

- **Data Handling and pre-processing:** Google Collab IDLE, Python, pandas, NumPy, Scipy, Regex, datetime, Scikit Learn, Scipy
- **Data Visualisation:** Matplotlib, Seaborn, Networkx, Plotly, Geopandas, Scipy
- **Clustering:** Scikit Learn

- **Predictive Analysis:** Scikit Learn
- **Market Basket Analysis:** Mlextend

D. Pre-processing

I) Initial Finding:

- An initial review revealed that the data had 5268 duplicate entries, 1454 null values in the ‘Description’ attribute, and 135080 null values in the ‘CustomerID’ attribute, respectively.

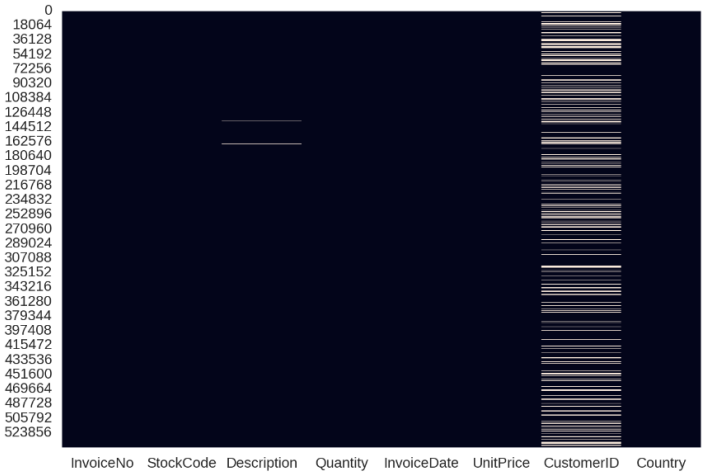

Fig 2: Heat map of null values

- In the attribute ‘Quantity,’ maximum and minimum values were both 80995, which indicates that the data entering order was reversed. The problem is that the dataset still contains both the original and reversal item.

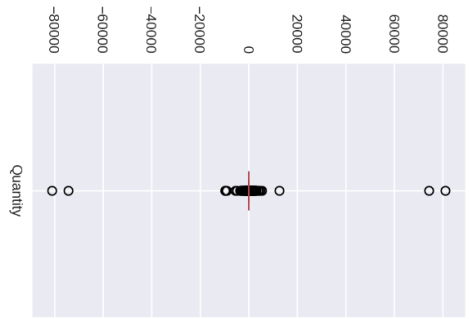

Fig 3: Box-plot of Quantity

- The attribute ‘UnitPrice’ has negative values, which is unusual since it would result in a corporation losing money. These transactions can reflect orders clients canceled or the company had to write off the lousy debt.
- Since market/customer segmentation necessitates categorizing each consumer, the absence of uniquely recognized customers might be difficult.

II) Feature Engineering

1. Handling Time and date

The transaction's date and time are included in the ‘InvoiceDate’ field. These data have been divided into independent columns to simplify future data processing and feature engineering.

2. Invoice No.

‘InvoiceNo’ provides information on the transaction's status (a "C" indicates a canceled transaction) and its identifier. Further feature engineering might be facilitated by extracting this information. Since lousy debt adjustments do not reflect actual sales, they are removed from the dataset. They are also not associated with any one consumer.

3. Unit Price

A new column ('TotalSum') might be added to show the entire amount paid by the client for the specific purchase; the unit price indicates the cost of a particular item. Rows with a total sum of zero (rows with 0 TotalSum) appear to be used to track various activities; these rows are deleted.

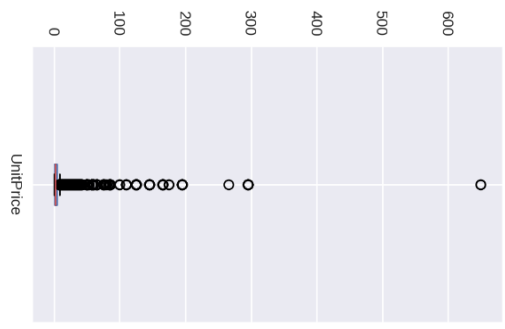

Fig 4: Box-plot of Unit Price

4. Extracting country location

Each nation's longitude and latitude locations might be included to show the distance between clients geographically (customers living near one another could have similar purchasing habits). There are 224 rows in which the nation is unspecified. The most common nation mentioned in these countries will be imputed, which is the United Kingdom.

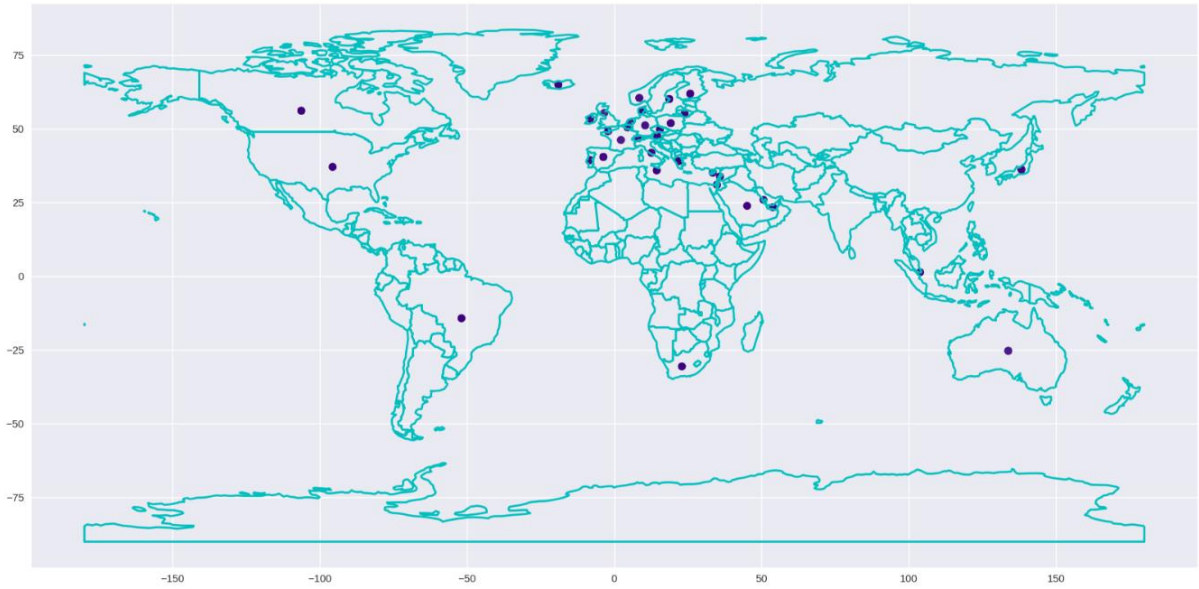

Fig 5: Distribution of Sales all over the world

5. Cohort Analysis

In this project, Cohort analysis is done by creating cohorts(groups) of customers based on the initial and frequent purchases to find the retention rate, average quantity, and average count.

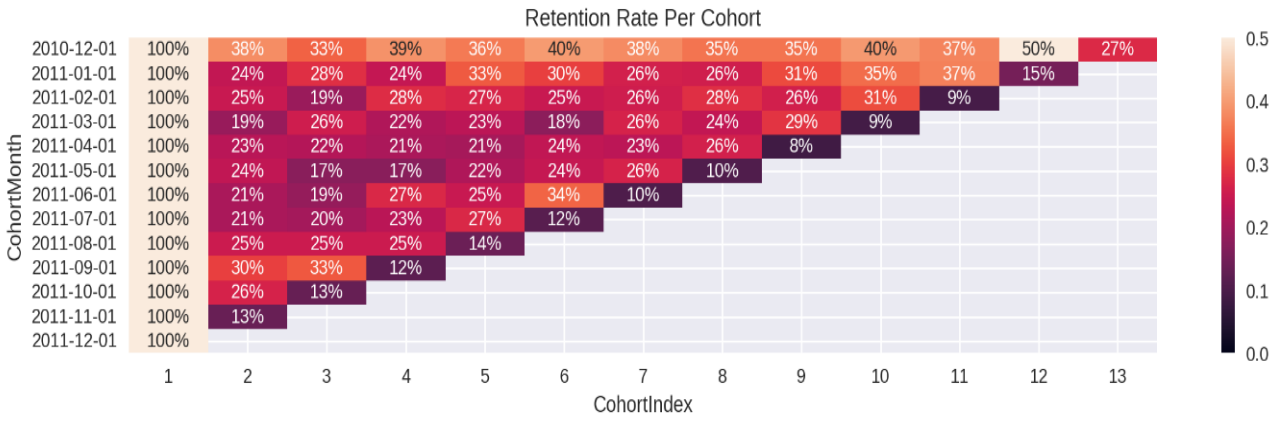

Fig 6: Customer retention rate

Fig 7: Average quantity per cohort

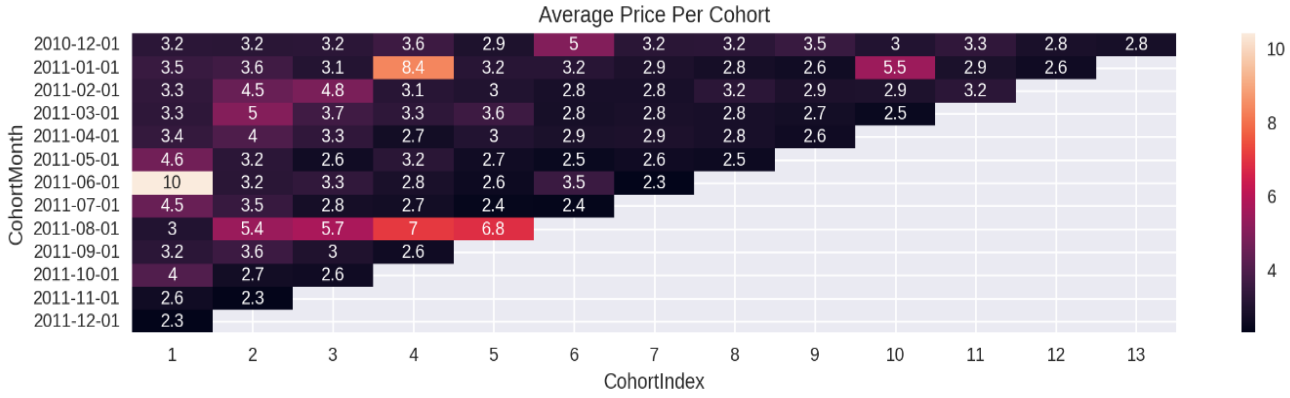

Fig 8: Average Price per cohort

Inferences that can be made from these graphs are:

1. Over time, the customer retention rate declines.
2. Two factors are measured by customer retention:
   - How many clients each cohort had at the beginning (acquisition month)
   - How many will be active in the upcoming months?
3. With occasional increases and dips in activity, about 25% of newly gained clients will make a repeat purchase, which stays relatively steady.
4. The number of goods purchased appears to grow throughout the Christmas season.
5. During the Christmas season, more significant purchases are made.
6. During the Christmas season, customers frequently make larger purchases; this might be because a more significant discount or incentive is offered.

6. Segmentation based on Recency, Frequency, and Money (RFM)

Recency: How recently did each customer's most recent purchase occur?

Frequency: The number of times the consumer bought in the previous 12 months

Monetary Value: How much did the client spend over the last 12 months?

The RFM values and the RFM score, which give a comprehensive idea about the three parameters (Recency, Frequency, and Monetary Value), are created for 12 months from 2010-12-10 to 2011-12-09.

CustomerID	Recency	Frequency	MonetaryValue
12346.0	326	2	0.00
12347.0	3	151	3598.21
12348.0	76	31	1797.24
12349.0	19	73	1757.55
12350.0	311	17	334.40

Table 1: Sample table of RFM values

- Quantiles are created to group the customers based on Recency, Frequency and Monetary values.

CustomerID	Recency	Frequency	MonetaryValue	Recency_Q	Frequency_Q	MonetaryValue_Q
12346.0	326	2	0.00	1	1	1
12347.0	3	151	3598.21	4	4	4
12348.0	76	31	1797.24	2	2	4
12349.0	19	73	1757.55	3	3	4
12350.0	311	17	334.40	1	1	2

Table 2: RFM table with quantile segments

- These attributes form RFM scores using all the quantiles, simply merging the quantile values.

CustomerID	Recency	Frequency	MonetaryValue	Recency_Q	Frequency_Q	MonetaryValue_Q	RFM_Segment	RFM_Score
12346.0	326	2	0.00	1	1	1	1.01.01.0	3
12347.0	3	151	3598.21	4	4	4	4.04.04.0	12
12348.0	76	31	1797.24	2	2	4	2.02.04.0	8
12349.0	19	73	1757.55	3	3	4	3.03.04.0	10
12350.0	311	17	334.40	1	1	2	1.01.02.0	4

Table 3: RFM values with calculated RFM score

- A higher RFM score indicates that the customer adds more value to the recency, frequency, and Monetary Values. Moreover, further customer segments can be created based on this RFM score which would give a comprehensive idea about the customer RFM values. The customers who have an RFM score greater than equal to 9 falls under the ‘Gold’ category; a score greater equal to 5 and less than 9 would indicate a ‘Silver’ category and the rest ‘Bronze’ category.

	Recency	Frequency	MonetaryValue	
	mean	mean	mean	count
General_Segment				
1.Gold	25.2	191.0	4036.1	1662
2.Silver	94.5	35.4	622.8	1867
3.Bronze	206.5	10.8	162.8	766

Table 4: Customised customer segments

E. Clustering Analysis

RFM, a grouping of the three variables, is frequently used in consumer segmentation for marketing objectives. This study uses three algorithms—Kmeans, Ward's Minimum Variance technique, and EM—along with consumer segmentation based on RFM. The three input variables (RFM) were created from the pre-processed, converted, and normalized customer data to eliminate positive skewness.

Skewness before transformation

```
Recency's: Skew: 1.2444937290057807, : SkewtestResult(statistic=26.458726592189453, pvalue=2.895613835933345e-154)
Frequency's: Skew: 18.12183686440274, : SkewtestResult(statistic=83.17480537059672, pvalue=0.0)
MonetaryValue's: Skew: 21.459150114624148, : SkewtestResult(statistic=86.8803096286766, pvalue=0.0)
```

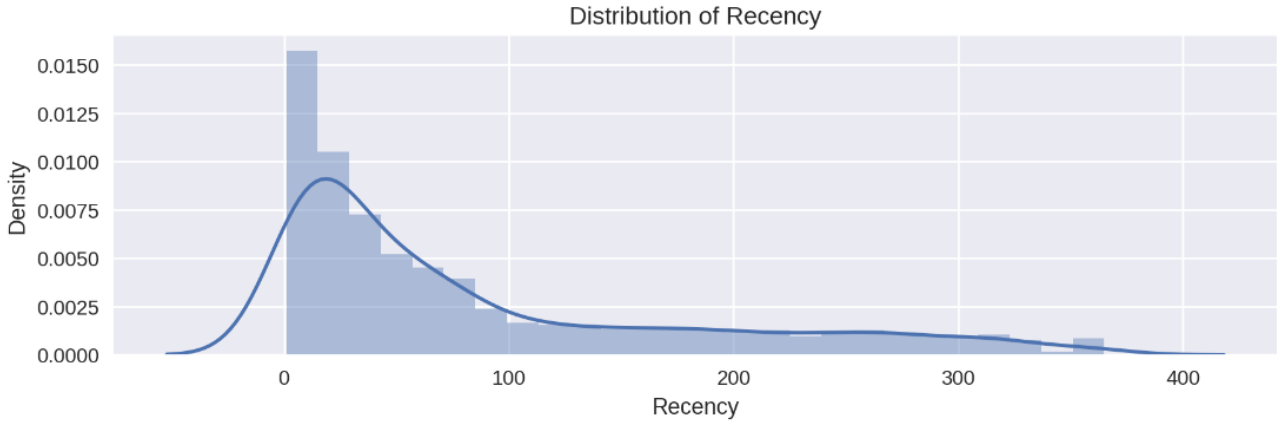

Fig 9: Density distribution of Recency

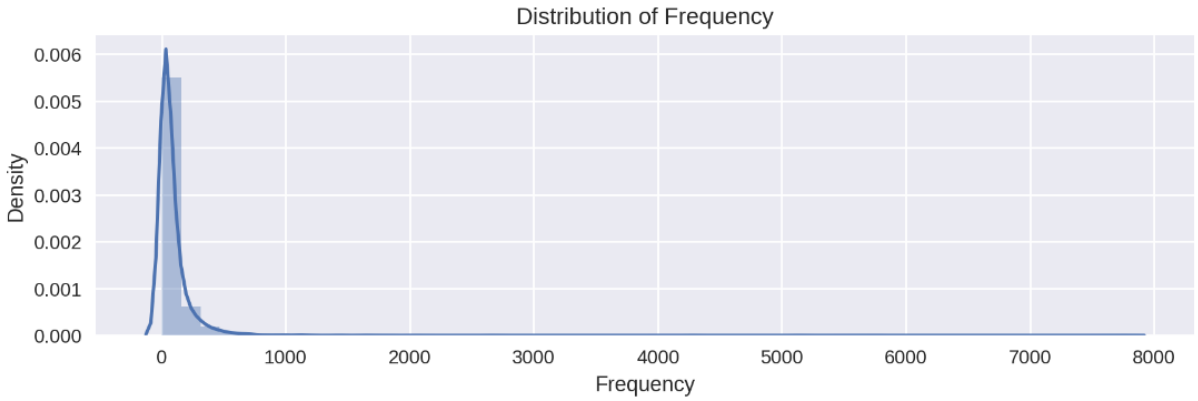

Fig 10: Density distribution of Frequency

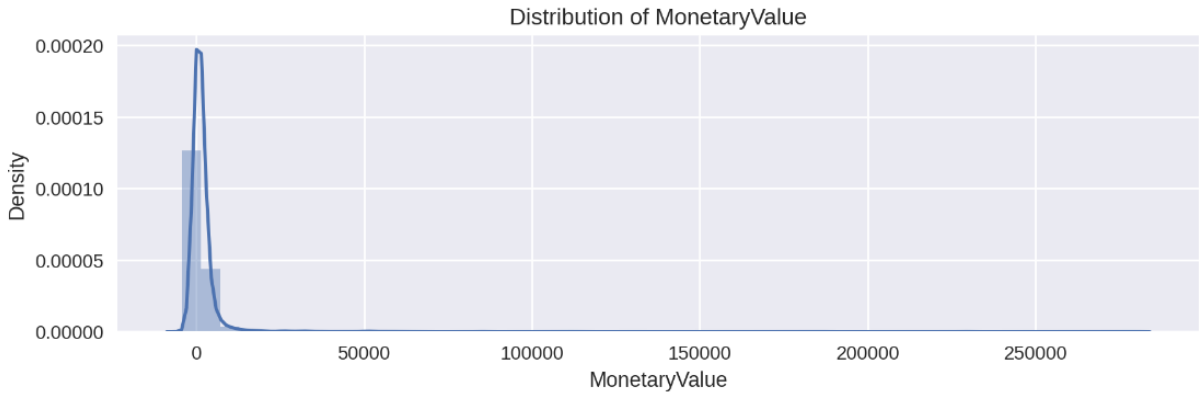

Fig 11: Density distribution of Monetary Value

Skewness after transformation

```
Recency's: Skew: -0.4547480558293254, : SkewtestResult(statistic=-11.634943795553278, pvalue=2.737704076465579e-31)
Frequency's: Skew: -0.239827897754714, : SkewtestResult(statistic=-6.339202735355403, pvalue=2.309571642177164e-10)
MonetaryValue's: Skew: 0.1289857073105935, : SkewtestResult(statistic=3.44287521745169, pvalue=0.0005755648170534836)
```

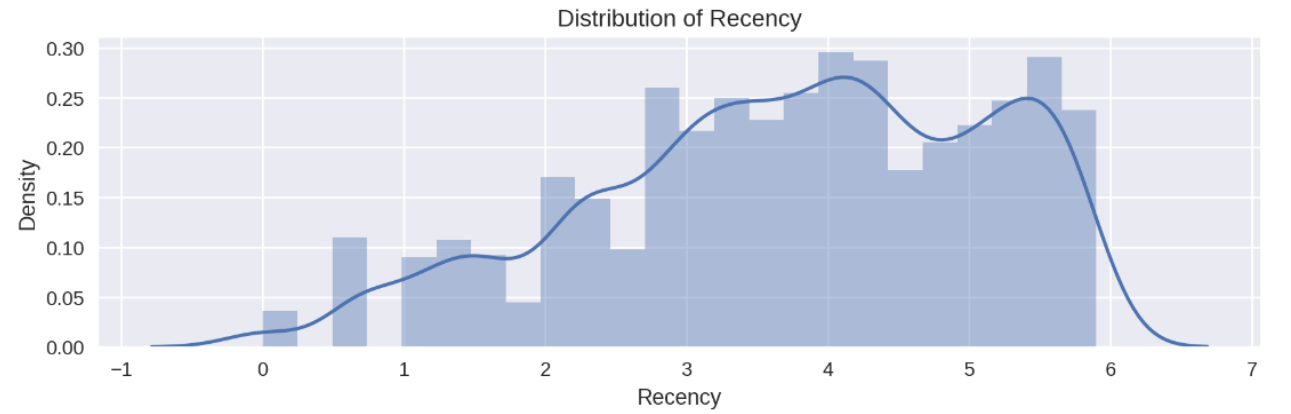

Fig 12: Density distribution of Recency

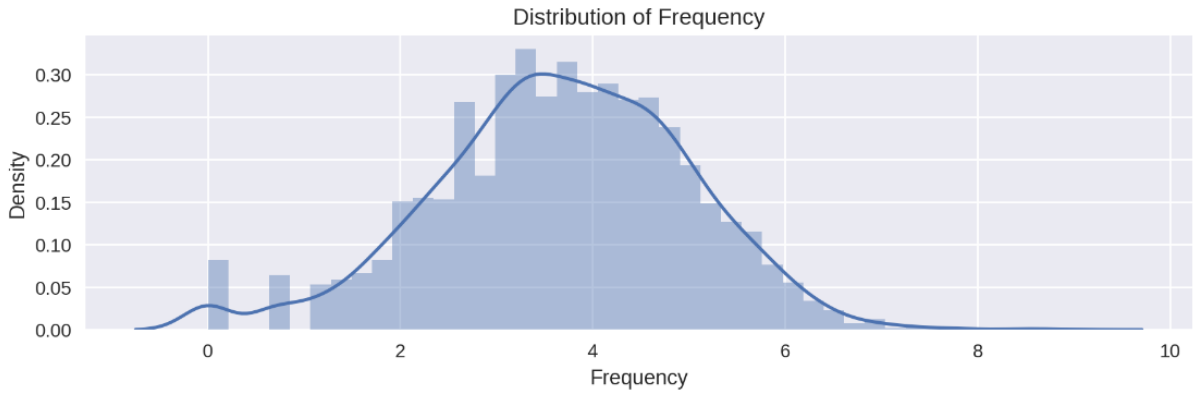

Fig 13: Density distribution of Frequency

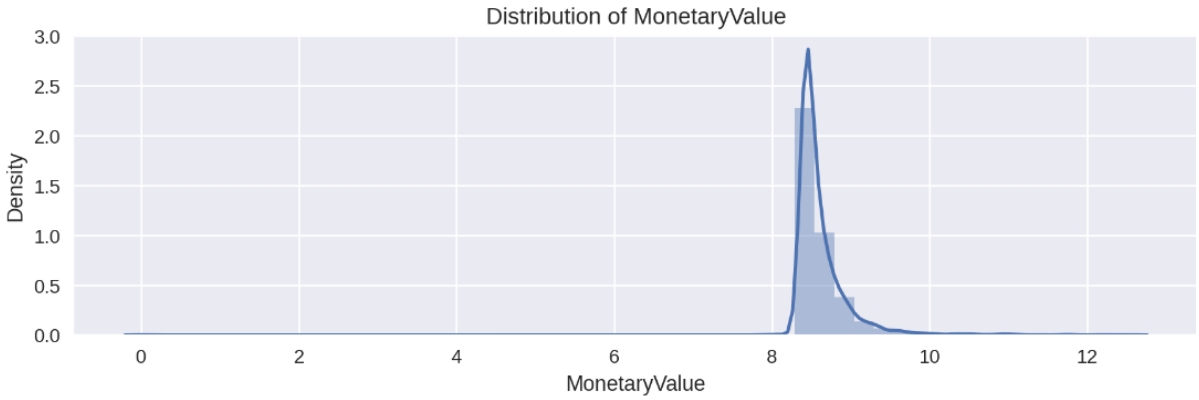

Fig 14: Density distribution of Monetary Value

I) K means Algorithm

K-means is a popular unsupervised learning algorithm used for clustering data. The K means algorithm assigns each data point to the cluster nearest to the centroid values. The coordinates of the center are the arithmetic means of each dimension across all of the cluster's points, making it the average of all the cluster's points. Each data value is assigned to the closest cluster at each iteration of the algorithm based on a similarity criterion like Euclidean distance. Following this, the partitions are recalculated using these hard assignments. A data value may swap clusters with each subsequent pass, changing the values of the cluster at each pass.

The algorithm followed in K means is as follows:

- Pick the k clustering value.
- Create k clusters randomly, then identify their centers, or create k random locations and use those as the cluster centers.
- Identify the closest cluster center for each location.
    - calculate the new cluster centers once more.
    - Until a convergence requirement is satisfied, repeat the previous two stages.

This method requires input data on how many clusters should be as different as feasible. The k-means approach will generate precisely k unique clusters with the highest level of

differentiation. The best number of clusters in KMeans can be found using the elbow method and silhouette analysis methods.

The elbow approach includes charting the total squared distances from every point to the nearest cluster center versus the number of clusters. The plot resembles an arm, and the point of diminishing returns, where adding more clusters does not appreciably reduce the sum of squared distances, lies at the "elbow" of the arm.

Silhouette analysis calculates a silhouette score for each point, which measures how similar a point is to its cluster compared to others. The average silhouette score is used to assess the clustering quality after the silhouette scores for each cluster are averaged. A higher score indicates better-defined clusters.

Initially, inertia/ within the sum of center squares was calculated for 2 to 10 clusters:

No. of Clusters	Inertia/ within the sum of center squares
2	7710.74
3	5976.01
4	5002.13
5	4204.51
6	3656.33
7	3127.28
8	2704.91
9	2454.64
10	2232.29

Table5: Inertia/ within the sum of center squares of clusters

As expected, the inertia/ within the sum of center squares decreases with the increase in clusters from 2 to 8. To know the favorable number of clusters, plotting a graph depicting the inertia vs. clusters is necessary. The chart will help analyze the saturation point of the distances, and the point where there is an abrupt decrease in the distance will be considered the elbow point.

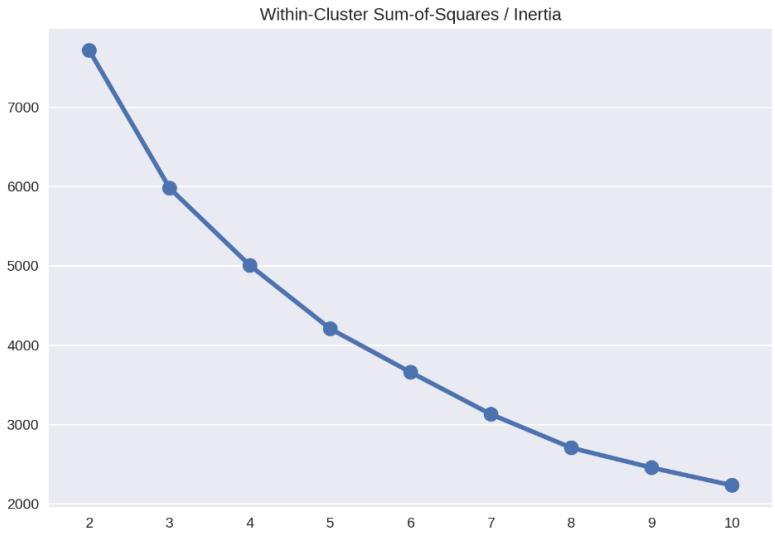

Fig15: Elbow plot

There isn't proper visibility of an elbow point, but it has to be a four or 5-cluster solution. The silhouette score is calculated for 3, 4, and 5 cluster solutions to test the cluster solution further.

No. of Clusters	Average silhouette score
3	0.35290092207973167
4	0.2878744760894411
5	0.29671389243672125

Table 6: Average silhouette score for 3, 4, and 5 cluster solution

According to the results, a 3-cluster solution is much better than a four or 5-cluster solution. Moreover, these results can also be interpreted using Silhouette plots of a particular cluster solution.

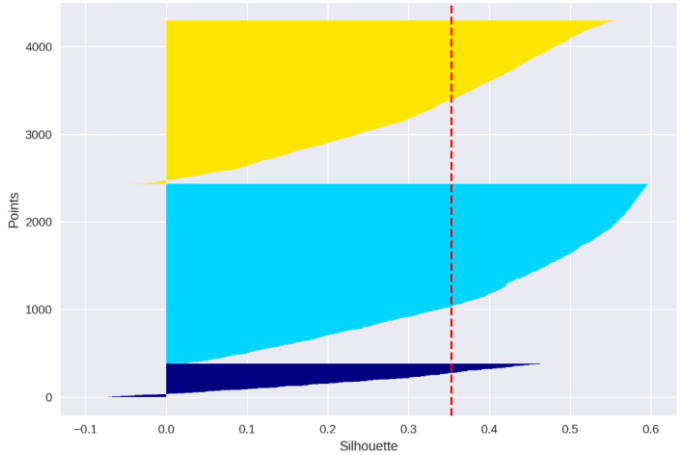

Fig16: Silhouette plot for 3 cluster solution

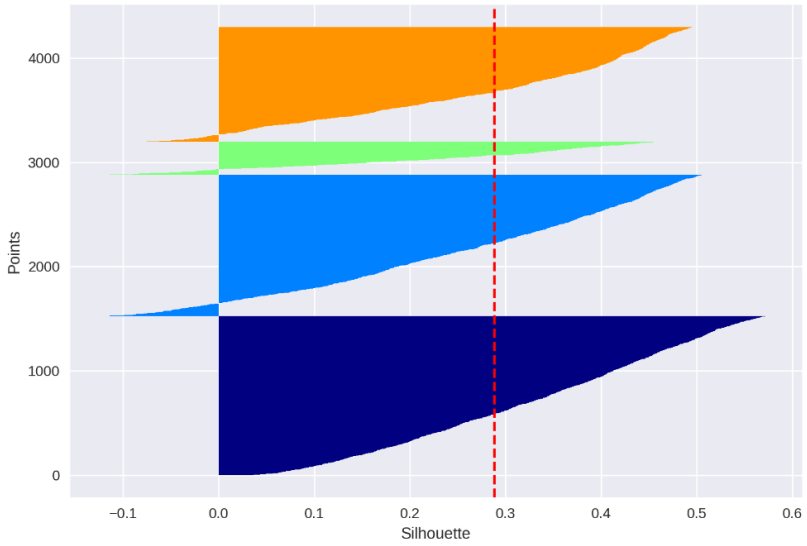

Fig17: Silhouette plot for 4 cluster solution

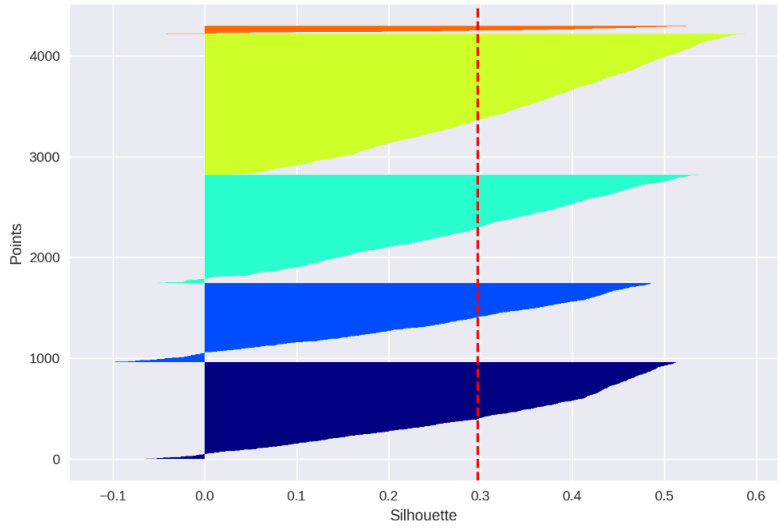

Fig18: Silhouette plot for 5 cluster solution

II) Expectation Maximization Algorithm

EM approaches are used to identify clusters in observations (or variables) and then place those observations inside the clusters. This clustering method's primary strategy and logic is to calculate a single continuous significant variable in a massive sample of observations. The distribution of values for the big continuous variable follows a normal distribution if the sample of the supplied dataset comprises clusters of observations, each having a different mean. Based on one or more probability distributions, the EM clustering method calculates the odds of cluster membership. The clustering algorithm's objective is to maximize the data's

overall probability or likelihood based on of the (final) clusters. The EM method may be used with categorical and continuous data.

Choosing the input divisions comes first. The Expectation step, which kicks off the EM cycle, is defined by the equation below:

$$E[z_{ij}] = \frac{p(x = x_i | u = u_j)}{\sum_{n=1}^{k} P(x = x_i | u = \mu_n)}$$

$$= \frac{e^{-\frac{1}{2}\sigma^2 \left(x_i - \mu_j\right)^2}}{\sum_{n=1}^{k} e^{\frac{-1(x_i - \mu_n)^2}{2\sigma^2}}}$$

According to this equation, the weight or expectations for pixel z concerning partition j are equal to the likelihood that x is pixel x_i provided that μ is partition μ_i divided by the total of the same probability over all partitions k. The covariance of the pixel data is represented by the sigma squared in the second equation. The M step, also known as the maximization step, starts after the E step has been completed and every pixel has been assigned a weight or expectation for each division.

$$u_j \leftarrow \frac{1}{m} \sum_{i=1}^{m} E[z_{ij}] x_i$$

This EM cycle is repeated until there is no noticeable change in the partition values. The optimal number of clusters in the EM algorithm can be found using measures like:

BIC (Bayesian Information Criterion): It is a metric for gauging how well a statistical model fits the data, considering both the data's probability and the model's complexity. BIC penalizes the number of parameters in the model while clustering, which helps prevent overfitting. A lower BIC score indicates a better model fit.

AIC (Akaike Information Criterion): AIC is a metric for the quality of fit of a statistical model, just like BIC is. The number of parameters is penalized less than BIC is, however. A lower AIC value indicates a better model fit.

No. of Clusters	BIC	AIC
2	27190.76	27069.83
3	25241.76	25057.16
4	24675.17	25057.16
5	24520.84	24208.95
6	24332.17	23956.62
7	23908.14	23468.94
8	23784.02	23281.17
9	23578.65	23012.14
10	23546.43	22916.28

Table 7: BIC and AIC scores for 2-8 clusters

Further plotting BIC and AIC scores vs. the number of clusters to get an elbow point.

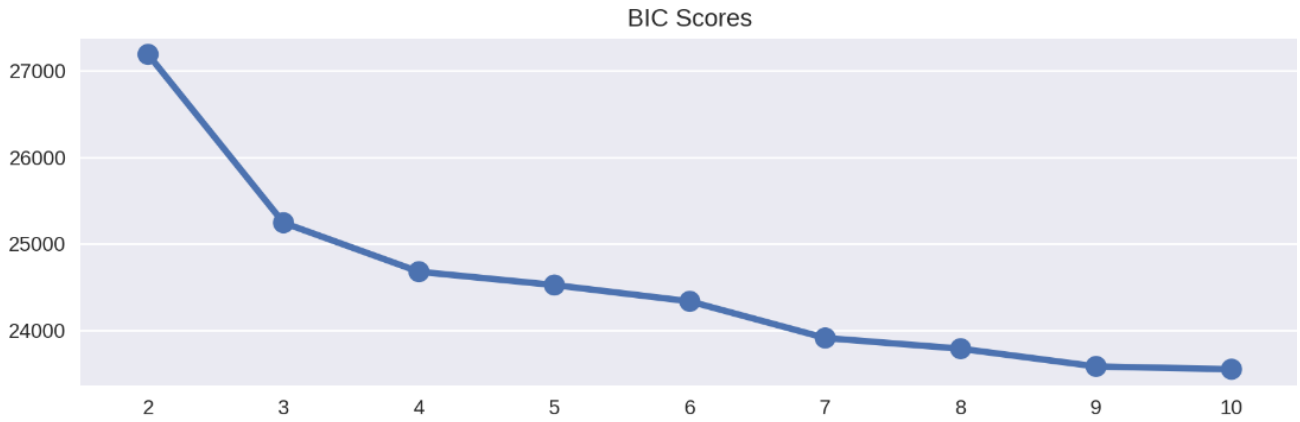

Fig19: BIC score vs no. of clusters

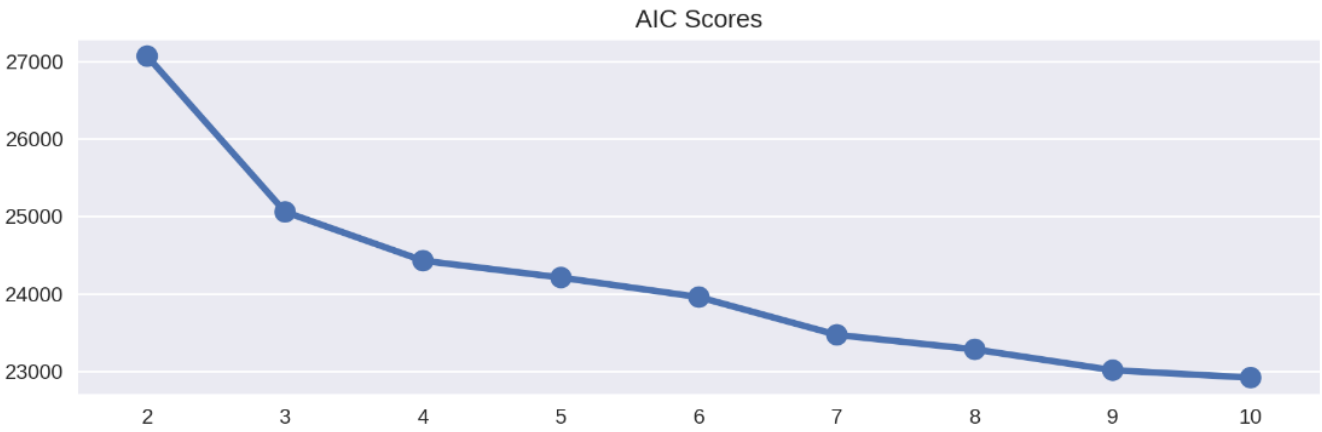

Fig20: AIC score vs no. of clusters

The plots do not give a discrete elbow point; the results get saturated among the 3, 4, and 5 cluster solutions. Further calculating Silhouette score and plot for 3, 4, and 5 cluster solutions.

No. of Clusters	Average silhouette score
3	0.25340880883944816
4	0.10744301843112017
5	0.12876687732434364

Table 8: Average silhouette score for 3,4 and 5 cluster solution

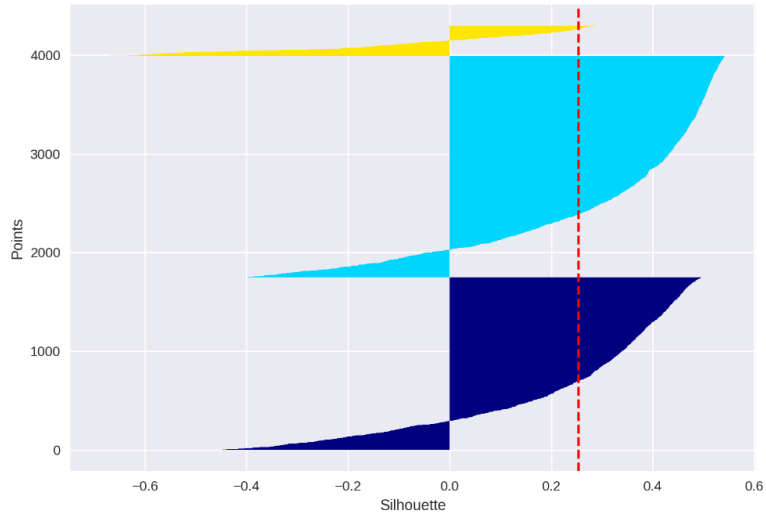

Fig21: Silhouette plot for 3 cluster solution

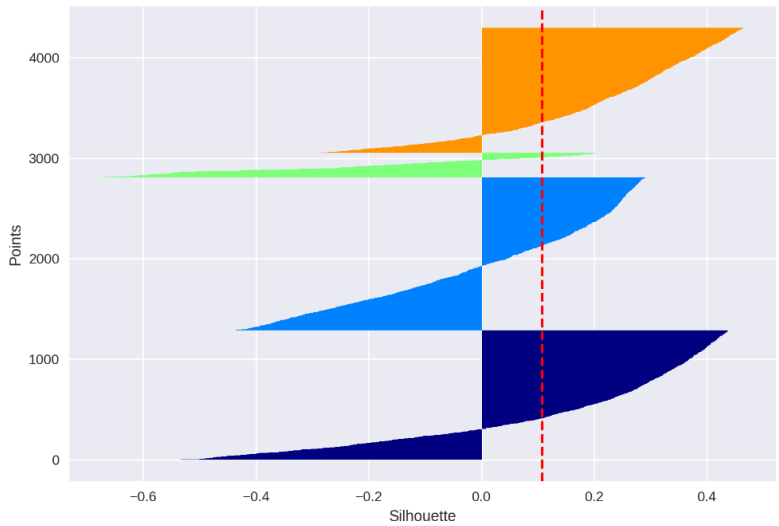

Fig22: Silhouette plot for 4 cluster solution

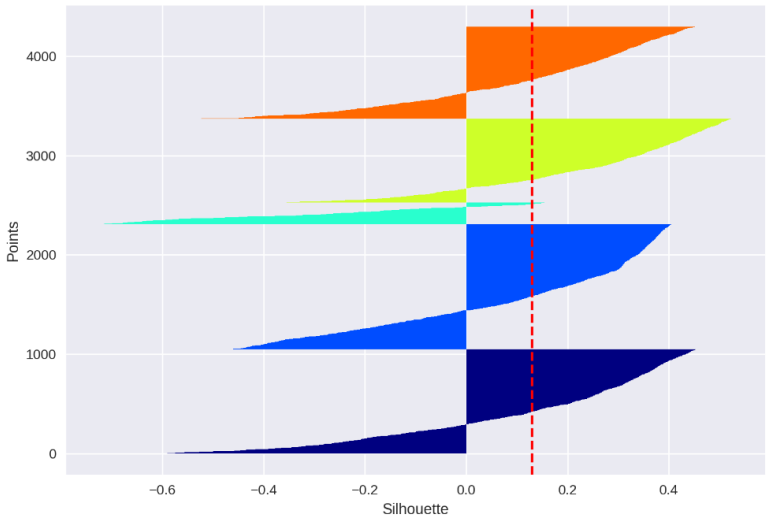

Fig23: Silhouette plot for 5 cluster solution

III) Ward's Minimum Variance Method

Ward's minimum variance method is a hierarchical clustering algorithm that seeks to reduce the overall within-cluster variation. The goal is to combine clusters with the minor overall within-cluster variance increase. The sum of the squared Euclidean distances between the newly merged cluster's centroids and the original clusters' centroids is used to calculate the increase in the overall within-cluster variance.

Each data point is first assigned to a separate cluster, after which the algorithm repeatedly merges the two clusters that cause the slightest increase in the overall within-cluster variance. This process is repeated until every data point is contained in a cluster. Ward's approach may yield clusters of various sizes and forms and is sensitive to the data scale.

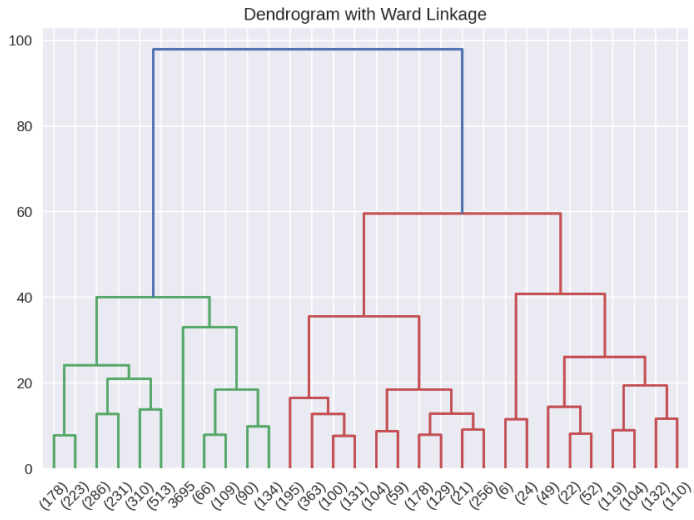

Fig24: Dendrogram for Ward's Linkage

The above figure shows that the optimal number of clusters from Ward's Linkage is 2. Further, a Silhouette score and plot can be figured out for a two-cluster solution.

The average Silhouette score for a two-cluster solution is 0.35350378584680436

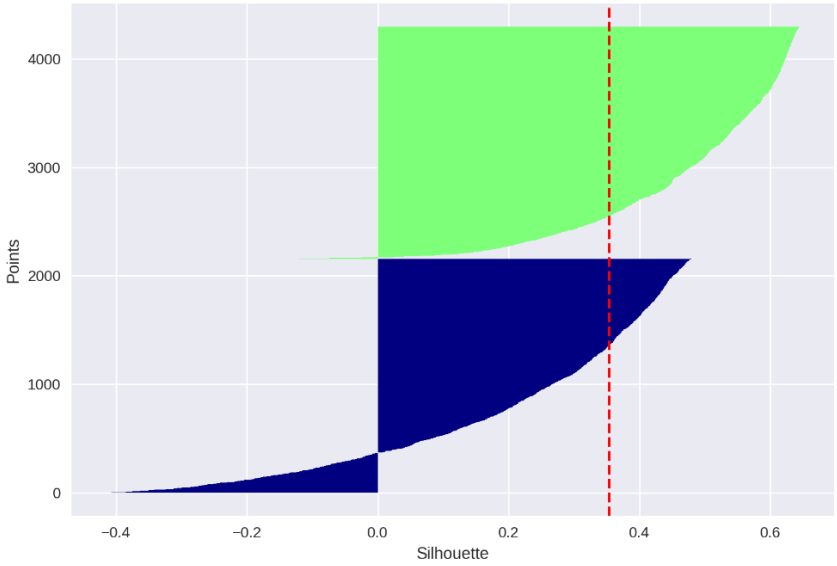

Fig25: Silhouette plot for 2 cluster solution

CONCLUSION OF CLUSTER ANALYSIS:

From all the mentioned techniques and their requisite measures, a 3-cluster solution best fits the data and the project's problem under consideration. The result of the 3-cluster Kmeans solution will be considered for further analysis.

F) PREDICTIVE ANALYSIS

This part will address the challenge of anticipating each customer's purchasing patterns, employing the primary classifier models, and assessing how well they function. The customer profile derived from the K-means clustering will be used. The dataset will be segmented into specific clusters and decided using clustering measures. Following this label, such as high-spending, low-spending, medium-spending, etc., labels will be assigned to the dataset. Predictive analysis can be leveraged in such datasets to understand their behavior and preferences. This can help make more informed business decisions, such as targeted marketing campaigns or product development initiatives tailored to each segment's needs and preferences. The dataset is divided into training and test sets, each accounting for 75% and 25% of the supervised task. Also, standardization has to be achieved to achieve a mean of 0

and a standard deviation of 1 unit. A 5-fold cross-validation is employed to discover the optimal values for specific hyperparameters for each model as part of a grid search.

I) Support Vector Machine

Support vector machines often deal with pattern classification. Hence this approach is mainly used to categorize various patterns. To separate the data points of different classes in a high-dimensional feature space, SVM seeks out the hyperplane that does so the best. The support vectors, the nearest data points from each category, are chosen to optimize the distance between the hyperplane and the support vectors.

The core principle of SVM is the creation of an ideal hyperplane for linearly separable patterns that may be utilized for classification. The best hyperplane maximizes the margin of the hyperplane or the distance from the hyperplane to the nearest point of each pattern. It is chosen from the collection of hyperplanes for categorizing patterns. The primary goal of SVM is to maximize the margin to accurately categorize the provided patterns, i.e., the higher the margin size, the more accurately it does so.

The equation shown below is the hyperplane representation:

The hyperplane, $aX + bY = C$

The basic concept of a hyperplane is depicted in the figure below, which shows what two distinct patterns look like when they are divided into three dimensions by a hyperplane Essentially, this plane consists of three lines, two of which are marginal lines and the other two of which are support vector lines on either side of the marginal lines.

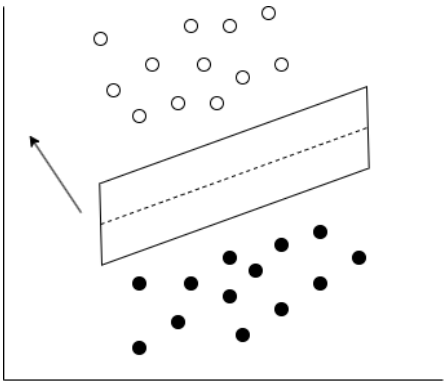

Fig26: plane classifying two classes

Both linear and non-linear classification problems may be completed with SVM. SVM discovers a linear hyperplane that divides the data points of several classes in linear classification. A linear hyperplane can divide the data points in non-linear classification using kernel functions to translate the data points to a higher dimensional feature space.

After the grid search, it was found that the best value for the regularization parameter is C = 1000. With this configuration, we achieved a training accuracy of 0.996274 and a test accuracy of 0.990688.

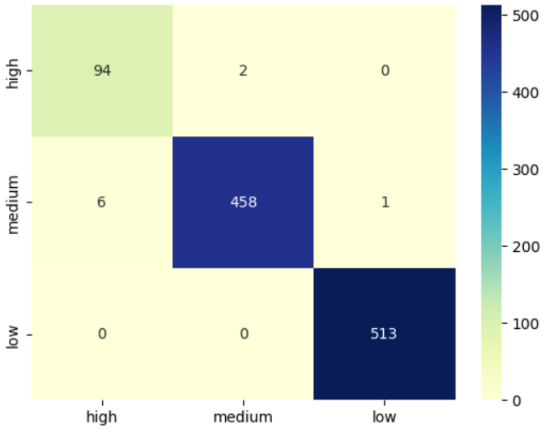

Fig27: Confusion matrix for SVC

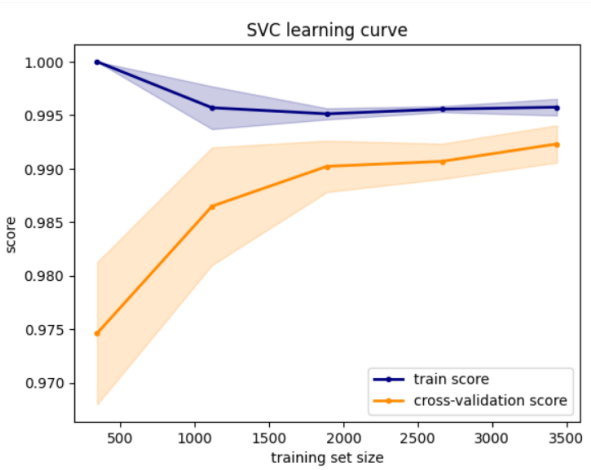

Fig28: SVC learning curve

II) Naïve Bayes Classifier

The Naive Bayes algorithm, based on the Bayes theorem, uses the assumption of independence between features to make classification decisions. The method determines the likelihood that a data point will fall into each potential class, then chooses the class with the highest likelihood to be the predicted class. The Bayes theorem serves as the foundation of the classifier.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

When B has already happened, we may use the Bayes theorem to calculate the likelihood that A will also occur. Here, A is the hypothesis and B is the supporting evidence. Here, it is assumed that the predictors and characteristics are independent. That is, one characteristic's existence does not change another's behavior. The term "naive" is a result.

Bayesian Inference

Let y be the class label and $X = (x_1, x_2, \ldots, x_n)$ be the parameters/ features. Bayes theorem can be rewritten as:

$$P(y|X) = \frac{P(X|y)\ P(y)}{P(X)}$$

expanding using the chain rule,

$$P(y|x_1, x_2, \ldots, x_n) = \frac{P(x_1|y)\ P(x_2|y) \ldots P(x_n|y)P(y)}{P(x_1)P(x_2) \ldots P(x_n)}$$

The denominator remains static, which can be removed,

$$P(y|x_1, x_2, \ldots, x_n) \alpha\ P(y)\Pi_{i=1}^{n} P(x_i|y)$$

There may be circumstances in which the categorization is multivariate. As a result, we must identify the class y with the highest probability.

$$y = argmax_y\ P(y)\Pi_{i=1}^{n} P(x_i|y)$$

The assumption of independence between features must be satisfied for Naive Bayes to perform and other classification algorithms while being quick and straightforward to construct. Naive Bayes classifiers come in three variations: Gaussian, Multinomial, and Bernoulli, each of which makes a different assumption about the input distribution. The project employs a GaussianNB, assuming the features' probability is Gaussian.

This model is worse than the previous ones; we achieve a training accuracy of 0.877290 and a test accuracy of 0.876753.

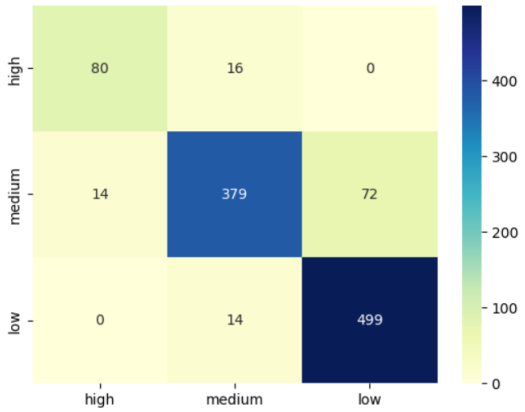

Fig29: Confusion matrix for GaussianNB

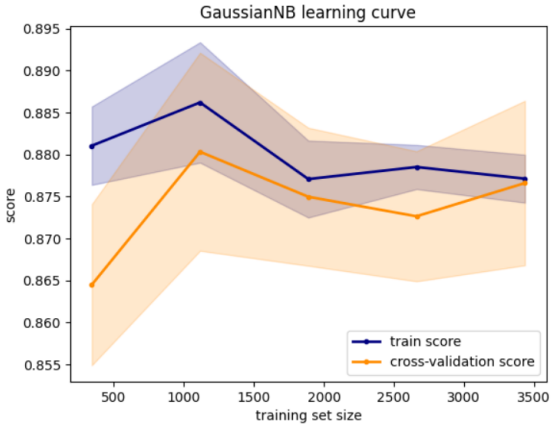

Fig30: GaussianNB learning curve

III) Ensemble Learning Classifier

Ensemble learning is a machine learning technique that combines multiple models to improve the overall performance of the predictive model. Ensemble learning's fundamental premise is to integrate the predictions of several models, each of which was trained on a different subset of the data, to provide a final prediction that is more reliable and accurate than any one individual model. There are various kinds of ensemble techniques, such as voting ensemble, bagging, boosting, and stacking.

1.) RANDOM FOREST

It is an ensemble model comprising several decision trees, and the outcome is determined by summing the predictions from each classifier. Due to its accuracy and resistance to overfitting, it is a powerful and well-liked machine learning method. Using a random selection of characteristics and training data, the random forest method creates many decision trees, which it then merges to get a final classification determination. This enhances the model's generalization capabilities and lowers variance.

In this case, the best hyperparameters are:

• max depth equal to 10

• max features equal to 1

• min samples leaf equal to 1

• min samples split equally to 2

• n estimators equal to 1000

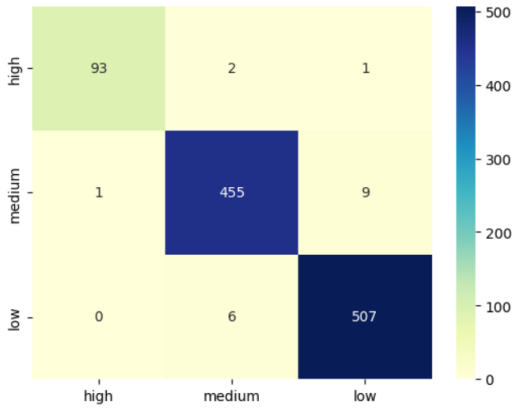

Fig31: Confusion matrix for Random Forest

Fig32: Random Forest learning curve

As a result, we have a training accuracy of 1.00 and a test accuracy of 0.984788; these results are confirmed by Figure 23 and 24, which shows that the model made some errors.

2.) VOTING ENSEMBLE

The Voting Ensemble is another popular ensemble method that combines the predictions of multiple individual models to improve the overall accuracy and stability of the model. This approach trains several base models with varying algorithms and hyperparameters on the same dataset. Each model produces its prediction during the prediction phase, and the voting ensemble aggregates them to produce the final prediction. The Voting Ensemble approach effectively reduces the model's bias and variance, leading to a better generalization performance on new, unseen data.

The model used has two base estimators- A support Vector Classifier (SVC) and Random Forest Classifier with the same configurations that gave the best results in individual cases, which are:

• max depth equal to 10

• max features equal to 1

• min samples leaf equal to 1

• min samples split equally to 2

• n estimators equal to 1000

• C = 1000

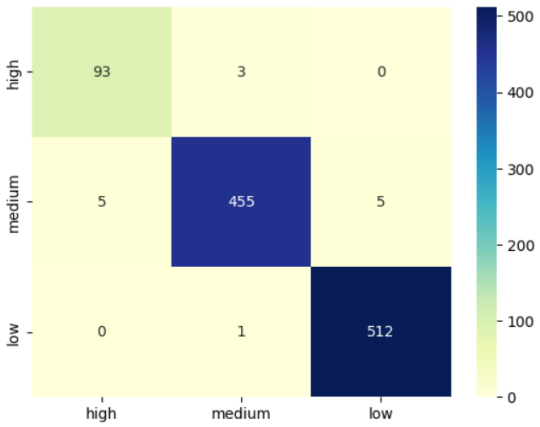

Fig33: Confusion matrix for Voting Ensemble

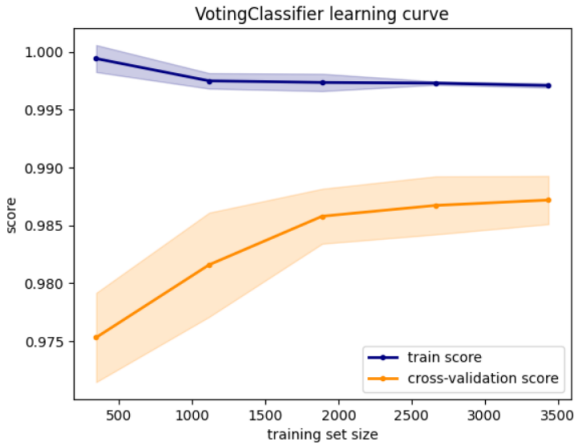

Fig34: Voting Ensemble learning curve

As a result, we have a training accuracy of 0.997594 and a test accuracy of 0.985408; these results are confirmed by Figure 25 and 26, which shows that the model made some errors.

3.) BAGGING ENSEMBLE

Bagging is a popular ensemble learning technique that involves creating multiple independent models on different random samples of the training data. These models are trained using bootstrap sampling, meaning each model is trained on a subset of the training data drawn randomly with replacement. Bagging helps to reduce variance and overfitting by generating a set of diverse models that work in unison to make a final prediction. The final prediction is obtained by taking the average of the predictions from all the models. Bagging is commonly used in decision tree-based models like Random Forest to improve their performance and robustness.

The model uses a base estimator as a Support Vector Classifier with the best setting of C=1000 and 1000 number of base estimators.

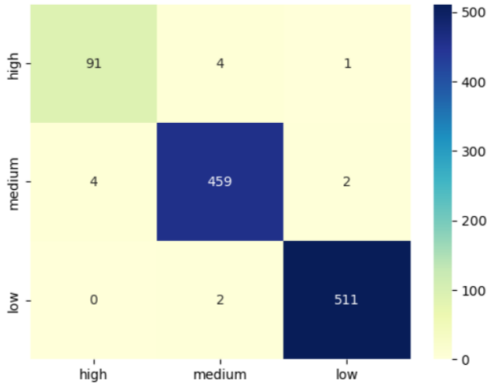

Fig35: Confusion matrix for Bagging ensemble

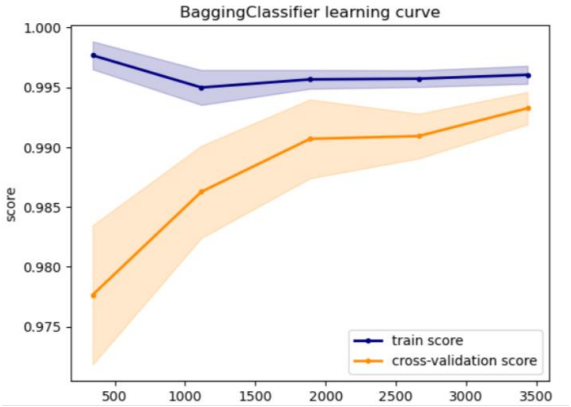

Fig36: Bagging ensemble learning curve

With this configuration, we achieved a training accuracy of 0.99658 and a test accuracy of 0.99006, which is slightly better than the others.

S.no	Classifier	Mean test accuracy
1	Support Vector Classifier	0.990688
2	Bagging Ensemble Classifier	0.99006
3	Voting Ensemble Classifier	0.985408
4	Random Forest Classifier	0.984788
5	Naïve Bayes Classifier	0.876753

Table 9: Comparison of supervised learners

CONCLUSIONS

The SVC is nearly flawless on the training and test sets and is the best model. However, the Bagging Ensemble, Voting Ensemble, and Random Forest have excellent performances and high accuracy. Overall, even if the findings from the other models are less accurate, except the Naïve Bayes classifier, they all have an accuracy of more than 93%, which is somewhat respectable. This is because the K-Means clusters are practically linearly separable, which indicates that the dataset is simple enough.

G) MARKET BASKET ANALYSIS

It is essential to determine the goods that customers could buy on their subsequent visit to the retail shop after identifying their clusters of customers. An association rule mining-based Apriori method will be implemented to find the linked rules unique to the goods in question. Using itemsets of length k-1, the Apriori method creates candidate itemsets of length k. The

algorithm first searches the database for each item's frequency. The process then produces frequent itemsets of length one and repeatedly increase their length to produce frequent itemsets of length k, continuing until no more frequent itemsets are produced. Association rules are then derived from each of the frequently created itemsets. The formula for an association rule is X →Y, where X and Y are itemsets and X ∩Y =∅. Only laws that meet a predetermined minimum support and confidence level are retained. Each rule's support and confidence are computed.

ITEMS BY FREQUENCY AND TOP-GROSSING ITEMS

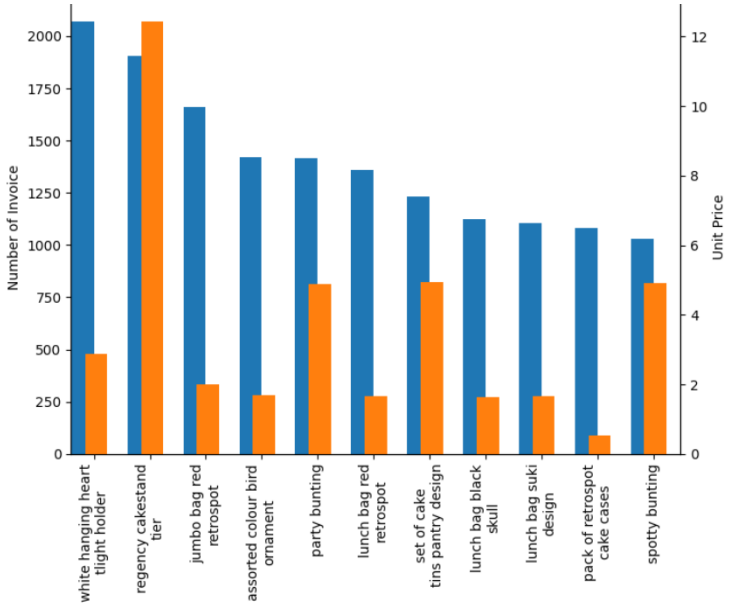

Fig37: items by frequency

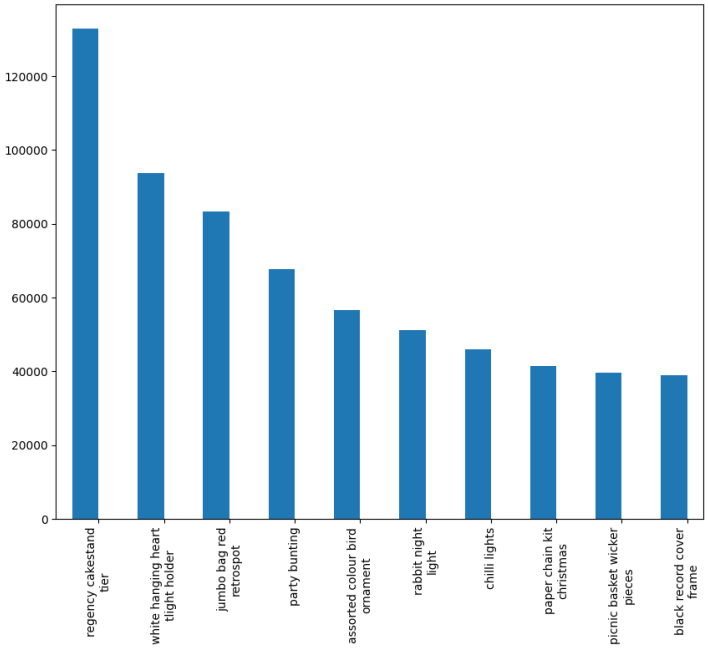

Fig38: top-grossing items

According to the chart above, the postal product "Regency cakestand tier" has the most sales. Another intriguing finding for the same item is that despite its high unit price, it frequently appears in several bills.

I) MODEL DESIGN

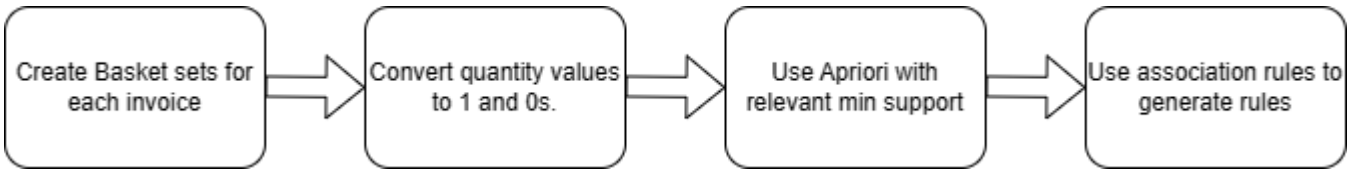

Fig39: MBA model design

Support: How many transactions include both items A and B in them. Range = [0,1]. A 1% minimum support level is utilized for the analysis.

Confidence: A given A frequency, the co-occurrence of items A and B. Range = [0,1]. A minimum confidence criterion of 40% is utilized for the study.

Lift: It reveals the impact of a rule's strength on the incidence of items A and B. The strength improves with increased lift. Range: [0, inf)

Following metrics computation and rule generation utilising a few threshold values, the association rules are as follows.

antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
{'ALARM_CLOCK_BAKELIKE_IVORY'}	{'ALARM_CLOCK_BAKELIKE_GREEN'}	0.021591	0.035136	0.012629	0.584906	16.646924	0.011870	2.324445
{'ALARM_CLOCK_BAKELIKE_GREEN'}	{'ALARM_CLOCK_BAKELIKE_PINK'}	0.035136	0.025003	0.014054	0.400000	15.998371	0.013176	1.624996
{'ALARM_CLOCK_BAKELIKE_PINK'}	{'ALARM_CLOCK_BAKELIKE_GREEN'}	0.025003	0.035136	0.014054	0.562118	15.998371	0.013176	2.203480
{'ALARM_CLOCK_BAKELIKE_GREEN'}	{'ALARM_CLOCK_BAKELIKE_RED'}	0.035136	0.038650	0.023118	0.657971	17.024025	0.021760	2.810728
{'ALARM_CLOCK_BAKELIKE_RED'}	{'ALARM_CLOCK_BAKELIKE_GREEN'}	0.038650	0.035136	0.023118	0.598155	17.024025	0.021760	2.401088

Fig40: association rules

To create a frequently occurring itemset, Apriori principles were applied. It was discovered that a minimal support criterion of 1% and a minimum confidence level of 40% produce enough rules after a few cycles.

II) INTERPRETING ASSOCIATION RULES

A rule that has high support and high confidence

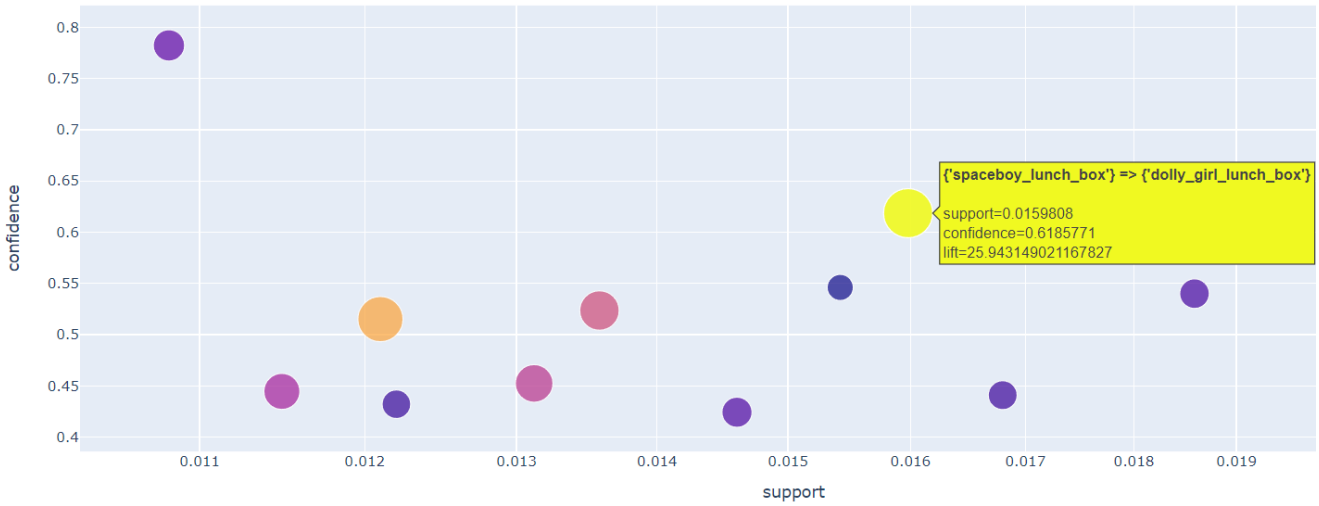

Fig41: rule that has high support and high confidence

A rule like that wouldn't even be interesting from a subjective standpoint because they happen frequently and are likely to be evident.

A rule that has reasonably high support but low confidence

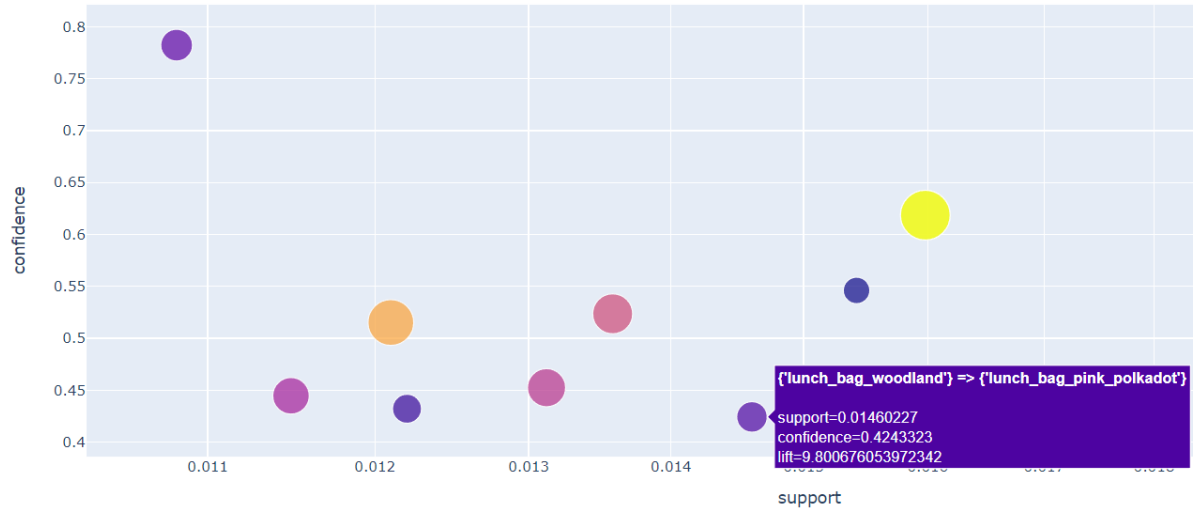

Fig42: rule that has reasonably high support but low confidence

Given that an itemset may regularly recur across transactions yet not be related to other itemsets, such a rule is likely to be uninteresting.

A rule that has low support and low confidence

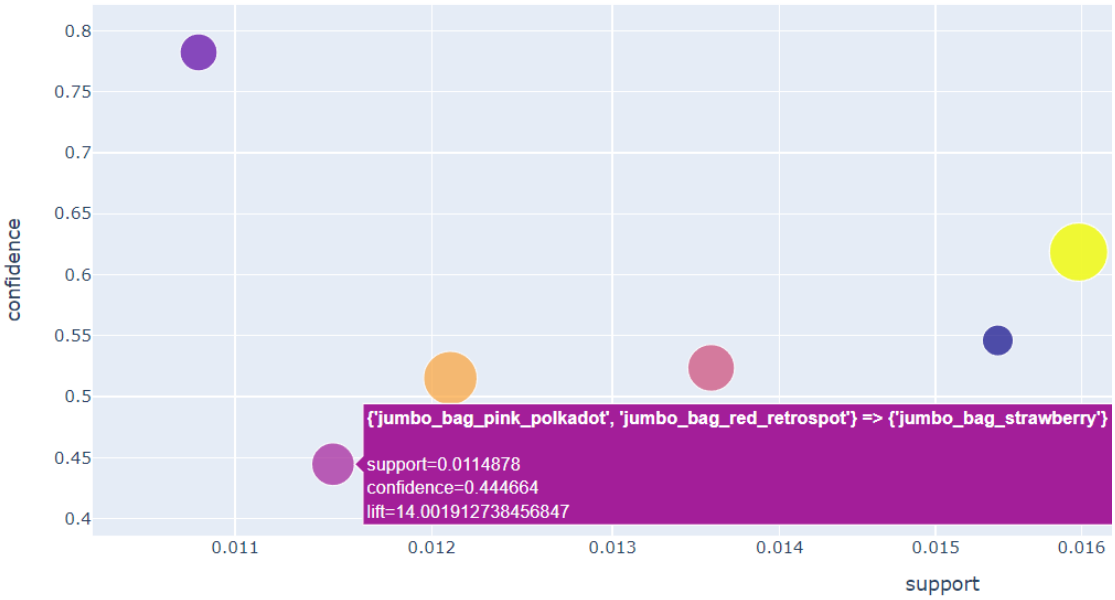

Fig43: rule that has low support and low confidence

Due to the itemset's rarity and need for connections, such rules lack appeal and value.

A rule that has low support and high confidence

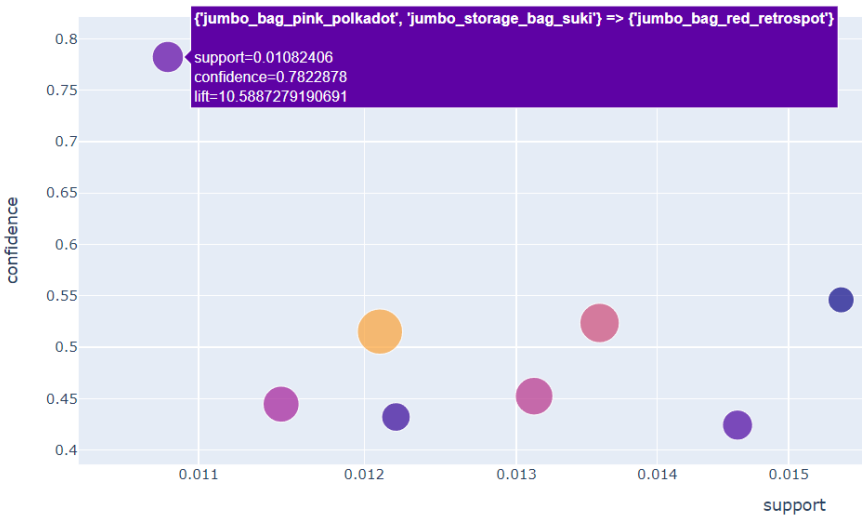

Fig44: rule that has low support and high confidence

These rules have the potential to be intriguing since they are uncommon but also have a close connection to other itemsets.

Finally, exploring the network graph below will help better to grasp the connections between all of the products:

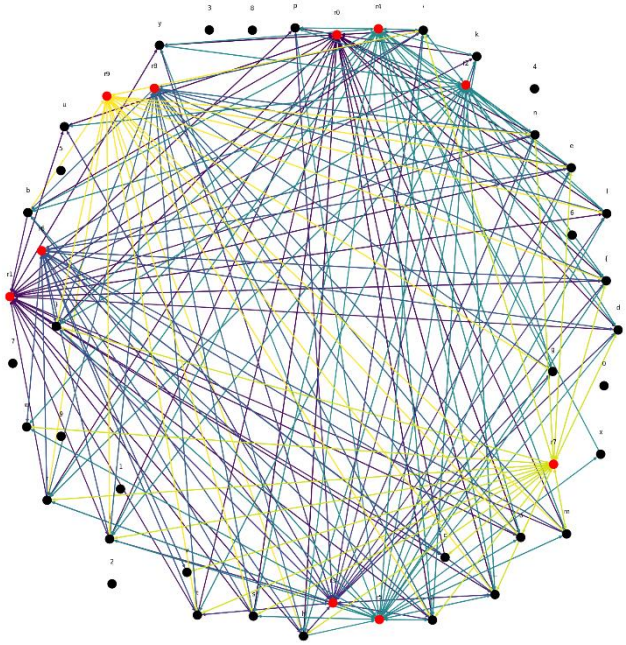

Fig45: network graph

CONCLUSIONS

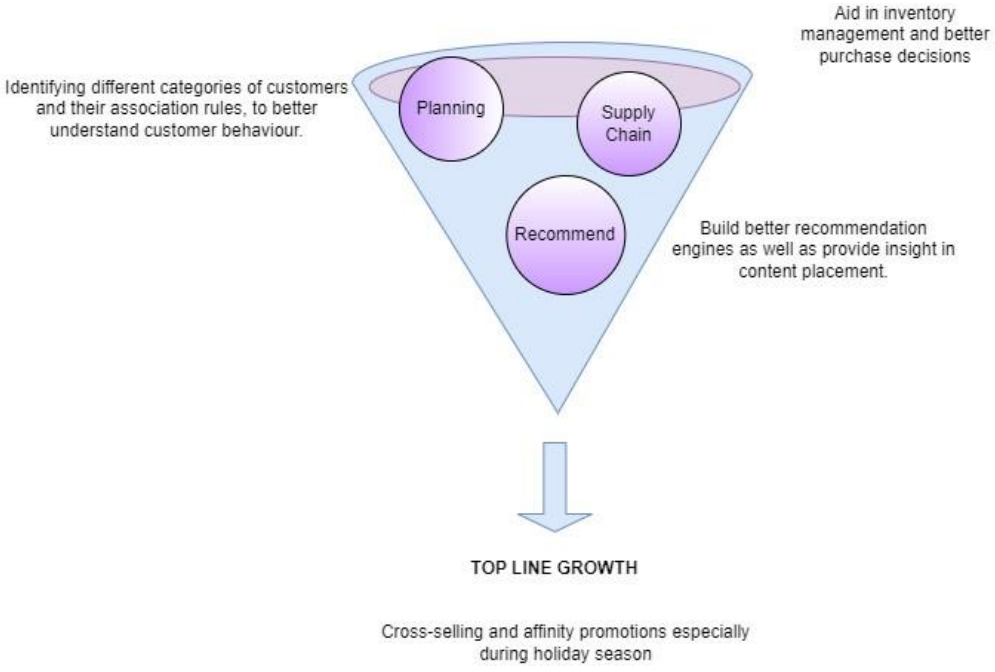

Fig 46: Conclusion of market basket analysis

- There are more than only association rules, as can be seen from the summary. Following the Gephi network graph's interest, it is evident that several elements are shared as antecedents and consequents. The rules that are more pertinent to the particular business purpose may be filtered using this network knowledge and the scatter plot produced by Plotly.
- In inventory management and cross-selling, the fundamental analysis may be used to achieve the business goal at both the bottom-line and top-line levels. 'Regency cake stand 3 tyre' and 'green regency tea cup and saucer' are two products that illustrate this.
- Additionally, we can see which product variants are most popular with UK customers based on all the association rules accessible for a particular product type out of all the available product variations.

5. RESULTS

Prior to conducting the analysis, a thorough pre-processing step was performed on the dataset. This involved addressing data quality issues such as duplicates, missing values, and anomalies. By ensuring the dataset's reliability and accuracy, the foundation for accurate and meaningful analysis was laid. Cluster analysis was conducted using multiple techniques to identify distinct groups of customers based on their purchasing behavior. Three different methods, namely K-means, Expectation Maximization (EM), and Ward's Minimum Variance, were employed. The optimal number of clusters was determined using various metrics, resulting in a 3-cluster solution for K-means and EM methods, and a 4-cluster solution for Ward's Minimum Variance. These clusters provided valuable insights into customer segmentation and formed the basis for targeted marketing efforts and resource allocation.

To predict customer churn, a predictive analysis was performed using machine learning algorithms. Features such as customer demographics, purchase history, and RFM scores were utilized to train and evaluate the models. Several algorithms, including support vector machines, naïve bayes classifier and ensemble learning methods, were tested to identify the most accurate model for predicting customer churn. The SVC is nearly flawless on the

training and test sets and is the best model. However, the Bagging Ensemble, Voting Ensemble, and Random Forest have excellent performances and high accuracy.

Market basket analysis was employed to uncover associations and patterns among purchased products. The Apriori algorithm was utilized to identify frequent item sets and generate association rules. These rules provided insights into product combinations that were frequently purchased together, enabling businesses to optimize product placement, cross-selling, and promotional strategies.

6. CONCLUSION

In conclusion, this project employed a robust methodology of pre-processing, cluster analysis, predictive modeling, and market basket analysis to gain valuable insights into customer behavior and preferences. The results provide significant implications for businesses aiming to enhance their marketing strategies, improve customer retention, and drive revenue growth.

The dataset was carefully cleaned through the pre-processing stage, ensuring data quality and reliability. This step was crucial in preparing the data for subsequent analyses, enabling accurate and meaningful interpretations. Cluster analysis revealed distinct customer segments based on purchasing behavior. By employing various techniques such as K-means, Expectation Maximization (EM), and Ward's Minimum Variance, meaningful patterns and groupings among customers were identified. These clusters offer valuable insights for businesses to tailor their marketing efforts, target specific customer segments, and customize their offerings to enhance customer satisfaction and loyalty. The predictive modeling aspect of the project focused on forecasting customer churn. By utilizing advanced machine learning algorithms and incorporating relevant features such as customer demographics, purchase history, and RFM scores, a robust predictive model was developed. This model can effectively identify customers at risk of churn, allowing businesses to implement proactive retention strategies and minimize revenue loss. Market basket analysis uncovered associations and patterns among purchased products. The application of the Apriori algorithm helped identify frequent itemsets and association rules, providing valuable insights into product combinations that are commonly purchased together. This information can guide businesses in optimizing product placement, cross-selling, and promotional strategies to increase sales and customer satisfaction.

Looking to the future, this project has several avenues for further exploration and improvement. The incorporation of additional data sources, such as social media or demographic data, could provide a deeper understanding of customer preferences and behavior. Exploring more advanced machine learning techniques and evaluating their effectiveness in predicting customer churn could lead to more accurate models.

7. References

Deepali Kamthania, A. P. (2018). Market Segmentation Analysis and Visualization Using K-Mode Clustering .

Dolnicar, S. (2002). A Review of Unquestioned Standards in Using Cluster Analysis for Data-Driven .

Iromi R Paranavithana, M. I. (2021). Unsupervised Learning and Market Basket Analysis in Market Segmentation.

Lim, T. P. (2019, Jan 7). *The Most Important Data Science Tool for Market and Customer Segmentation.* Retrieved from Medium: https://towardsdatascience.com/the-most-important-data-science-tool-for-market-and-customer-segmentation-c9709ca0b64a

Loraine Charlet Annie M.C., A. K. (2012). Market Basket Analysis for a Supermarket based on Frequent Itemset Mining.

Manpreet Kaur, S. K. (2016). Market Basket Analysis: Identify the changing trends of market data using association rule mining.

Ms. G. Nathiya, M. S. (2010). An Analytical Study on Behavior of Clusters Using K Means, EM and K* Means Algorithm.

Muthalagu, S. J. (2020). Market segmentation for profit maximization using machine learning algorithms.

Qiankun Zhao, S. S. (2003). Association Rule Mining: A Survey.

Savi Gupta, R. M. (2014). A Survey on Association Rule Mining in Market Basket Analysis.

Thomas, J. W. (2019). Market Segmentation.