Customer Segmentation

CSCI 6443: Data Mining

Fall 2023

Instructor: Dr. Abdelghani Bellaachia

Term Project Report

Jadhav, Shubham
G30570862
Department of Computer Science
School of Engineering
George Washington University

Ramagiri, Jahnavi
G46015075
Department of Computer Science
School of Engineering
George Washington University

GitHub Repository: https://github.com/shubhjadhav/Customer-Segmentation

December 07, 2023

Contents

List of Figures

List of Tables

1 Introduction

In today's intensely competitive business landscape, existing enterprises face the imperative of implementing effective marketing strategies to survive amidst cutthroat competition and the continuous emergence of new businesses. Adaptation or risk obsolescence has become a fundamental principle in contemporary marketing. With the expanding customer base, catering to diverse needs has become a formidable challenge for companies. This is where data mining assumes a pivotal role by uncovering concealed patterns within a company's database. Customer segmentation, a data mining application, plays a crucial role in categorizing customers with similar patterns into clusters, simplifying the management of a vast customer base. This segmentation directly or indirectly influences marketing strategies by unveiling new avenues, such as identifying suitable products for specific segments, tailoring marketing plans accordingly, offering targeted discounts, and deciphering previously unknown customer-object relationships. Customer segmentation empowers companies to understand customer purchasing behavior, leading to improved service and satisfaction. It also aids in identifying target customers and refining marketing tactics to enhance revenue generation.

2 Problem Statement

Fresh Mart's CEO has sought Jahnavi and Shubham's Data Mining expertise to leverage their substantial customer data repository. With a comprehensive collection of transactional invoices, Fresh Mart aims to discern customer behavior, optimizing it for two strategic objectives. Firstly, the company aims to segment its customer base to tailor promotions and incentives effectively. Secondly, leveraging insights from this segmentation, Fresh Mart plans to predict the purchasing patterns of new and first-time customers.

Through this initiative, Fresh Mart anticipates improving customer engagement and strategic decision- making, fostering enhanced business performance and customer satisfaction.

3 Key Considerations

Customer Segmentation: The project focuses on segmenting customers based on their behaviors and preferences, which is crucial for targeted marketing strategies and personalized customer experiences.

Product Categories: Understanding the product categories and their clustering is vital for inventory management, product recommendations, and marketing strategies.

Customer Classification: By leveraging customer segmentation, businesses can categorize new customers based on a deeper understanding of their purchasing behavior.

4 Literature Survey

1. **Business Rule**-based customer segmentation involves creating segments by applying predefined rules or criteria based on specific business considerations. These rules are often derived from domain expertise, marketing strategies, or key performance indicators (KPIs). The segmentation process relies on established guidelines to categorize customers into distinct groups.

2. **Quantile Membership**: Divides customers into groups based on Recency, Frequency, and Monetary data.

3. **Supervised clustering with decision tree-based** customer segmentation combines the interpretability of decision trees with the flexibility of clustering. Using predefined labels or target variables, this method leverages decision tree models to guide the creation of meaningful customer segments aligned with specific business goals [2].

4. **Unsupervised Clustering** cluster-based customer segmentation involves employing clustering algorithms on customer data without predefined labels or target variables. The objective is to identify inherent patterns and group customers based on similarities in their features or behaviors. Common clustering algorithms for this purpose include K-means, hierarchical clustering, and DBSCAN [3].

5. **A-priori** based customer segmentation refers to the utilization of the Apriori algorithm, a classic algorithm in association rule mining, for the purpose of categorizing customers. In this context, the algorithm analyzes transaction data to identify frequent itemsets, representing sets of items often purchased together. These itemsets are then used to create associations and patterns, facilitating the segmentation of customers based on their common purchasing behaviors [4].

6. **Post-hoc** customer segmentation involves categorizing customers based on their behaviors or characteristics after data has been collected, contrasting with a priori methods where segmentation criteria are established before data collection. In post-hoc segmentation, advanced analytics, clustering algorithms, or statistical techniques are often applied to identify patterns or groups within the customer data retrospectively. This method allows businesses to discover naturally occurring segments and uncover insights that were not predefined [4].

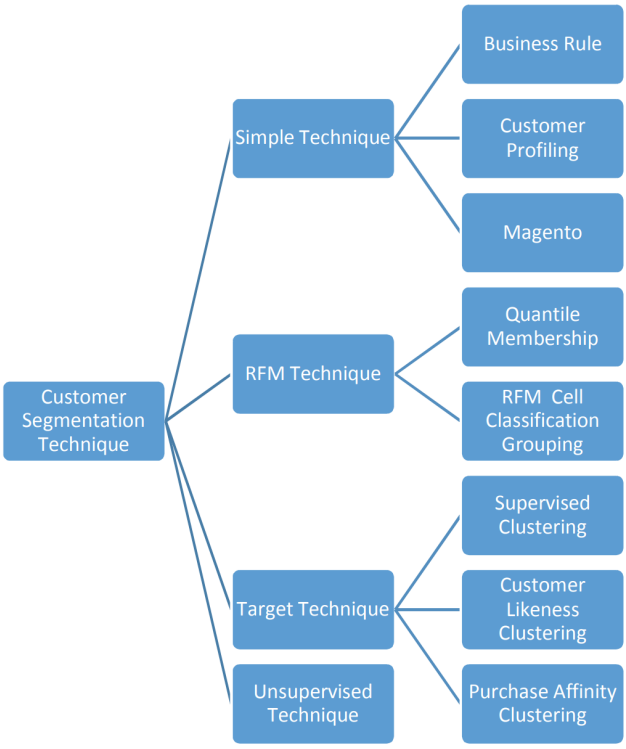

Figure 1: Customer Segmentation Classification [1]

5 Approach

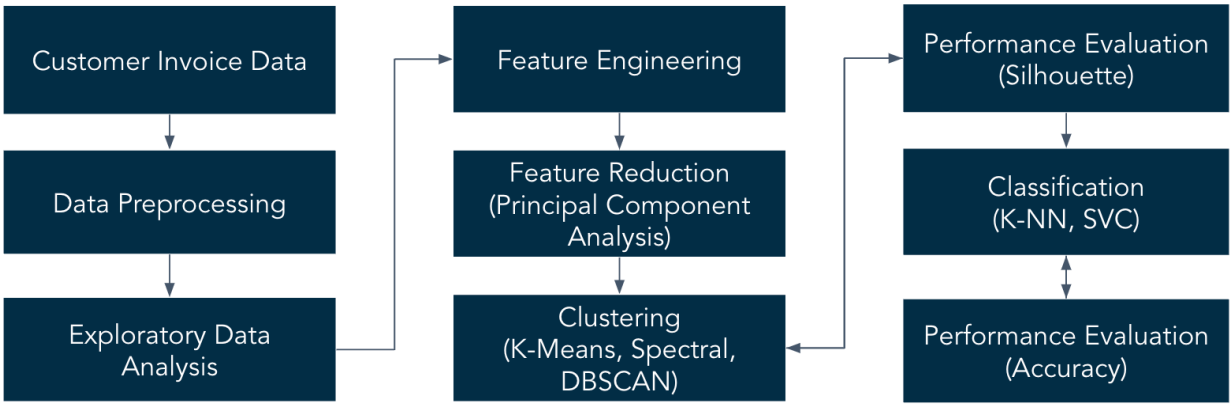

Figure 2: Project Approach

6 Data Description

1. **Invoice**: A 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.

2. **StockCode**: A 5-digit integral number uniquely assigned to each distinct product.

3. **Description**: Product (item) name.
4. **Quantity**: The quantities of each product (item) per transaction.
5. **InvoiceDate**: The day and time when each transaction was generated.
6. **UnitPrice**: Product price per unit in sterling.
7. **CustomerID**: A 5-digit integral number uniquely assigned to each customer.
8. **Country**: Name of the country where each customer resides.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
5	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	2010-12-01 08:26:00	7.65	17850	United Kingdom
6	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	2010-12-01 08:26:00	4.25	17850	United Kingdom
7	536366	22633	HAND WARMER UNION JACK	6	2010-12-01 08:28:00	1.85	17850	United Kingdom
8	536366	22632	HAND WARMER RED POLKA DOT	6	2010-12-01 08:28:00	1.85	17850	United Kingdom
9	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	2010-12-01 08:34:00	1.69	13047	United Kingdom

Table 1: Sample raw data

7 Exploratory Data Analysis

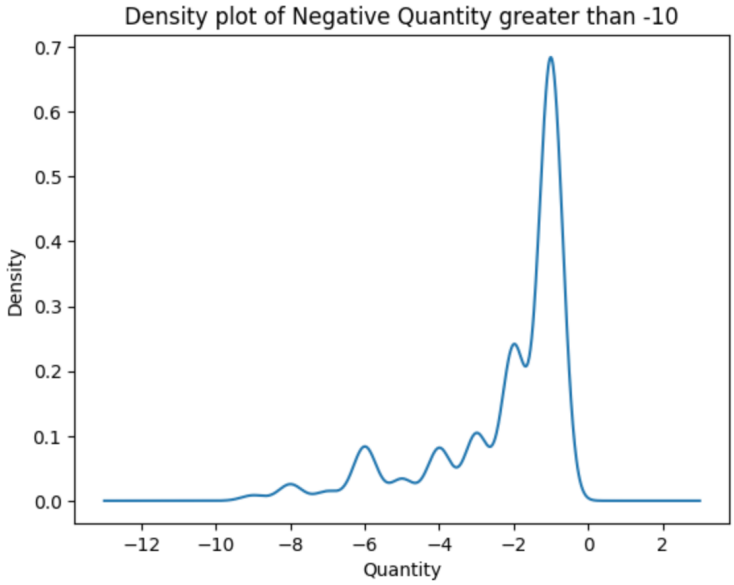

Figure 3: Density plot of all transaction quantity

Observation: By restricting negative quantity to greater than -10, we can observe the variation of unit price.

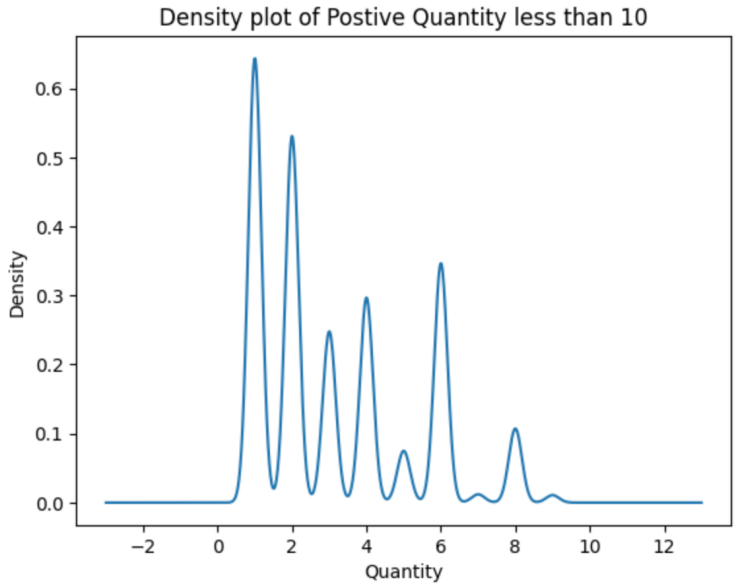

Figure 4: Density plot of all transaction negative quantity less than 10

Observation: We can observe that most of the quantity in the transaction data has a huge variation.

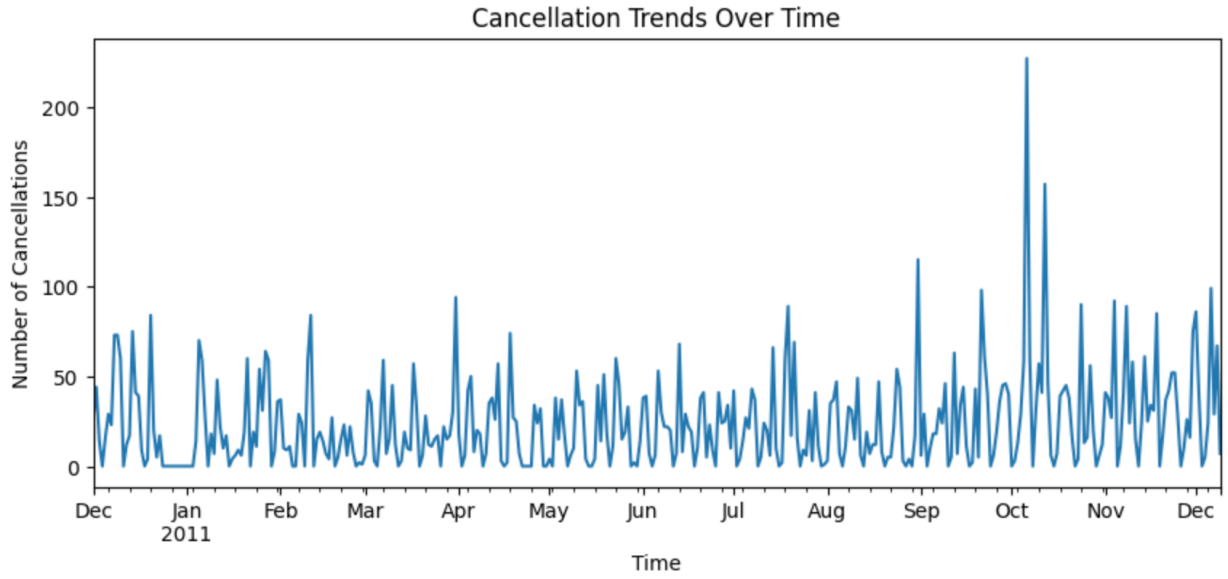

Figure 5: Cancellation trend over time

Observation: We can observe from the time series plot of cancellation that most of the cancellations are below 50 with periodic spikes.

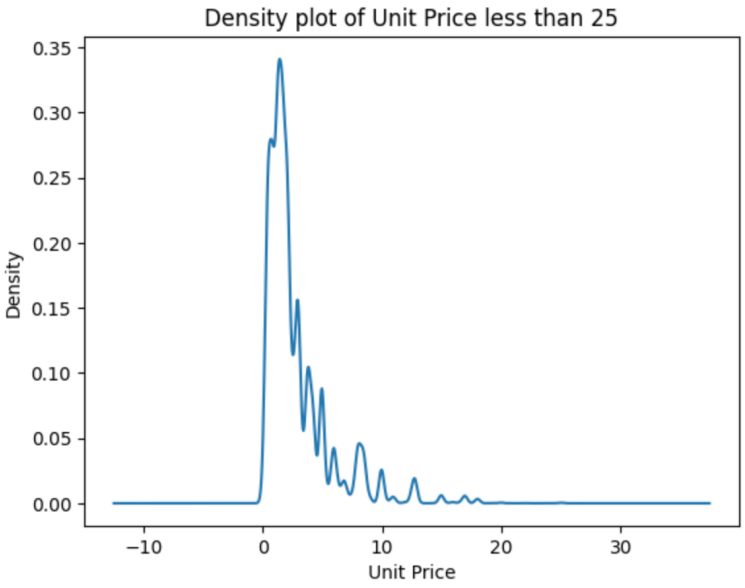

Figure 6: Density plot of unit price less than 25

Observation: By restricting the unit price to be less than 25, we can observe the variation of unit price.

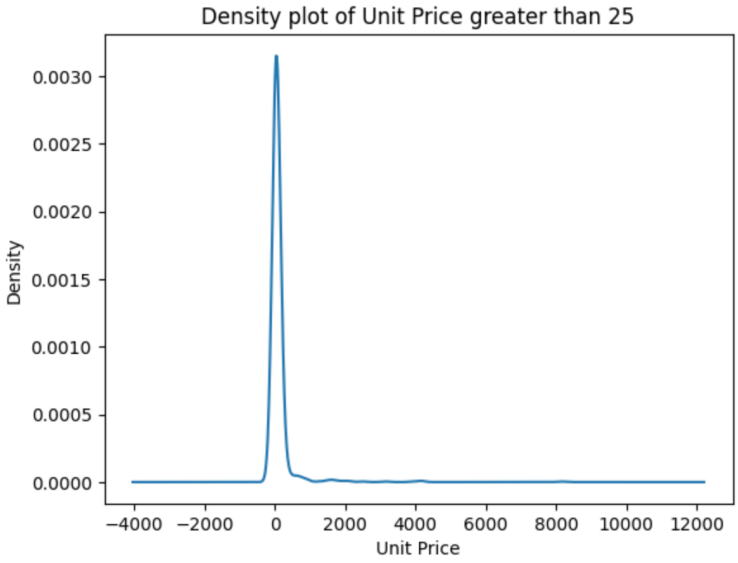

Figure 7: Density plot of unit price greater than 25

Observation: By restricting the unit price to be greater than 25, we can observe that there is no variation of unit price data and hence we can conclude that most of the data has unit price value less than 25.

8 Data Cleaning

We performed multiple data cleaning steps to ensure that the data is pure and analysis is not affected by the impure data. The steps taken to clean data is as follows:

1. **Duplicate Records**: There were 5268 duplicate observations and we dropped them from the data set.
2. **Missing Observations**: There were no observations with all values as NULL.
3. **Missing Values**: There were 1454 and 135080 observation with missing descriptions and customer id respectively. After we drop missing customer id, the null descriptions does not exist.
4. **Negative Values**: There were 10624 observations that have negative unit prices values. To remove these negative unit price, we checked if we have a existing record from the same customer with similar values but positive unit price. That is we wanted to check if the cancellations are for a transaction of purchase. Then we removed all the records with negative unit price along with matched postive purchase records.

9 Feature Engineering

9.1 Basket Price

Basket price is defined as total price of a customer for each invoice.

$$BasketPrice = Sum\ per\ Invoice\ (UnitPrice * Quantity)$$

Futher, we created multiple features from BasketPrice variable by aggregating the values for each customer as follows:

1. Count of Baskets
2. Minimum Basket price for each Customer
3. Maximum Basket price for each Customer
4. Average Basket price for each Customer
5. Total Basket price for each Customer

9.2 Time slots

We have transaction time in format: DD:MM:YY HH:MM. We segmented this data into 4 categories as follows:

1. Morning (6AM - 12PM)

2. Afternoon (12PM - 6PM)
3. Evening (6PM - 12AM)
4. Night (12AM - 6AM)

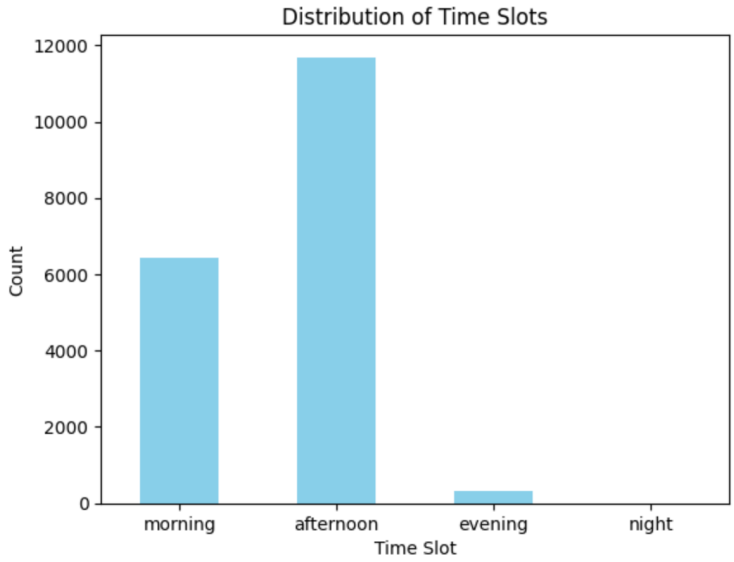

Figure 8: Histogram plot of Time Slot category

Observation: We can observe that most of the transactions are in afternoon slot followed by morning. There are no transactions in night which is logical.

9.3 Product Categories

The objective of product categorization is to establish meaningful product categories by analyzing product descriptions. Each transaction within the dataset is associated with a stock code, and each stock code is linked to a description. These descriptions offer valuable insights into the nature of the product, making them instrumental in the categorization process. To enhance the quality of the data, we employ basic preprocessing techniques to clean the descriptions. Subsequently, we utilize a count vectorizer to convert the text data into a numerical format. The vectorized input is then given to K-means clustering.

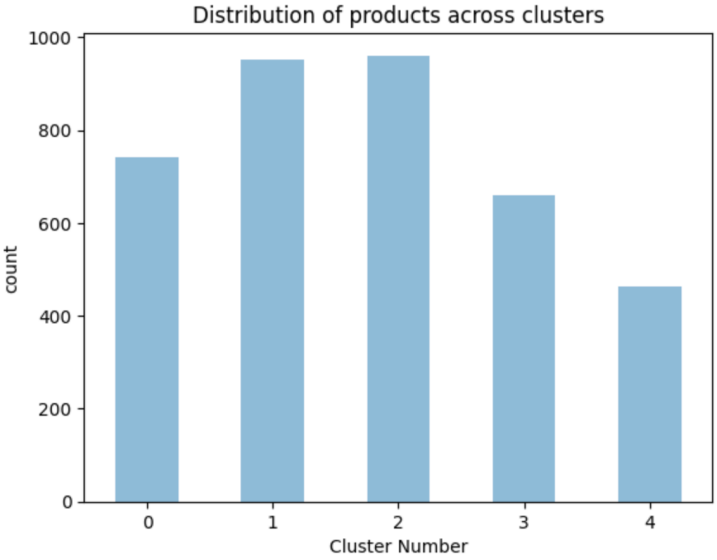

Figure 9: Histogram plot for distribution of products across categories

Observation: The resulting outcome reveals products organized into five distinct clusters. The distribution of products across these clusters varies, with each cluster containing between 600 and 1000 elements. Word clouds visually represent the most frequently occurring words within each cluster, providing additional insights into the characteristics of products within each category.

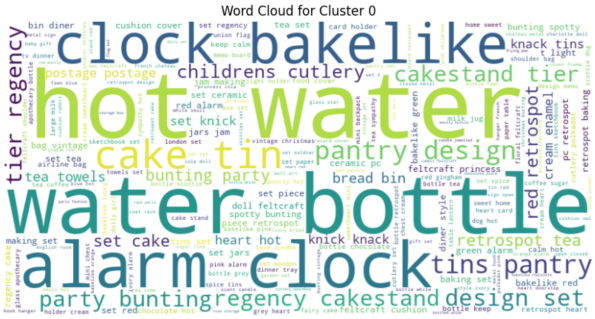

Figure 10: Word Cloud of Product category 0

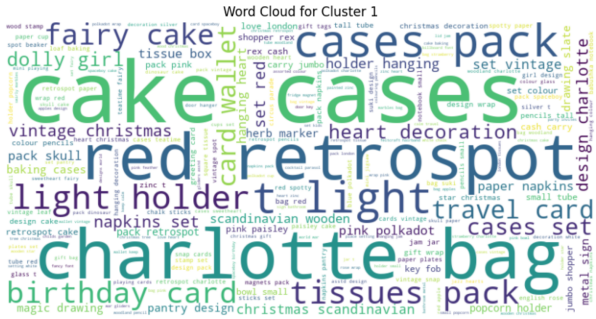

Figure 11: Word Cloud of Product category 1

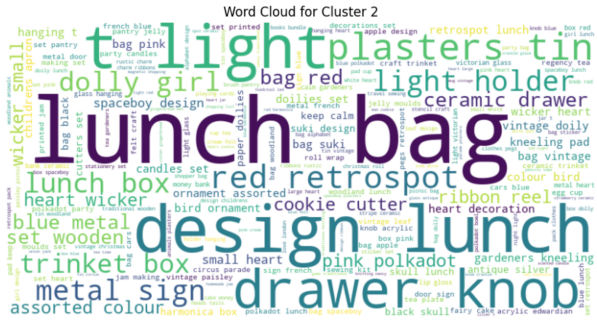

Figure 12: Word Cloud of Product category 2

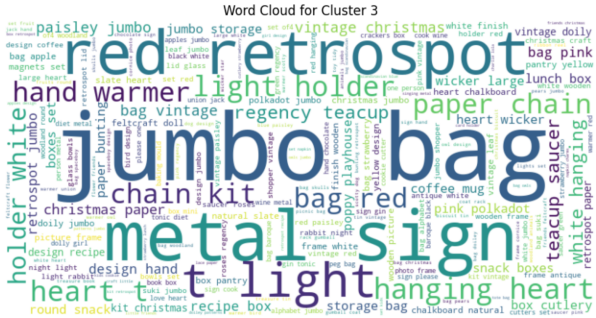

Figure 13: Word Cloud of Product category 3

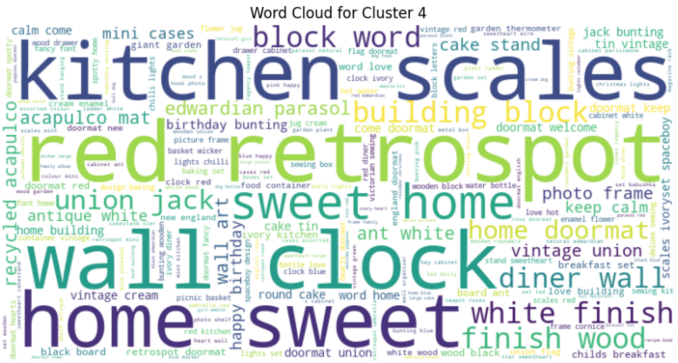

Figure 14: Word Cloud of Product category 4

9.4 Recency

Recency score is calculated based on a customer's last purchase date. Here we calculate last purchase date for each customer and find difference in days from the last transaction date in the entire dataset.

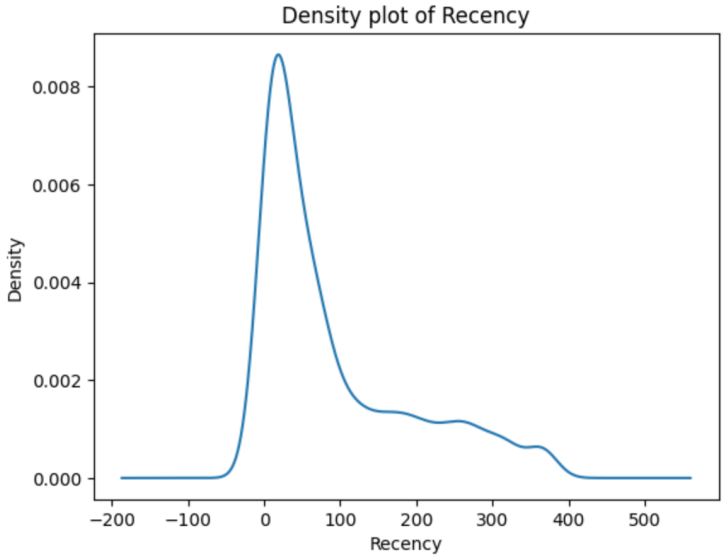

Figure 15: Histogram plot of Recency across customers

Observation: We can observe that most of the customer's recent transaction days are within 100 days of last transaction date. We can also see that there are few customers who have not made any transaction for entire year.

9.5 Cancellation

Some of the customers have had cancellations which accounts to 2% of the entire transaction data. We constructed three metrics from the cancellations records as follows:

1. Number of Cancellations
2. Total cancellation amount
3. Revenue Lost Ratio - Total cancellation amount by total revenue from customer

10 Dimensionality Reduction

Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional form by identifying and prioritizing the principal components, which capture the most significant variance in the original dataset. This process simplifies data representation while retaining essential patterns, facilitating improved analysis, visualization, and computational efficiency.

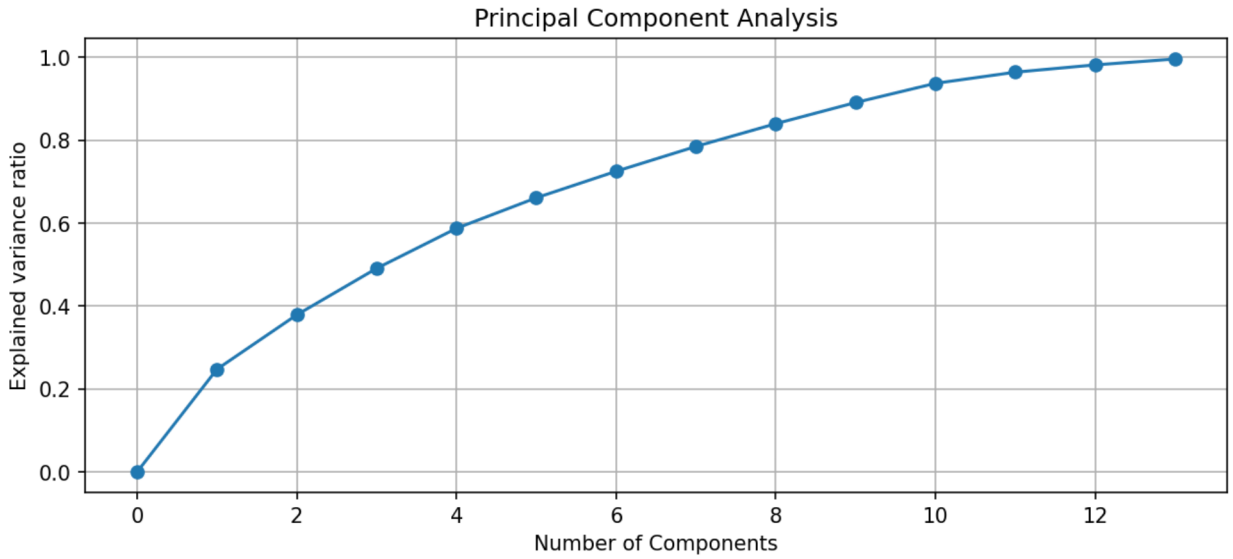

Figure 16: Plot of Explained Variance Ratio across PCA components

Observation: Figure 16 shows the plot of Explained variance ratio across PCA components ranging from 0 to 12 components. The ratio explains the information captured from the corresponding components. Here we can observe that with 8 PCA components we can extract around 80% information.

11 Clustering

11.1 Grid Search

GridSearch, short for Grid Search Cross-Validation, is a hyperparameter tuning technique widely used in machine learning. It systematically searches through a predefined grid of hyperparameter values for a given model and evaluates the model's performance using cross-validation. The goal is to find the optimal combination of hyperparameters that maximizes the model's performance on a specific metric, such as accuracy or precision.

11.2 K-Means Clustering

K-Means Clustering is a popular unsupervised machine learning algorithm used for partitioning a dataset into distinct, non-overlapping groups or clusters. The algorithm iteratively

assigns data points to clusters based on the similarity of features and updates cluster centroids until convergence. The term "K" refers to the predetermined number of clusters in which the data is to be grouped.

11.2.1 All Features

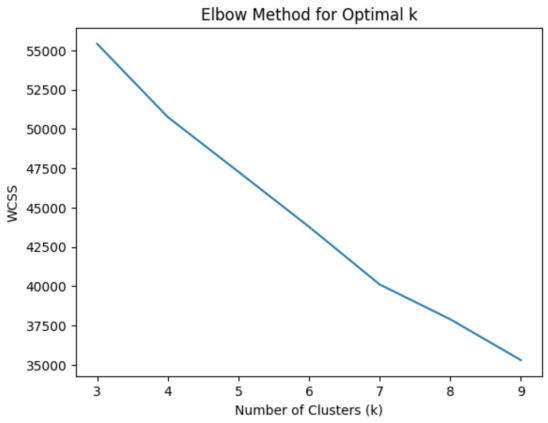

Figure 17: Plot from K-Means GridSearch results

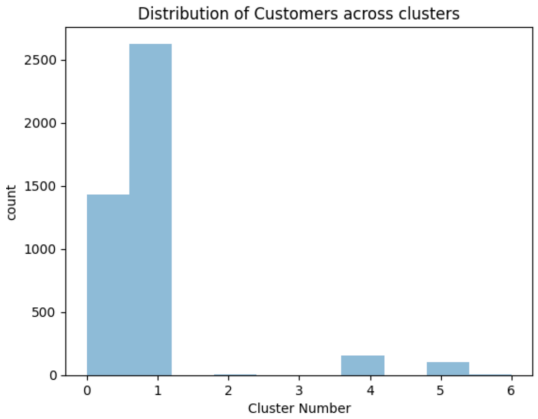

Figure 18: Histogram plot for distribution of customer categories

Observation: In this experiment K-Means GridSearch is performed on the entire dataset, with Figure 17 illustrating the corresponding Within-Cluster Sum of Squares (WCSS) scores. Figure 18 displays a histogram presenting the distribution of customer categories. Notably, the majority of customers are concentrated in two clusters, while the remaining clusters exhibit minimal to zero customer representation. The optimal number of clusters (K) identified in this analysis is 7.

11.2.2 Basket Price based Features

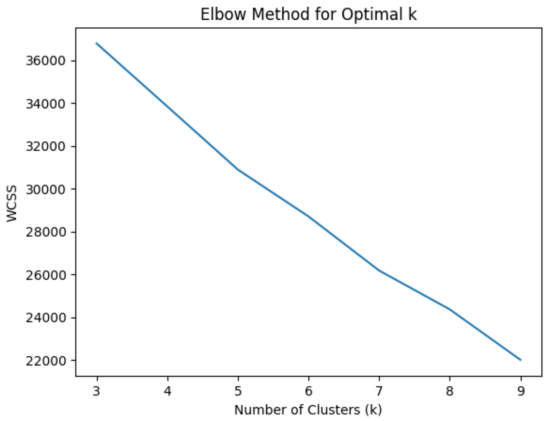

Figure 19: Plot from K-Means GridSearch results

Figure 20: Histogram plot for distribution of customer categories

Observation: In this experiment K-Means GridSearch is executed specifically on features derived from Basket Price (BP), namely BPMax, BPmin, BPmean, BPTotal, C0, C1, C2,

and C3. Figure 19 visually represents the associated Within-Cluster Sum of Squares (WCSS) scores, while Figure 20 provides a histogram illustrating the distribution of customer categories. Notably, customers are observed to be distributed across clusters 0, 2, 5, and 6, with minimal customer representation in the remaining three clusters.

11.2.3 Time based Features

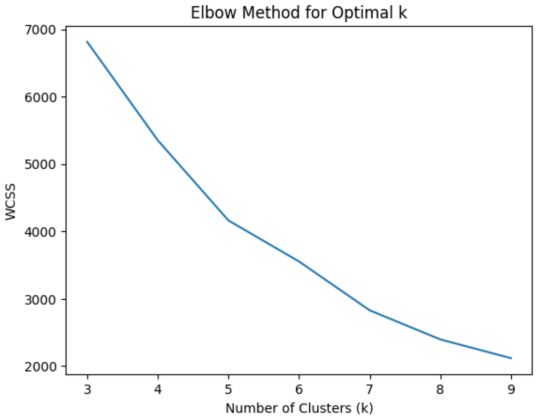

Figure 21: Plot from K-Means GridSearch results

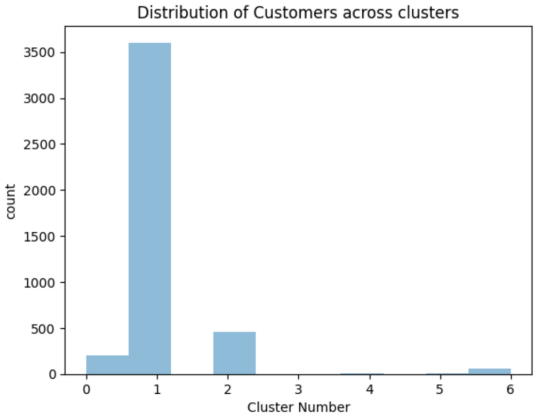

Figure 22: Histogram plot for distribution of customer categories

Observation: a K-Means GridSearch is specifically conducted on features derived from Time-Based Features. Figure 21 presents the Within-Cluster Sum of Squares (WCSS) scores resulting from this analysis, and Figure 22 offers a histogram depicting the distribution of customer categories. Notably, the distribution of customer categories appears highly skewed, with all customers predominantly falling into the same cluster.

11.2.4 Features from Pearson's Correlation Analysis

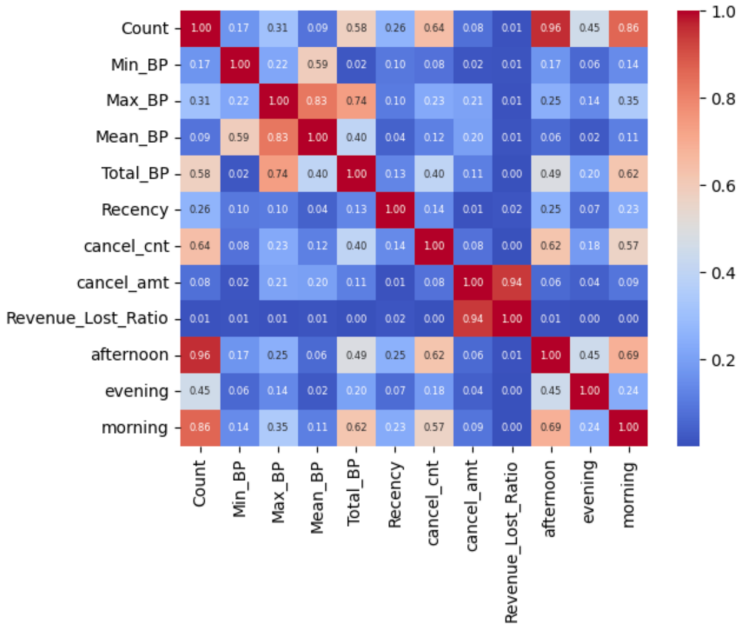

Figure 23: Heat Plot of Correlation Matrix of All features

Observation: Figure 23 displays a Heatmap of Pearson's Correlation Matrix for the dataset features, illustrating the linear relationships between pairs of variables. The analysis reveals generally low correlation among input features, signifying their independence. Notably, 'Feature count' demonstrates a strong correlation with 'afternoon' and 'evening,' while the Basket Price features exhibit notable correlation within themselves. Employing a threshold of 0.7, features surpassing this correlation limit are considered for elimination to mitigate multicollinearity concerns.

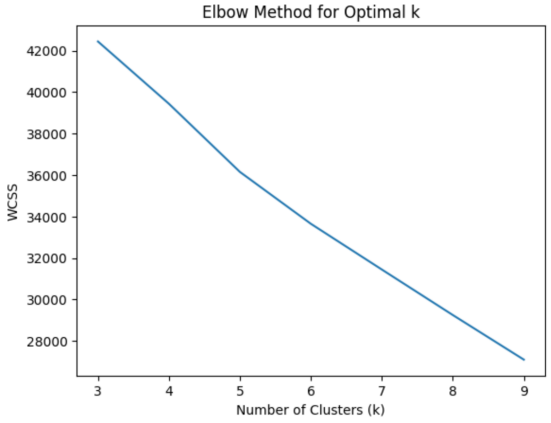

Figure 24: Plot from K-Means GridSearch results

Figure 25: Histogram plot for distribution of customer categories

Observation: In this experiment K-Means GridSearch is executed specifically on features derived from Collinearity Analysis. Figure 24 visually represents the associated Within-Cluster Sum of Squares (WCSS) scores, while Figure 25 provides a histogram illustrating

the distribution of customer categories. Notably, customers are observed to be distributed across diverse clusters showing a good distribution.

11.2.5 PCA components

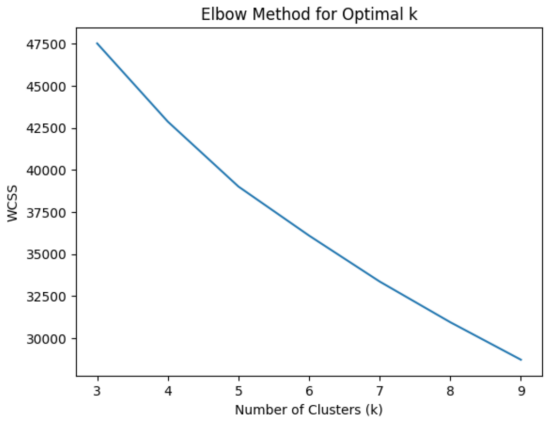

Figure 26: Plot from K-Means GridSearch results

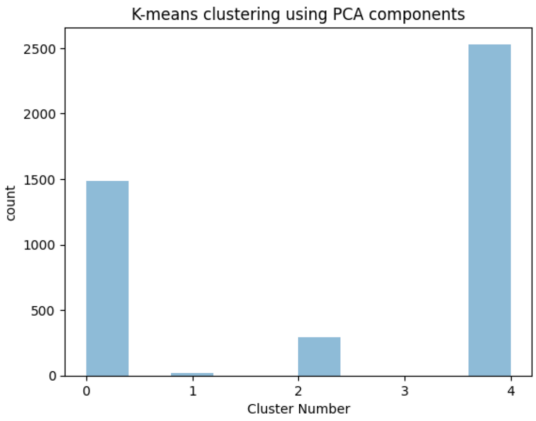

Figure 27: Histogram plot for distribution of customer categories

Observation: In this experiment K-Means GridSearch is executed specifically on features derived from Principal Component Analysis on 12 componenets. Figure 26 visually represents the associated Within-Cluster Sum of Squares (WCSS) scores, while Figure 27 provides a histogram illustrating the distribution of customer categories. Notably, customers are observed to be distributed across clusters 0, 2, 4, with minimal customer representation in the remaining 2 clusters.

11.3 Spectral Clustering

Spectral clustering is a versatile and effective algorithm used for partitioning data into distinct groups based on similarity. By leveraging the eigenvalues and eigenvectors of a similarity graph constructed from the data, spectral clustering uncovers inherent structures, making it particularly suited for non-linear and complex datasets.

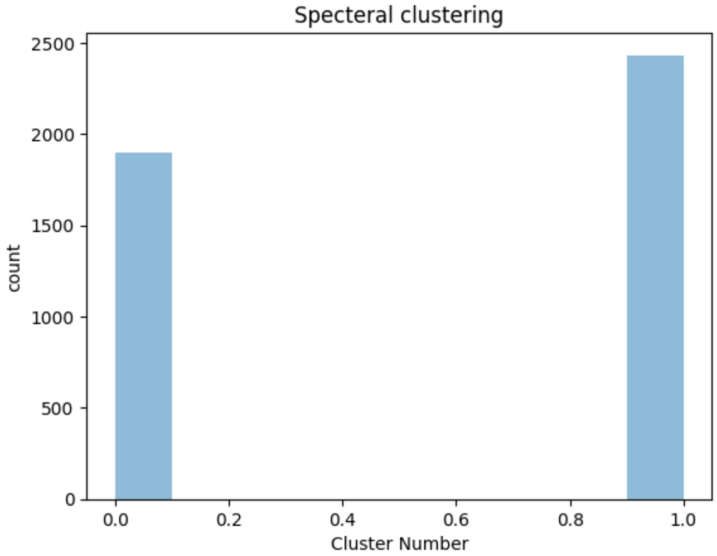

Figure 28: Histogram plot for distribution of customer categories

Observation: Figure 28 illustrates the outcomes of spectral clustering applied to the dataset features, revealing the partitioning of customers into two distinct clusters with an equal distribution. Spectral clustering has successfully segmented the customer population into two groups. The balanced distribution among the clusters indicates a relatively even assignment of customers to each group, suggesting that spectral clustering has effectively captured inherent patterns or structures within the dataset.

11.4 Density-Based Spatial Clustering of Applications with Noise

DBSCAN is a clustering algorithm commonly used in data analysis. Unlike traditional methods that assume clusters have a specific shape, DBSCAN identifies clusters based on dense regions in the data space. It works by defining neighborhoods around data points and grouping them into clusters if they have sufficient density. DBSCAN is effective in discovering clusters of arbitrary shapes and is particularly robust in handling noise and outliers. It's a valuable tool for uncovering hidden structures in data, especially when the shapes and sizes of clusters are not known beforehand.

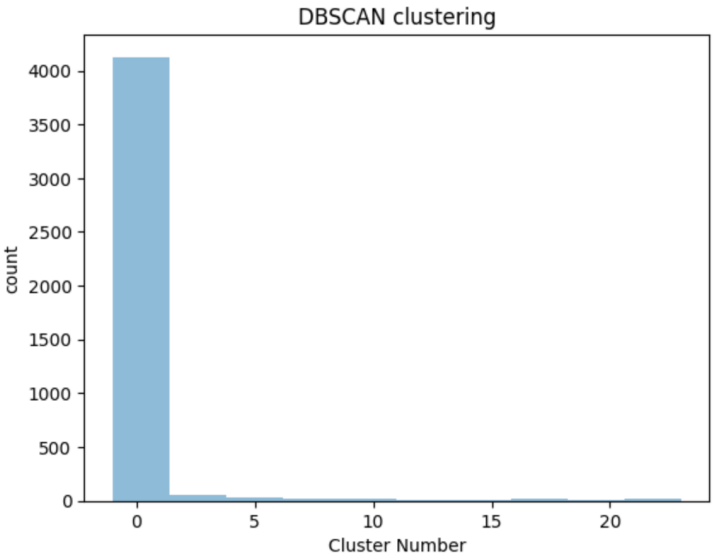

Figure 29: Histogram plot for distribution of customer categories

Observation: Figure 28 showcases the results of DBSCAN (Density-Based Spatial Clustering of Applications with Noise) on the dataset features, revealing that all customers are predominantly grouped into a single cluster. Unlike some clustering algorithms that aim for a predefined number of clusters, DBSCAN identifies dense regions in the data and assigns data points to clusters based on density connectivity. The observation of a skewed distribution implies that DBSCAN has identified a dominant cluster encompassing the majority of customers, potentially indicating a clear and dense pattern within the dataset.

12 Classification

12.1 K-Nearest Neighbours

K-Nearest Neighbors (KNN) is a simple and intuitive algorithm in machine learning used for classification and regression tasks. It operates on the principle of proximity, where an unknown data point is assigned the label or value of the majority of its k nearest neighbors in the feature space. KNN is non-parametric, meaning it doesn't make assumptions about the underlying data distribution, and it is instance-based, as it stores the entire training dataset for predictions. The choice of the parameter 'k' determines the number of neighbors considered during the classification or regression process. KNN is widely used for its simplicity and effectiveness in various applications.

```
Best Parameters: {'n_neighbors': 1}
Training Accuracy with Best Parameters: 100.00%
Validation Accuracy with Best Parameters: 58.89%
Classification Report:
              precision    recall  f1-score   support

           0       0.55      0.60      0.57       207
           1       0.00      0.00      0.00         2
           2       0.64      0.70      0.66       361
           3       0.85      0.81      0.83        21
           4       0.83      0.83      0.83         6
           5       0.47      0.39      0.43       115
           6       0.55      0.44      0.49       154

    accuracy                           0.59       866
   macro avg       0.56      0.54      0.54       866
weighted avg       0.58      0.59      0.58       866
```

Figure 30: Classification results for KNN using GridSearch

Observation: Figure 30 presents the classification results obtained from K-Nearest Neighbors (KNN) using Gridsearch. The optimal model, determined to be the one with a single neighbor, demonstrates perfect training accuracy at 100%. However, the validation accuracy, recorded at 58.99%, suggests a substantial drop in performance when applied to unseen data, indicating a clear case of overfitting.

12.2 Support Vector Classifier

Support Vector Classification (SVC) is a supervised machine learning algorithm used for binary and multiclass classification tasks. As part of the support vector machines (SVM) family, SVC works by finding the optimal hyperplane that separates different classes in the feature space. It focuses on identifying the support vectors, which are the data points closest to the decision boundary. SVC aims to maximize the margin between classes, enhancing the model's generalization to new, unseen data. It is particularly effective in scenarios with non-linear decision boundaries through the use of kernel functions. Overall, SVC is a powerful and widely applied algorithm in classification tasks due to its robust performance and flexibility.

```
Best Parameters: {'C': 100}
Training Accuracy with Best Parameters: 73.07%
Validation Accuracy with Best Parameters: 74.60%
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.51      0.67       207
           1       0.00      0.00      0.00         2
           2       0.68      0.99      0.80       361
           3       0.91      0.48      0.62        21
           4       0.00      0.00      0.00         6
           5       0.68      0.92      0.78       115
           6       1.00      0.45      0.62       154

    accuracy                           0.75       866
   macro avg       0.61      0.48      0.50       866
weighted avg       0.81      0.75      0.72       866
```

Figure 31: Classification results for SVC using GridSearch

Observation: Figure 30 showcases the classification outcomes derived from the Support Vector Classifier (SVC) using Gridsearch. The identified optimal model, characterized by a hyperparameter C value of 100, achieves a training accuracy of 73%. Additionally, the validation accuracy is reported at 74.6%. These results suggest a reasonable level of generalization from the training set to the validation set, as the training and validation accuracies are relatively close.

13 Conclusion

In conclusion, our customer segmentation project successfully harnessed advanced techniques to derive meaningful features from the existing dataset. Leveraging clustering algorithms with different feature groups, coupled with dimensionality reduction, we optimized our approach through Gridsearch for hyperparameter tuning. The resulting customer categories effectively represented the underlying customer population. Moreover, validation through classification, employing SVC models, demonstrated the robustness of our segmentation by yielding a superior classifier with satisfactory results. This project not only enhances our understanding of customer behavior but also provides a practical and effective framework for businesses to tailor their strategies based on distinct customer segments.

References

[1] J. Nurma Sari, L. Nugroho, R. Ferdiana, and P. Santosa, "Review on customer segmentation technique on ecommerce," *Advanced Science Letters*, vol. 22, pp. 3018–3022, 10 2016.

[2] Y. Wang, X. Yang, L. Zhang, X. Fan, Q. Ye, and L. Fu, "Individual tree segmentation and tree-counting using supervised clustering," *Computers and Electronics in Agriculture*, vol. 205, p. 107629, 2023.

[3] T. Kansal, S. Bahuguna, V. Singh, and T. Choudhury, "Customer segmentation using k-means clustering," in *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, pp. 135–139, 2018.

[4] H. C. v. T. Laura Kazbare and J. K. Eskildsen, "A-priori and post-hoc segmentation in the design of healthy eating campaigns," *Journal of Marketing Communications*, vol. 16, no. 1-2, pp. 21–45, 2010.